



Diseño de procesamiento de datos

Encuesta de Microemprendimiento (EME)

Valentina Andrade, analista, y Nicolás Godoy Márquez, practicante

2022-02-10

Tabla de contenidos

1	Presentación	3
1.1	Enumeración de subprocesos	3
2	Integración de datos	5
2.1	Combinar fuentes	6
2.2	Disponer base	12
3	Clasificación y codificación	14
3.1	Ejectuar clasificación y codificación	14
3.2	Validar clasificación y codificación	19
3.3	Consideraciones	21
4	Revisión y validación	23
4.1	Revisar y validar	23
4.2	Indicadores	28
5	Edición e imputación	29
5.1	Revisión	29
5.2	Editar e imputar datos	30
6	Derivación de nuevas variables y unidades	32
7	Cálculo de ponderadores	33
8	Cálculo de agregados	34
9	Finalización de los archivos de datos	35
10	Bibliografía	36

1 Presentación

El presente documento tiene por objeto especificar las tareas y actividades de los ocho subprocesos, así como sus respectivas actividades y tareas, que corresponden al procesamiento de los datos recolectados en el marco del **Piloto de la VII Encuesta de Microemprendimiento (EME)** realizada por el *Instituto Nacional de Estadísticas* en 2022. Todo ello se basa en el **Modelo Genérico del Proceso Estadístico (GSBPM)** en su versión 5.1, adaptándolo a la realidad institucional.

1.1 Enumeración de subprocesos

A partir del mapa de procesos correspondientes al segmento de Negocio del INE, los procesos a seguir son los siguientes:

1. Gestión de calidad y metadatos
2. Infraestructura estadística
3. Detección y evaluación de necesidades.
4. Diseño y planificación.
5. Construcción.
6. Recolección de datos.
7. Procesamiento.
8. Análisis de resultados.
9. Difusión.
10. Evaluación y retroalimentación.
11. prueba

Sin embargo, como ya se señaló, este documento especificará la ruta a seguir para cumplir con el proceso de procesamiento de datos recolectados en el marco de la prueba piloto de la VII EME. Es decir, se enmarca en el proceso de **Diseño y planificación**; en particular, en la primera etapa “*Diseñar procesamiento*” del subproceso *Diseñar el procesamiento y análisis*. En ese sentido, se describirán las actividades, roles responsables y mecanismos de control utilizados durante la ejecución de cada actividad, a la vez que se presentará el glosario de definiciones relacionadas, el listado de documentos aplicables de referencia y el marco legal del subproceso. De tal modo, los subprocedimientos particulares a detallar son los siguientes:

1. Integrar datos
2. Clasificar y codificar
3. Revisar y validar
4. Editar e imputar
5. Derivar nuevas variables y unidades
6. Calcular ponderaciones
7. Calcular agregados
8. Finalizar archivos de datos

Todo ello tiene por objetivo **integrar, clasificar, verificar, limpiar y transformar los datos**, de modo que estén listos para ser *analizados y difundidos*. Las distintas fases del análisis, si bien tienden a ser secuenciales, pueden ser iterativas y paralelas. Es relevante considerar que los resultados del piloto estadístico pueden generar transformaciones en los subprocesos a detallar, así como en sus respectivas tareas y actividades.

Como resultado final del procesamiento, se esperan dos productos principales

1. **Base Full VII EME**: contiene toda la información de los módulos A al K, exceptuando la glosa detallada correspondiente a la pregunta f2, que entrega información correspondiente a las y los trabajadores del microemprendimiento.
2. **Base Empleo VII EME**: presenta información detallada del empleo generado por cada microemprendimiento, que ha sido capturada en la pregunta f2.

El procesamiento también incluirá bases de datos que permitirán analizar la calidad del proceso de recolección de datos. En particular, una base de datos que incluye la **Hoja de ruta** de los encuestadores, y la **Disposición final de casos**.

2 Integración de datos

Este primer subproceso busca **integrar y combinar** la información de una o más fuentes de datos, con el objeto de generar una base de datos integrada. En específico, si bien el piloto de la VII EME cuenta con sólo una submuestra, se emplearon tres formatos de recolección de información: presencial en papel, presencial en DMC y telefónico en DMC. Siguiendo las siglas de cada método de recolección:

- *Papel and Pencil Personal Interviewing* (**PAPI**)
- *Computer Assisted Personal Interviewing* (**CAPI**)
- *Computer Assisted Telephonic Interviewing* (**CATI**)

Además, como ya se señaló en el apartado de presentación, se trabajará con cuatro bases de datos diferentes:

1. **Hoja de ruta:** contiene información disponible en las hojas de ruta de las y los encuestadores, respecto de todos los informantes seleccionados, tanto de las encuestas logradas como de las que no se lograron. Incorpora variables de identificación del informante, de su hogar y de su vivienda, así como de caracterización de las visitas, observaciones constatadas en la hoja de ruta, e identificadores del encuestador y sus supervisores. En este caso, cada fila corresponde a una visita realizada a un informante determinado, por lo cual el total de observaciones corresponde al total de informantes elegibles multiplicado por el total de visitas realizadas a cada uno de ellos.
2. **Disposición final de casos:** busca presentar el código de disposición final del levantamiento de los datos, con independencia del estado de logro del levantamiento. Esta base tiene una importancia crucial para la construcción de los factores de expansión, pues permite distinguir entre informantes a) elegibles que responden; b) elegibles que no responden; c) no elegibles; y d) elegibilidad desconocida. Ello necesita parte de la información presente en la Tarjeta de Registro de Hogares (TRH) y del cuestionario final. Así, comprende variables de identificación del informante; variables sociodemográficas relevantes; el código de disposición final del levantamiento y la supervisión; variables de observaciones de la hoja de ruta; y variables de identificación de encuestador y el equipo de supervisión.
3. **Base Full VII EME:** contiene toda la información recolectada a través del cuestionario aplicado a los informantes, por lo que provee información que detalla las características de las y los microemprendedores y sus microemprendimientos o actividades por cuenta propia, salvo aquella incluida en la pregunta f2. El número de observación es igual al número de encuestas logradas o interrumpidas.
4. **Empleo VII EME:** contiene toda la información recabada en la grilla relacionada con la pregunta f2. Ello implica que el total de observaciones será igual al número de trabajadores declarados por cada uno de los informantes. Así, las y los microemprendedores sin trabajadores no serán incluidas/os en esta base.

Así, el subproceso **5.1 Integración de datos** será aplicado en cada una de las cuatro bases de datos, a modo de asegurar los estándares de calidad correspondientes en todos los casos.

La información recabada en los tres medios deberá ser unificada, lo cual se realizará en R, con librerías vinculadas al paquete *tidyverse*. Es relevante considerar que un problema en el cuestionario significó la repetición de algunas encuestas, por lo cual será necesario eliminar los folios duplicados, conservando sólo la segunda aplicación del cuestionario a cada informante. Ello también se realizará en R, a partir de librerías vinculadas a *tidyverse*. El presente subproceso contempla dos etapas, con diversas actividades: **combinar fuentes** y **disponer base**.

Es fundamental destacar que los scripts en que se realizó el subproceso de integración en su conjunto se encuentran en la carpeta R del repositorio de GitHub eme-ine/procesamiento. El archivo **01integrar.R** presenta el trabajo de integración general, mientras que en los archivos **01integrar-cdf-eme.R**, **01integrar-hr-eme.R**, **01integrar-full-eme.R** y **01integrar-empleo-eme.R** se encuentran los procedimientos específicos para la integración que resulta en las bases de Código de Disposición Final, Hoja de Ruta, Full y Empleo, respectivamente.

2.1 Combinar fuentes

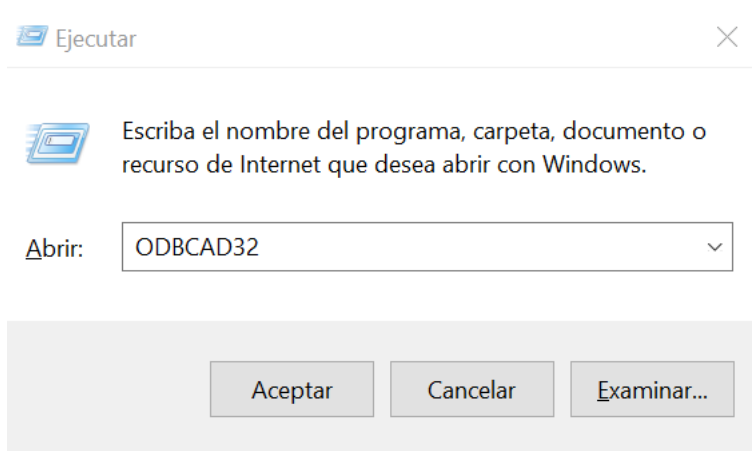
Esta etapa consta de combinar los datos provenientes del proceso de recolección de datos. El resultado es la conformación de cuatro bases de datos integradas.

2.1.1 Solicitar datos

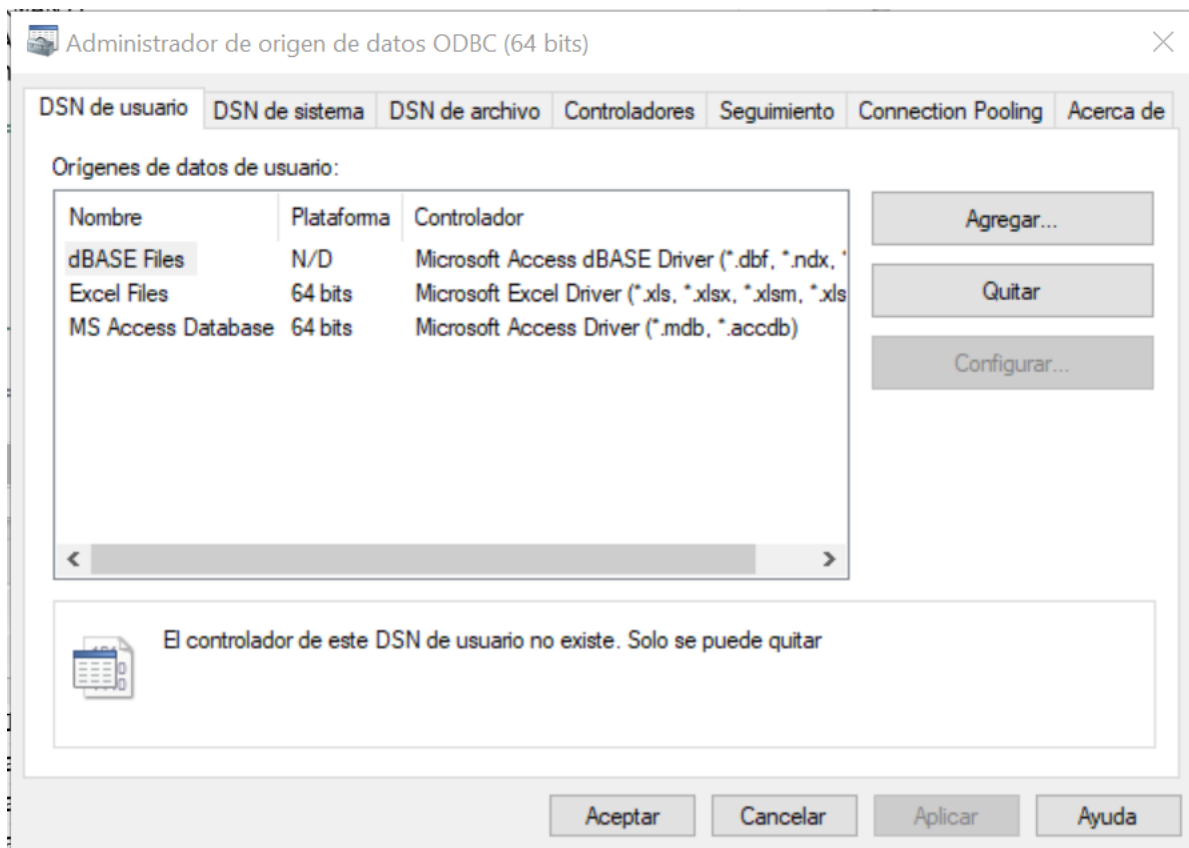
Para la presente actividad, se debe recibir la base de datos provenientes del proceso de Recolección de datos, a partir de la solicitud elevada por parte de la Subdirección Técnica (SDT) a la Subdirección de Operaciones (SDO). En el caso de los datos recolectados en DMC, la solicitud será enviada vía correo electrónico por parte del encargado del equipo responsable de la VII EME a la SDO, especificando la estructura deseada para los datos, así como el total de observaciones para cada base y las variables que cada una debe incluir. Todo ello está detallado en los documentos **Estructura Bases de Datos VII Encuesta Microemprendimiento (VII EME)**, tanto en formato Word como en planila Excel.

En el caso de los datos recolectados en formato papel, la descarga será realizada por la analista encargada del procesamiento, directamente desde el servidor (SQL Server) de la Institución. Los pasos a seguir son los siguientes:

- a) Teclear el botón de Windows + R, e ingresar el código ODBCAD32.

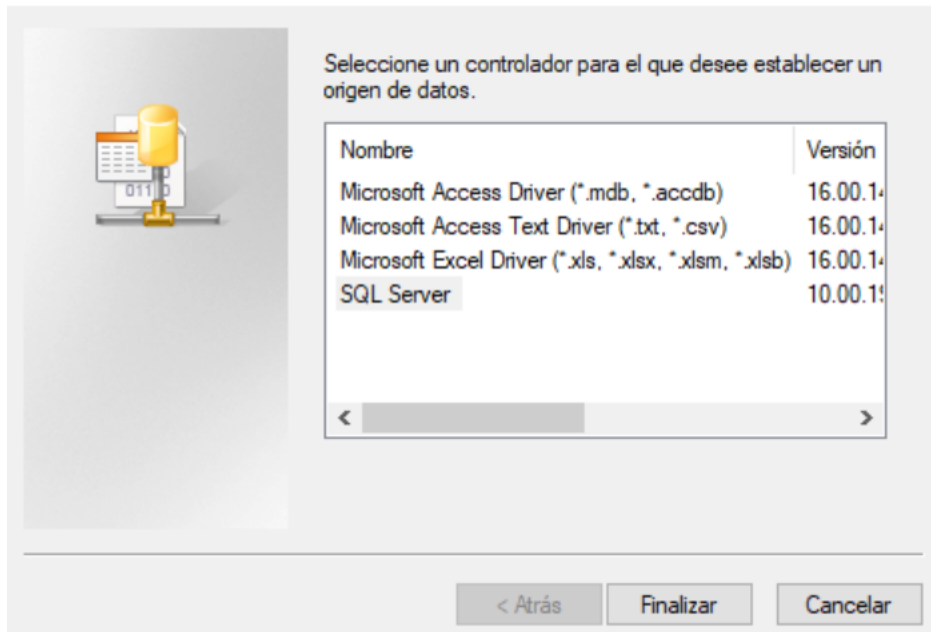


- b) Se abrirá una ventana emergente, llamada Administrador de origen de datos ODBC. Debe seleccionarse *dBASE Files*, y luego *Aceptar*.

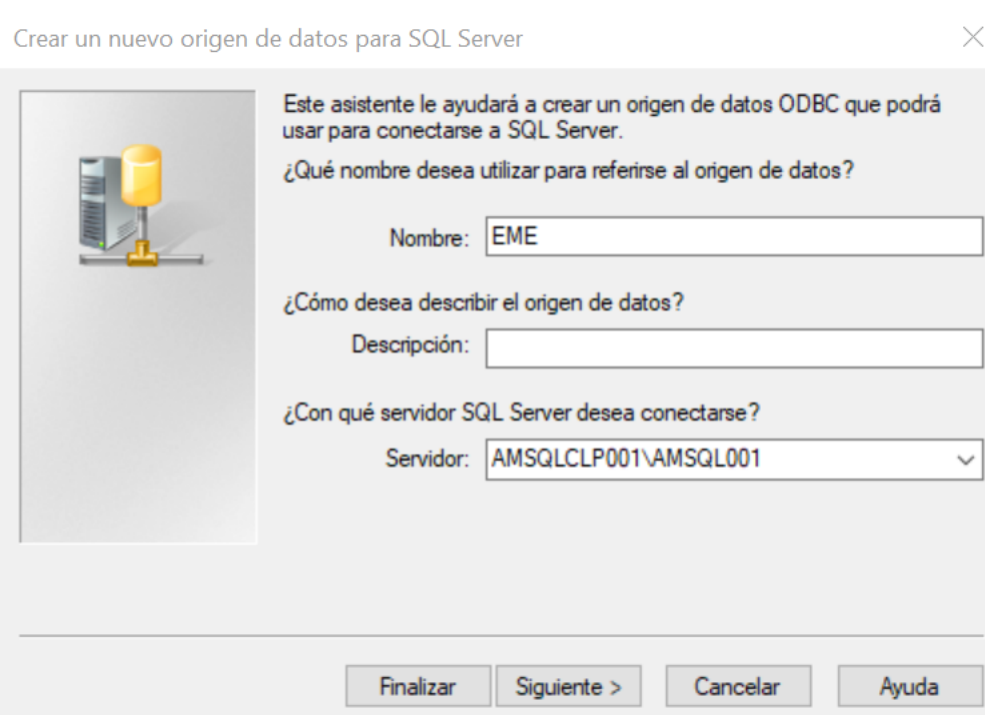


- c) Posteriormente, se ha de seleccionar **SQL Server** y *Finalizar*.

Crear nuevo origen de datos

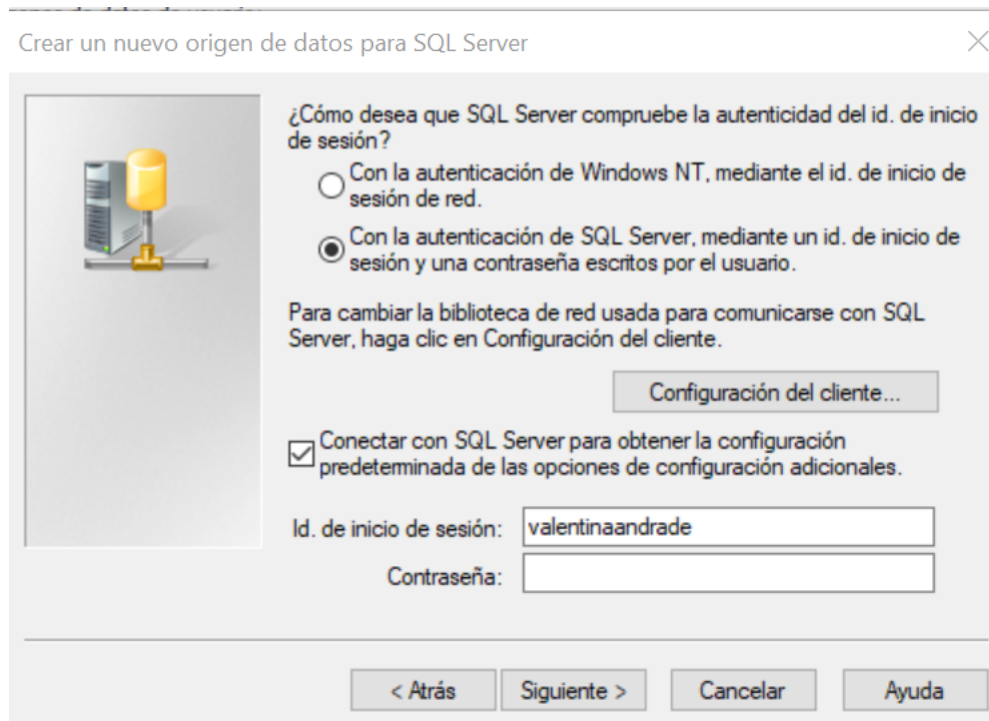


- d) Posteriormente, se seleccionó el nombre **EME** para referirse al origen de los datos. Asimismo, se elige conectarse al servidor SQL Server **AMSQLCLP001 AMSQL001**, según lo indicado en la imagen. Luego, seleccionar *Siguiente >*.



- e) Luego, se debe seleccionar que la comprobación de autenticidad del id. de inicio de sesión realizada por SQL Server se realice **con la autenticación de SQL Server, mediante**

un id. de inicio de sesión y una contraseña escritos por el usuario. En este caso, la Id. de inicio de sesión es el usuario de la analista. Hay que asegurarse de seleccionar la opción *Conectar con SQL Server para obtener la configuración predeterminada de las opciones de configuración adicionales.* Luego seleccionar *Siguiente >*.



Crear un nuevo origen de datos para SQL Server

¿Cómo desea que SQL Server compruebe la autenticidad del id. de inicio de sesión?

☐ Con la autenticación de Windows NT, mediante el id. de inicio de sesión de red.

☒ Con la autenticación de SQL Server, mediante un id. de inicio de sesión y una contraseña escritos por el usuario.

Para cambiar la biblioteca de red usada para comunicarse con SQL Server, haga clic en Configuración del cliente.

[Configuración del cliente...](#)

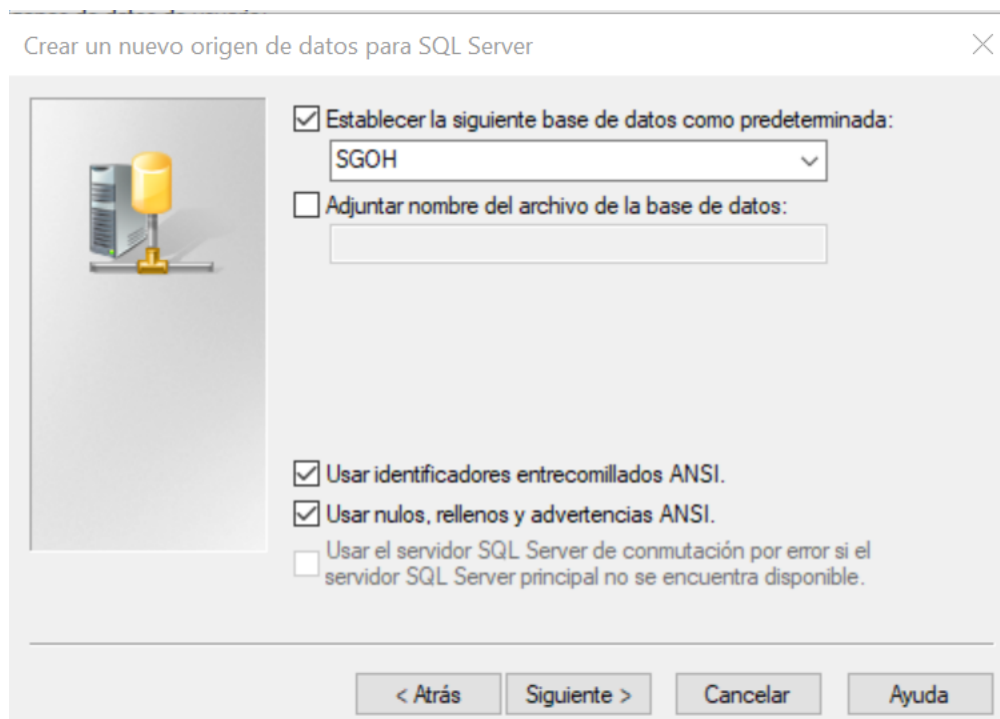
☒ Conectar con SQL Server para obtener la configuración predeterminada de las opciones de configuración adicionales.

Id. de inicio de sesión:

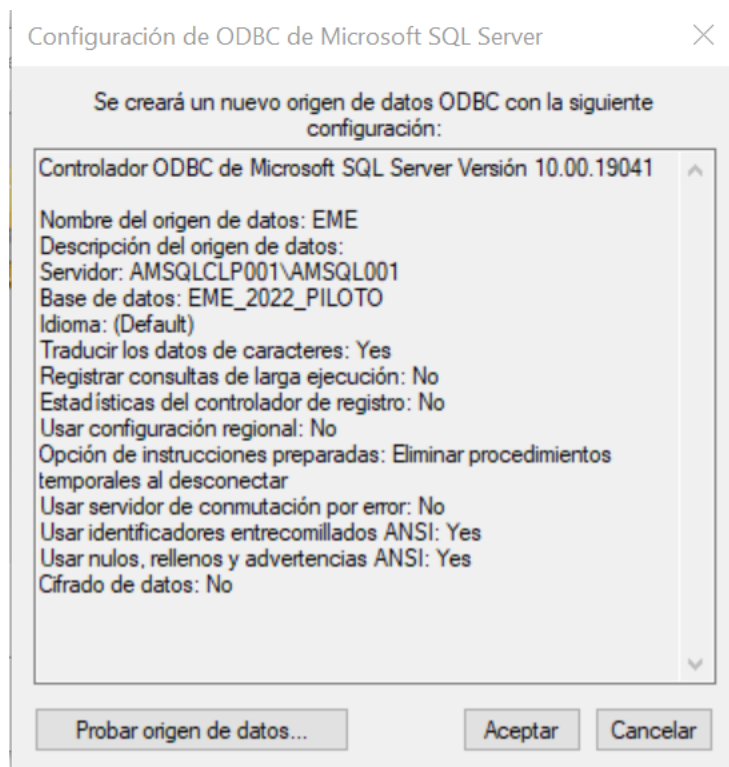
Contraseña:

< Atrás Siguiente > Cancelar Ayuda

- f) Después, se debe establecer **SGOH** como la base de datos predeterminada. Se ha de estar seguro de que están seleccionadas las opciones **Usar identificadores entrecomillados ANSI** y **Usar nulos, rellenos y advertencias ANSI**.

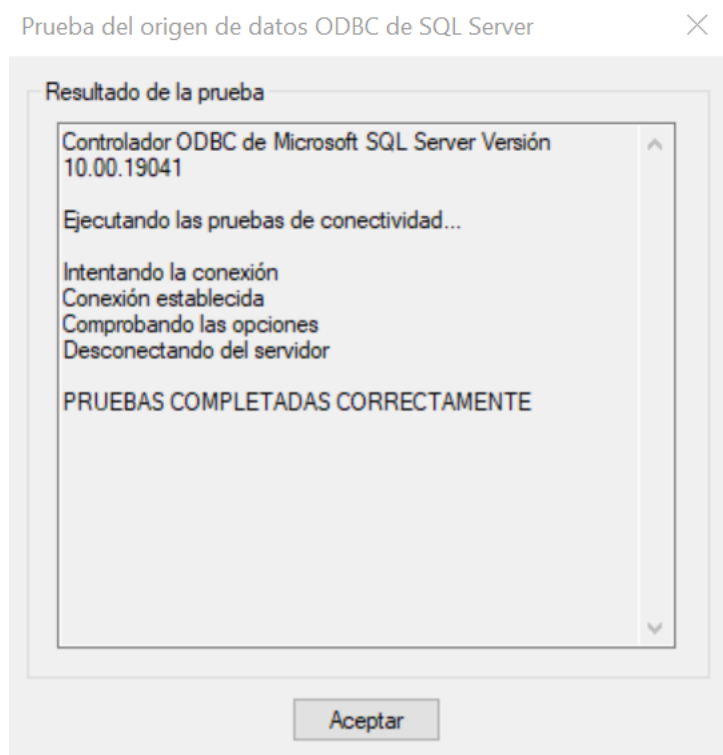


g) Si el proceso se ha desarrollado adecuadamente, aparecerá la siguiente ventana emergente. Hacer click en **Aceptar**.



h) Por último, se ejecutarán las pruebas de conectividad. Hay que esperar a que aparezca el

texto **PRUEBAS COMPLETADAS CORRECTAMENTE** y, por último, hacer click en **Aceptar**.



Es fundamental destacar que los datos serán recibidos a medida que se lleve a cabo el proceso **4. Recolección de datos**, por lo que el equipo de procesamiento debe encargarse de actualizar los datos recibidos de manera periódica.

2.1.2 Compilar datos

Las bases de datos solicitadas en la actividad anterior serán compiladas en el repositorio de GitHub eme-ine/procesamiento. En particular, las bases de datos están alojadas en la carpeta **input/data**. Es fundamental actualizar los datos alojados en el repositorio, a medida que se reciban los datos recolectados en el proceso de Recolección.

2.1.3 Preparar/verificar datos

Los datos alojados en el repositorio de GitHub en la actividad anterior han de ser revisados en R, a modo de confirmar que se cuente con la información necesaria para construir las cuatro bases de datos. Es fundamental verificar, con cada actualización de las bases de datos, que estas contemplen la totalidad de variables y observaciones definidas en los documentos de Estructura de bases de datos. Posteriormente se ha de llevar adelante una **validación en tercer nivel**, que incluye la aplicación de métodos automatizados o manuales que permitan identificar fenómenos irregulares o anómalos originados en el proceso de Recolección

de datos, como puede ser información incompleta, falta de información, eliminación de datos duplicados, entre otros. Tal validación se realizará en R, y los respectivos scripts se encuentran en la carpeta R del repositorio de GitHub anteriormente mencionado. Luego, en caso de hallar inconsistencias, es necesario enviar un correo electrónico al encargado/a o responsable en la Subdirección de Operaciones (SDO) de entregar tales datos, o al encargado del SGAD o SDT, identificando de manera explícita los datos observados, y solicitando subsanar las observaciones indicadas para entregar la base de datos corregida. Como resultado, se obtendrá una base de datos verificada por el o la analista especialista SDT. Hay que cerciorarse de *verificar el porcentaje de registros duplicados o con errores*.

La verificación semi-manual de los datos arrojó diversas inconsistencias, así como la ausencia de algunas variables. El informe respectivo se encuentra en la carpeta `input/docs/01estructura_datos/revision-full-eme-final.rtf`, del repositorio de GitHub `eme-ine/procesamiento`.

2.1.4 Integrar datos

El input de esta actividad es la base de datos verificada por las/os analistas especialistas de la SDT. Las tareas a realizar son, en primer lugar, garantizar que la estructura de los datos a unificar sea compatible - por ejemplo, asegurándose de que los tipos de datos de las variables a unificar sean iguales en cada base de datos -, así como asegurarse de que cada una cuenta con un campo de identificación para las y los informantes, encuestadores y supervisores, según corresponda. Una vez comprobada la compatibilidad, se procede a programar la integración de los datos recogidos en papel y DMC a través de R. Luego se transforman los datos que necesiten la aplicación de procedimientos de sustitución o reducción de datos en las variables, para posteriormente revisar la correcta aplicación de datos procedimientos. Es necesario identificar la proporción de registros duplicados, así como la proporción de unidades enlazadas manualmente. La Jefatura del Subdepartamento de Estadísticas Estructurales ha de revisar los datos para comprobar que la actividad se realizó en razón del procedimiento estipulado. El resultado de esta actividad es la conformación de una base de datos integrada, almacenada en el repositorio de GitHub `eme-ine/procesamiento`.

El procedimiento de integración se realizó en un script de R general llamado `01integrar.R`, presente en la carpeta R del repositorio de GitHub anteriormente señalado. Además, los procedimientos específicos necesarios para la integración de cada una de las bases de datos se encuentran descritos en sus respectivos scripts: `01integrar-cdf-eme.R`, `01integrar-hr-eme.R`, `01integrar-full-eme.R` y `01integrar-empleo-eme.R`.

2.2 Disponer base

La segunda etapa tiene por objeto garantizar que la integración de las bases de datos cumpla con dimensiones de estructura de acuerdo a las condiciones del medio digital donde se almacenará la información, quedando disponible para el procesamiento del microdato. A partir de esta estructura de datos, corresponde iniciar en paralelo actividades de procesos y sub-

procesos posteriores, como puede ser el análisis de microdatos y la detección de componentes de comportamiento atípico (*outliers*).

2.2.1 Verificar base integrada

A partir de la base de datos integrada en la etapa anterior, las tareas a realizar constan de la revisión de la correcta integración de tales datos, para luego identificar irregularidades originadas por la integración de los datos, como su completitud y consistencia, a través de métodos manuales o automatizados. La revisión se realizará en R. Por último, se verifica la conformidad del procedimiento de integración de bases respecto de los protocolos establecidos. En caso de encontrar errores, inconsistencias o incongruencias en la integración, el profesional responsable en la SDT ha de informar por correo al responsable de la actividad Integrar datos de la etapa anterior, para iniciar nuevamente el proceso a partir de la compilación de datos. Es fundamental verificar el porcentaje de registros duplicados o con errores en los datos integrados, así como la realización de una revisión por parte de la Jefatura de Departamento y/o Subdepartamento, que asegura que la verificación se realizó correctamente. Como resultado de esta actividad se tienen las bases de datos integradas en la etapa anterior, con controles de revisión aprobados, y almacenada en el repositorio de GitHub `eme-ine/procesamiento`.

2.2.2 Entregar base integrada

Partiendo de las bases de datos integradas, y con controles de revisión aprobados, hay que proceder a formalizar la disposición de los datos según el protocolo para Liberar base, cumpliendo con el medio de divulgación, los plazos y la identificación de cargos que podrán acceder a los datos integrados. Así, las bases integradas y revisadas serán alojadas en la carpeta `input/data` del repositorio de GitHub `eme-ine/procesamiento`. Es relevante verificar que se cumpla con el protocolo de liberar la base, así como que esta y el proceso de integración sea revisado por la Jefatura del Subdepartamento de Estadísticas Estructurales, confirmando la realización del proceso en conformidad del procedimiento estipulado. El producto de esta actividad es la base integrada, almacenada en el repositorio de GitHub `eme-ine/procesamiento/output/data`.

3 Clasificación y codificación

Tal como lo indica su nombre, este proceso tiene por objeto *clasificar y codificar los datos*, procesos distintos mas complementarios para la garantización de la precisión y confiabilidad de las estadísticas elaboradas en los distintos niveles de análisis. Mientras que clasificar refiere a asignar un código de clasificación a cada informante, la codificación implica transformar la información textual de una observación en un código que identifica una categoría correcta para tal observación.

El presente subproceso cuenta con dos etapas. Primero, **ejecutar la clasificación y codificación**, que tiene por objeto generar datos clasificados y codificados en razón de los lineamientos establecidos en el proceso de Diseño y Planificación; luego, **garantizar que la clasificación y la codificación** ejecutada en la base de datos cumpla con los criterios de calidad señalados en el proceso de Diseño y Planificación.

3.1 Ejecutar clasificación y codificación

3.1.1 Revisar y preparar insumos

Como input de esta actividad, es necesario contar con el modelo del algoritmo Support Vector Machines (SVM) así como con su respectivo script, y recibir la base de datos integrada y validada en el subproceso anterior. Las tareas a realizar constan de, en primer lugar, la preparación y revisión de los insumos principales para ejecutar la clasificación y codificación de textos, en particular, el script y los datos integrados en el subproceso anterior, asegurándose de su completitud en términos tanto de cantidad de variables como de registros. Posteriormente, se procede a la actualización periódica del aprendizaje de la codificación y la clasificación de las glosas, derivada del análisis especializado en clasificadores para, por último, cargar en el sistema informático (**GITHUB?**) los insumos derivados de tal tarea, para ejecutar el modelo SVM sobre los textos digitados. Es fundamental aplicar controles de completitud a la base de datos integrada, a modo de verificar que el procedimiento de revisión cumpla con los lineamientos establecidos. Los productos de esta actividad son un script que instruye las fases del proceso de clasificación, así como la base de datos integrada y validada en el subproceso de integración de los datos.

3.1.2 Ejecutar codificación automática

Para esta actividad es necesario tener a disposición tanto la base de datos integrada como la de entrenamiento, y el modelo de algoritmos SVM junto con su respectivo script. Entre las tareas que realizar hay que, en primer lugar, cargar en el repositorio de GitHub emeine/procesamiento todos aquellos insumos que derivan del entrenamiento de los modelos, junto con las respectivas glosas a codificar. Luego, ejecutar un script de manera local donde se realicen los procesos a) *clean data* o normalización de nuevas glosas a codificar, y b) *tokenización* de las glosas por codificar, es decir, la identificación de palabras individuales

observadas en las glosas, vinculadas a la vez con las categorías del clasificador. Posteriormente, se procede a ejecutar la función de predicción a partir de una matriz que represente numéricamente la glosa y el modelo determinado por cada operación estadística. Después se ejecuta la codificación automática a partir del modelo SVM. Por último, si hay una baja probabilidad de aceptar la glosa en la codificación en razón de lo estipulado en el proceso de Diseño y planificación, hay que enviarla a la siguiente actividad Revisar, para completar la codificación de las glosas. Como control de la actividad, hay que verificar que se ejecutó la clasificación y codificación de todas las glosas correspondientes. Entre los resultados de la presente actividad se tendrán una base de datos con registros clasificados y codificados automáticamente, y una solicitud de clasificación y codificación para aquellas glosas con baja probabilidad, o que presentan inconsistencias a la hora de ser aceptadas.

3.1.3 Revisar base de datos

A partir de los datos integrados en el subproceso anterior, así como de las respuestas a las consultas que surgiesen en la actividad precedente, de los clasificadores estadísticos empleados y la tabla de correspondencia del subproceso de Diseño del procesamiento y análisis, la presente actividad consta de dos tareas. Primero, revisar tanto la completitud como las condiciones de los insumos recibidos para ejecutar el procedimiento de clasificación y codificación de los registros de los datos integrado; en segundo lugar, en caso de no cumplir con los criterios de calidad y completitud establecidos, se habrá de regresar los datos al subproceso anterior, especificando qué es lo que se deberá subsanar. Ello endrá como resultado una base de datos revisada y lista para dar inicio a la clasificación y codificación. Es fundamental verificar que el procedimiento de revisión cumpla con los lineamientos establecidos en el subproceso de Diseño.

3.1.4 Asignar carga de trabajo

A partir de los datos revisados para iniciar la clasificación y codificación, se procede a, primero, verificar la disponibilidad de personal para realizar el procedimiento de clasificación y recodificación, para luego revisar y aplicar los criterios que permiten asignar carga laboral al codificar y, por último, entregar el detalle de la carga laboral de cada codificar, así como los plazos de cumplimiento según la forma y tiempos establecidos. Es fundamental revisar que la asignación fue correcta, en la línea de cada operación estadística. Ello tendrá como resultado la asignación de carga individual del equipo de codificadores.

3.1.5 Clasificar y codificar los registros

Las entradas de la presente actividad son los clasificadores estadísticos, el compilado de respuestas de consultas en actividades anteriores, la tabla de correspondencia para la codificación, la carga laboral individual de los codificadores, la base de datos revisada y lista para iniciar la clasificación y codificación, las glosas codificadas provenientes de la actividad

Revisar, y los registros sin codificar provenientes de la etapa 2 Validar clasificación y codificación. Las tareas a realizar son: primero, revisar las glosas para interpretar y lograr ubicarlas dentro del sistema de clasificación; segundo, asignar el código alfanumérico apropiado, coherente y proviamente establecido según el sistema de clasificador determinado, incorporando el código digitado a la base de datos definitiva; tercero, si el codificar no puede asignar algún código por falta de información, inconsistencia en la glosa o por registros desconocidos o difíciles de clasificar, se debe acudir a los revisores enviando un correo electrónico en que se señalen los casos identificados, quienes resolverán las dudas planteadas. Hay que comprobar que se clasificaron todos los registros asignados. El resultado de la actividad es una base de datos con registros clasificados y codificados automática y manualmente.

En particular, tomando la experiencia de versiones anteriores de la EME, hay dos tipos de variables que requieren de un proceso de codificación específico:

3.1.5.1 Rama de actividad económica (c1) y ocupación u oficio (c2) Las preguntas que recaban información sobre la actividad económica y la ocupación u oficio de las y los informantes son preguntas abiertas, cuyos datos son posteriormente clasificados según las clasificaciones internacionales **CAENES** (Clasificador de Actividades Económicas Nacional para Encuestas Sociodemográficas) y **CIUO** (Clasificador Chileno de Ocupaciones, o *CIUO08.CL*), a nivel de subgrupo principal (dos dígitos). Si bien en la VI EME el trabajo de codificación estuvo, principalmente, a cargo del Departamento de Procesamiento de la Encuesta Nacional de Empleo (ENE), en la presente versión de la encuesta se utilizará la experiencia de tal equipo para que sea el propio equipo de procesamiento aquel encargado de realizar la codificación.

El subproceso de codificación de la rama de actividad económica y del oficio y ocupación en que se desempeñan las y los informantes consta de cinco pasos:

1. Codificación automática de las respuestas registradas: se realizará para cada submuestra a partir del modelo Support Vector Machine (SVM), el mismo utilizado por la ENE.
2. Selección de casos a auditar: luego de la codificación automática, se auditarán aquellos casos que indicaron seguir trabajando como independientes, pero habiendo cambiado su actividad económica ($a1 = 2$), así como a los casos que mantuvieron su actividad económica ($a1 = 1$), pero cuyas flosas de clasificador a 2 dígitos para la ENE y para la EME difieren.
3. Clasificación de glosas por parte de Auditoría de Unidad de Nomenclatura: se presentarán los casos seleccionados para la auditoría a la Unidad de Nomenclatura, que establecerá su acuerdo o disconformidad con la codificación automática realizada en el paso 1.
4. Re-revisión de las glosas, en caso de ser distintas: el equipo de procesamiento de la EME revisará los casos en que la Unidad de Nomenclatura proponga un código distinto al asignado por medio de la codificación automática, codificando a la vez aquellos casos que la Unidad de Nomenclatura no pudo clasificar.
5. Análisis de consistencia de las glosas con otras preguntas: es importante revisar los casos clasificados como “Hogares como empleadores,” en tanto no pueden ser trabajadores

independientes, y deben dejarse fuera del marco muestral. Asimismo, es fundamental chequear que quienes hayan sido clasificados en la rama económica agrícola perciban ingresos provenientes de actividades agrícolas indicadas en la grilla d5.

3.1.5.2 Especifique El cuestionario del piloto de la VII EME cuenta con 29 preguntas con la alternativa “*Otro/a, especifique*”, de texto abierto, en que la o el informante detalla específicamente su respuesta, en caso de que esta no se presente entre las alternativas principales propuestas en el cuestionario original.

En específico, las preguntas que presentan la alternativa “*Otro/a, especifique*” son las siguientes:

1. A3 ¿Por qué razón terminó su trabajo como Empleador/a o Cuenta Propia?
2. A5 ¿Por qué razón no volvería a trabajar de forma independiente?
3. A8 ¿Actualmente a qué se dedica?
4. A9 ¿Cuál es el principal motivo por el cual se encuentra trabajando como asalariado/a?
5. B2 ¿Antes de iniciar con este negocio o actividad, usted...?
6. B3 ¿Cuál fue la motivación principal por la cual inició su actual negocio o actividad por cuenta propia?
7. C9 ¿Donde lleva a cabo principalmente su negocio o actividad por cuenta propia?
8. C10 ¿El local o las instalaciones en las que trabaja son?
9. E4 ¿Cuál es la principal razón por la que no ha iniciado actividades de su negocio o actividad por cuenta propia ante el Servicio de Impuesto Internos (SII)?
10. E6 ¿Cuál es la principal razón por la que inició actividades de su negocio o actividad por cuenta propia ante el Servicio de Impuestos Internos (SII)?
11. F3 ¿Cómo suele fijar los salarios de sus trabajadores/as?
12. G6 ¿Cuál es la principal razón por la que no ha solicitado un crédito?
13. G7 ¿Cuáles fueron los motivos por los que no obtuvo crédito? *Selección múltiple, considerar todos los créditos no obtenidos* → Considerar créditos COVID y razones de no elegibilidad.
14. G13 ¿Cómo financia actualmente los gastos regulares del negocio (compra de materias primas, salarios, cuentas, entre otros)?
15. I2 Señale la principal razón por la cual no utiliza internet en su negocio o actividad por cuenta propia.
16. I4 ¿Dónde almacena la información que utiliza para el funcionamiento de su negocio o actividad por cuenta propia (ya sea esta contable, relativa a sus clientes, etc.)?
17. I5 Del siguiente listado, ¿qué usos le da a internet en su negocio o actividad por cuenta propia?
18. I6a Para realizar sus compras ¿qué tipo de canales digitales usa?
19. I6b Para realizar sus ventas ¿qué tipo de canales digitales usa?
20. I7 ¿Trabaja para una aplicación que contacta a sus clientes (ej: Uber, Cabify, AirBnb, Glovo, Rappi, Cornershop y otros similares)?
21. J2 ¿Cuál fue el principal beneficio de la última capacitación realizada?
22. J3 ¿Cuál fue la principal fuente de financiamiento de esta capacitación? → OJO
23. J4 ¿Cuál es la razón principal por la que no ha recibido ningún tipo de capacitación?

24. K1 ¿Cuáles son los dos aspectos más importantes que usted cree limitan al crecimiento de su negocio?
25. K2 Indique las dos afirmaciones que mejor reflejan los principales beneficios de ser independiente.
26. K12 ¿Cuáles de las siguientes formas de pago acepta para vender sus productos o prestar sus servicios?
27. K13 ¿Tiene conocimiento de los servicios o beneficios ofrecidos por alguna de las siguientes instituciones? → recién aparece financiamiento COVID, pero pregunta si los conoce.
28. K15 ¿Conoce alguno de los siguientes programas o beneficios?
29. K17 Además de los fondos mencionados, ¿destinó recursos personales para mantener en funcionamiento su negocio?

En este caso, el objeto es **categorizar** la mayor cantidad de respuestas literales posibles, a modo de perder la menor cantidad de información posible. Asimismo, ello puede contribuir a *modificar el cuestionario final de la VII EME*, en la medida que la categorización de las respuestas dadas por las y los informantes permitan presentar alternativas principales que resulten más exhaustivas, mejorando el proceso de respuesta de las y los entrevistados, evitando a su vez la redundancia entre las respuestas “Otro” y las categorías originales.

3.1.6 Revisar

Esta actividad parte de las consultas generadas en la actividad anterior, del compilado de respuestas a tales consultas, así como de los casos de glosas con baja probabilidad o que presentan inconsistencias para su aceptación, provenientes de la codificación automática. Las tareas para dar cumplimiento a esta actividad pasan por, primero, la revisión de la completitud de los antecedentes de la consulta; luego, la revisión del registro observado, a modo de asegurar que no es codificable; tercero, en caso de lograr la codificación, el revisor debe enviar una respuesta al responsable de la codificación manual para que ingrese tal codificación del registro; cuarto, si no fue posible lograr con tal tarea, el revisor ha de enviar un correo electrónico al responsable para dar respuesta del proceso de Recolección de datos, solicitando responder la consulta para solucionar los problemas de codificación; por último, en caso de haber consultas complejas para una correcta clasificación y codificación de las glosas, se ha de enviar un correo a la sección de Nomenclaturas, perteneciente al Subdepartamento de Calidad y Estándares. Como control, hay que verificar que todas las glosas correspondientes fueron revisadas, así como el preestablecimiento de procedimientos que permiten realizar procesos de entrenamiento a los modelos de codificación automática, la supervisión a la codificación manual, y tomar en cuenta las respuestas y retroalimentación a las consultas metodológicas realizadas a la sección de Nomenclatura, a partir de las auditorías de los clasificadores estadísticos. El output de esta actividad es la respuesta al responsable de la codificación manual para que ingrese la codificación del registro; la solicitud de recuperación de información a los equipos de trabajo de recolección; y las consultas a la sección de nomenclaturas.

3.1.7 Generar respuesta

A partir de la solicitud de recuperación de información o consultas, hay dos tareas a realizar. Primero, si es que se requiere recuperar la glosa, ha de enviarse una solicitud al responsable de la sección de supervisión y verificación (SDO). En tal caso, hay que recibir y revisar la consistencia del reporte, para luego chequear cada uno de los casos, determinado si se les ha clasificado en el código correcto o si, de lo contrario, ha de ser reclasificado. Si la glosa no es lo suficientemente coherente como para determinar su clasificación de manera certera, se ha de realizar una revisión cruzada en la cual distintos analistas realizan la clasificación, para luego comparar coincidencias en el código asignado. En caso de persegir la duda, se debe contactar a la sección de Nomenclaturas, compartiendo todos los antecedentes necesarios para definir una eventual reclasificación. Por último, la respuesta ha de ser enviada al equipo revisor Subdirección Técnica (SDT). En segundo lugar, si se identifican glosas insuficientes o inconsistentes, se debe enviar una consulta al responsable de la sección de Nomenclaturas, perteneciente al Subdepartamento de Calidad y Estándares. Para ello, primero se debe recibir y revisar la consistencia de la consulta, determinando el caso a partir de los estándares de calidad de la sección de Nomenclaturas, para finalmente enviar la respuesta al revisor por medio de un correo electrónico, respetando los plazos acordados entre las partes. Como resultado de esta actividad se obtiene la respuesta de consulta con la información recuperada o resuelta de manera satisfactoria.

3.1.8 Evaluar respuesta

La actividad de evaluar la respuesta debe realizarse a partir de la respuesta a la consulta realizada en la actividad anterior, con la información recuperada proveniente - dependiendo del caso - del proceso de Recolección de datos o de la sección de Nomenclaturas, así como del compilado de respuestas de consultas. Luego, se ha de recibir y evaluar la respuesta de la consulta: si tal evaluación es satisfactoria, se envía la respuesta a codificar, señalando la codificación correspondiente al registro, ingresando la consulta y respuesta al compilado; si tal evaluación no es satisfactoria, es decir, si no permite realizar la codificación, se habrá de re-enviar la consulta al encargado del proceso de Recolección o a la sección de Nomenclaturas, según corresponda. Hay que asegurarse de verificar que los reportes de respuesta contengan información relativa a todos los casos consultados. Los resultados de la actividad son, bien la solicitud de consulta al responsable de Recolección de datos o de la sección de Nomenclaturas, en caso de que la respuesta no sea satisfactoria; bien una respuesta satisfactoria al procedimiento de consulta.

3.2 Validar clasificación y codificación

3.2.1 Revisar base de datos

A partir de los datos clasificados y codificados en la etapa anterior. Entre las tareas a realizar están, primero, la revisión y verificación de que la base de datos recibida corresponda a la operación estadística y al período reportado, así como de la completitud de los insumos

recibidos a modo de ejecutar la revisión del procedimiento de clasificación y codificación de los registros de la base de datos. En caso de que esta última no cumpla con los criterios de completitud establecidos, se habrá de iniciar el procedimiento de completar la codificación de todas las glosas pendientes en la etapa correspondiente. Como output de esta actividad se tendrá la base de datos revisada.

3.2.2 Validar base revisada

Posteriormente, partiendo de la base de datos revisada, se ha de volver a chequear su completitud en relación con la cantidad de variables y registros, dando cuenta de que todas las glosas se codificaron y clasificaron de manera correcta. Si se requieren procedimientos de auditoría a partir de lo establecido en el proceso de Diseño y planificación, o por requerimientos de la sección de Nomenclaturas, se ha de seguir con la actividad *Entregar base de datos para auditar*; mientras que, de no contarse con una auditoría, se ha continuar con la actividad *Disponer base*. El resultado de la presente actividad es una base de datos validada.

3.2.3 Entregar base de datos para auditar

Una vez se cuente con la base de datos validada, se ha de solicitar por medio de un correo electrónico el inicio del procedimiento de auditoría a la unidad responsable, de acuerdo con el protocolo de aplicación de auditoría de clasificadores estadísticos, adjuntando la base de datos con los registros para ser auditados a través de Stata o R, lo cual tiene como resultado una base de datos para auditar. Es fundamental seguir el cronograma establecido, verificando el cumplimiento de la actividad en el periodo acordado entre las partes.

3.2.4 Auditar clasificación y codificación

A partir de la base de datos para auditar, y contando con el protocolo de proceso de auditoría de clasificadores estadísticos, se debe recibir la base de datos a auditar, aplicando la auditoría según el **protocolo Proceso de auditoría de clasificadores estadísticos**, para luego enviar por correo electrónico la base de datos auditada al responsable de la operación estadística. Es fundamental asegurarse de que todas las glosas hayan sido auditadas de manera adecuada. Todo ello tiene como resultado una base de datos auditada, con las observaciones y registros *conformes*, *no conformes*, *conformes con observaciones* y *sin clasificar (SC)*, así como un informe o minuta de auditoría.

3.2.5 Revisar base de datos auditada

Una vez se tenga la base de datos auditada, con las observaciones y registros *conformes*, *no conformes*, *conformes con observaciones* y *sin clasificar (SC)*, así como el protocolo de proceso de auditoría de clasificadores estadísticos, se ha de acceder a tal base de datos para revisar las observaciones y corregir en caso de considerar correctas las modificaciones sugeridas. Luego, se ha de informar a la unidad responsable de la auditoría la conformidad

del proceso, lo cual permite mejorar las bases de entrenamiento que, a su vez, preparan el modelo de clasificación automática de textos. Si no se está conforme con las observaciones de auditoría, se ha de enviar una respuesta con el detalle para dar inicio a un nuevo proceso de auditoría, identificando todos aquellos casos catalogados como no conformes. Se pondrá fin a la auditoría en virtud de los lineamientos del protocolo del proceso de auditoría de clasificadores estadísticos. Como resultado, se obtendrá una base de datos auditada y corregida con las observaciones del procedimiento de auditoría.

3.2.6 Disponer base de datos

A partir de la base de datos auditada y corregida con las observaciones del procedimiento de auditoría, o de la base de datos validada proveniente de la actividad Validar base revisada, según corresponda, se formalizará la disposición de la base según el protocolo establecido en el proceso de Diseño y Planificación, de modo que se debe cumplir con el medio de divulgación, plazos e identificación de los cargos que tendrán acceso a la base clasificada y codificada, así como con el protocolo institucional de seguridad de la información. Como resultado, se obtendrá una base de datos clasificada y codificada, validada y alojada en el repositorio de GitHub eme-ine/procesamiento.

3.3 Consideraciones

Es relevante revisar si hay preguntas que presenten una concentración atípicamente elevada en una de sus categorías de respuesta, dado que ello puede implicar una falta de exhaustividad en las alternativas: es decir, al no estar presente la categoría de respuesta que verdaderamente desean responder, es posible que las y los informantes opten por una de las alternativas a modo de suplir tal ausencia.

Asimismo, en la pregunta A3 (“¿Por qué razón terminó su trabajo como Empleador/a o Cuenta Propia?”), los motivos *trabajo de temporada* y *falta de clientes*, que son conceptualmente distintos, se encuentran juntos, por lo que habría que evaluar separarlos en próximas versiones del cuestionario.

Por su parte, la pregunta B6 (“¿Cómo financió o financiaron la puesta en marcha de esta actividad económica?”) no presenta la alternativa “Otro/a,” a la vez que no considera una o más categorías que aludan a los financiamientos que han surgido en el contexto de la Pandemia de Covid-19, como los retiros de fondos de las Administradoras de Fondos de Pensiones (AFP), el Ingreso Familiar de Emergencia (IFE), el Fondo de Garantía para Pequeños Empresarios (FOGAPE), o el Bono Clase Media, entre otros. Ello puede generar una alta concentración de datos en alternativas como No sabe (88) o No responde (99), lo cual redundaría en pérdida de información valiosa. Algo similar sucede con las preguntas G1 (“En los últimos dos años, ¿Ha solicitado usted o un tercero, alguno de los siguientes tipos de préstamos o créditos para fines de su actual negocio o actividad por cuenta propia?”) y G8 (“¿Tiene actualmente alguna de las siguientes deudas para fines del negocio o actividad por cuenta propia?”): ninguna incluye alternativas que reflejen la realidad económica gestada en el escenario pandémico.

Es fundamental guardar cuidado con las respuestas asociadas a la grilla de preguntas que se desprenden de la pregunta D3 (“*En los últimos 12 meses, ¿tuvo que incurrir en alguno de los siguientes gastos del negocio?*”), pues puede haber confusiones y errores de medición al ser una batería de preguntas compleja. Considerar un análisis factorial que permita determinar el grado de validez interna del constructo *gastos*.

Por último, en el módulo K no se presentan preguntas referentes a los precios de, por ejemplo, los insumos que emplean las y los emprendedores en el proceso productivo, y de posibles variaciones de tales en un periodo determinado.

4 Revisión y validación

El subproceso de revisión y validación de los datos tiene por objetivo la examinación de los datos a modo de *identificar potenciales problemas, errores y discrepancias* como valores atípicos, falta de respuesta de los ítems y codificación errónea, en un procedimiento iterativo a partir de reglas predefinidas. En ese sentido, se tomarán como base los procedimientos constatados en el informe final de la VI EME. Es fundamental considerar que este subproceso se encarga de la **identificación** de problemas en los datos, mientras que es en el subproceso siguiente en que se realizan las modificaciones a tales discrepancias.

4.1 Revisar y validar

4.1.1 Revisar insumos

Las entradas de la presenta actividad son la descarga de los datos integrados, clasificados y codificados desde el repositorio de GitHub *eme-ine/procesamiento*; el Manual o Instructivo de procesamiento con la especificación de la revisión y validación de los datos, creado en el subproceso Diseñar el procesamiento y análisis; el Manual metodológico *Documento de tratamiento de valores atípicos en VI EME*; la ficha metodológica de la operación estadística del proceso de la VII EME; el correo electrónico de los analistas de los subprocesos anteriores, donde se notifican las soluciones a los errores encontrados en la integración, clasificación y codificación; y la planificación interna del proceso de la VII EME.

4.1.2 Revisar a nivel de microdatos

La actividad busca revisar y validar los datos a nivel de microdatos, a modo de aseguar el cumplimiento de los criterios de calidad estipulados.

En este caso, la realización de la actividad de revisión de casos atípicos requiere de la base de datos clasificada y codificada en el subproceso anterior, junto con el presente manual de procesamiento, así como de la grilla técnica de revisión y validación de datos. También se necesita el manual metodológico y la ficha metodológica de la operación estadística.

La primera tarea por realizar es ejecutar las validaciones de primer y segundo nivel en R, lo cual incluye revisar la consistencia de los datos, así como los datos atípicos o fuera de rango, las observaciones de la encuesta, y por último revisar los datos a nivel de fuente o unidad, chequeando los criterios de flujos y rangos, validando las variables y la relación entre variables, esperando que se comporten según lo establecido en versiones anteriores. Ello incluye, por ejemplo, revisar la correlación entre la variable de gastos estimados y la gastos declarados, esperando que esta sea lo más alta posible. Asimismo, siguiendo la experiencia de la VI EME, es relevante revisar variables de

- a) Funcionamiento del negocio (c7);
- b) Remuneraciones (grilla d3_ (a, b y c) y variables f2_h sobre empleo);

- c) Impuestos (e6_1 sobre IVA, e6_2 sobre impuesto declaración de impuesto a la renta, e3_1, e3_5 y e3_6 sobre inscripción en el SII, y preguntas d3_ referentes a gastos);
- d) Herramientas (valorización de equipo de otros (h3 en razón de $h2_3 > 0$), propiedad de vehículos de transporte (h1_2 y h3_2) y de maquinaria específica (h1_3));
- e) Socios y trabajadores (considerar descuadres respecto de cantidad de socios por sexo y membresía del hogar (c4), compartir ganancias con socios que no pertenecen a la vivienda (d8), declaración de parte de utilidades en lugar de ganancias totales (d7), incongruencias en contratación de trabajadores (f1) y manera de fijar salarios (f3) en relación con trabajadores declarados (f2)); y
- f) Cuotas y créditos (problemas de flujo por no solicitud de crédito (g7), número de cuotas por pagar (g11), formas de pago (k11) e inconsistencias en el uso de internet (i1 e i3)).

4.1.3 Identificación de valores atípicos

Se entiende como valor atípico (u *outlier*) a toda aquella observación que, encontrándose dentro de los rangos establecidos para una variable, resultan atípicas, ya sea por su distancia respecto del resto de las observaciones dentro de la misma variable, como por su relación con valores reportados para otros atributos de la misma observación. Es decir, corresponde a una observación no ajustada al modelo de la mayoría de los datos, pudiendo encontrarse en la cola de la distribución estadística, o lejos del centro de los datos.

Siguiendo la experiencia de la VI EME, se empleará el modelo estadístico **bacon** (*blocked adaptive computationally efficient outlier nominators*), generando perfiles de informantes a partir de las variables **sexo** (nominal), **número de trabajadores/as asalariados** (numérica), **registro en el SII** (nominal) y **CAENES** (nominal).

En el caso de las *preguntas cuantitativas* (b9, d1, d2, d3c_1, d5d_1, d6c, d7_1, d7_2, d12, g9_1, h3_1), se realizará un conteo general de outliers. En el caso de las variables con distinta periodicidad (d3, d5 y d6) se realizará una anualización de todos los valores. Hay que constatar que, para cada uno de los perfiles, se obtendrán rangos distintos a la hora de identificar los outliers.

4.1.4 Tratamiento de outliers

En virtud de lo realizado en VI EME, se seguirán cuatro pasos para tratar a los outliers

1. Para cada caso identificado como outlier, se revisarán todas las observaciones. En caso de contar con un valor que justifique el valor identificado como atípico, entonces se mantendrá tal valor. De todos modos, se solicitará al equipo encargado de la recolección la revisión de la digitación en caso de que el valor atípico sea identificado en un cuestionario en formato de papel, a modo de rectificar que lo digitado se condiga con lo recolectado.
2. En caso de no justificar el valor atípico en el paso anterior, se revisará si el caso cuenta con algún otro outlier. Se asumirá coherencia en caso de que un informante presentase

dos o más outliers en el mismo módulo del cuestionario, por lo cual se mantendrá el valor original.

3. De no ocurrir lo constatado en el paso anterior, se evaluará la existencia de un outlier en las variables de ingreso anual, gasto anual y ganancia anual al percentil 30; de ser así, el dato será validado. Dado que la exigencia del modelo para identificar valores atípicos es muy alta, se utiliza este método para flexibilizar el criterio.
4. En caso de no satisfacerse las tres situaciones anteriores, se determinará que el valor no puede ser validado. Por ello, la respuesta del informante se reemplazará por un 96, es decir, '*sin dato*'. Esto, pues no es posible corroborar la veracidad de que el dato estuviese fuera de rango, yendo en la línea del tratamiento a outliers realizado en otras encuestas de la Institución como *ESI* y *ENUT*.

La siguiente tarea consta de la identificación y registro del resultado del procedimiento de revisión anterior. Así, los datos pueden alcanzar los siguientes estados

- a) **Dato válido:** datos que cumplen con los criterios de calidad establecidos en el marco de la operación estadística, de acuerdo con el manual correspondiente al procedimiento del subproceso Revisar y validar.
- b) **Dato con observaciones:** datos que no cumplen con los criterios de revisión (datos atípicos, inconsistencia, fuera de rango y omisiones, entre otros). El resultado se registrará en el reporte de errores, a modo de solicitar la recuperación de información a los responsables del proceso de recolección.
- c) **Dato rechazado:** datos que no cumplen con el mínimo de calidad para realizar un procedimiento de recuperación.

Como resultado de la actividad, se tendrá una base de datos revisada al nivel del micro dato, junto con la identificación de valores observados, rechazados y validados, así como un reporte de errores.

4.1.5 Solicitar corroboración de microdatos

A partir de los datos revisados a nivel de microdatos, junto con el reporte de los errores observados que requieren de corroboración, precisión o recuperación de información, así como el manual de procedimiento de Ejecutar, supervisar y finalizar la recolección, se procederá a corroborar la información revisada en la actividad anterior, con el objetivo de recuperar información.

Para ello se debe, primero, identificar los datos observados que necesitan de corroboración para, posteriormente, revisar el reporte de errores, asegurándose de que este contenga todos los casos que necesitan una recuperación o precisión de información en el subproceso de Ejecutar, supervisar y finalizar la recolección. Asimismo, es necesario verificar que los casos observados registren la observación o consulta con todos los campos completos, a modo de identificar la trazabilidad del caso, así como el procedimiento que se siguió para la validación

del dato. Para la recuperación de información, se debe enviar un correo electrónico al jefe encargado de la Subdirección de Operaciones (SDO), en que se solicite la corroboración de los datos identificados, adjuntando el reporte de errores y los plazos de entrega para el procedimiento de recuperación de información. También cabe la posibilidad de que las solicitudes de corroboración de datos se haga al interior de la Subdirección Técnica (SDT), sin necesidad de interactuar con la SDO.

Como resultado de la actividad se obtendrá el reporte con los casos que requieren corroboración, junto con un correo electrónico en que se solicita a la SDT o la SDO la corroboración de los datos.

4.1.6 Recuperar información

La siguiente actividad se realiza a partir de la solicitud de corroboración de datos, el reporte de valores detectados, el manual del subproceso Ejecutar, supervisar y finalizar la corrección, y la nueva solicitud de corroboración de información en caso de que haya respuestas rechazadas. El objetivo es realizar el proceso de recuperación de información para la validación de los datos, lo cual contempla las siguientes tareas.

En caso de que la responsabilidad sea del SDO, se debe, primero, concurrir a la dirección asociada a el o la informante o, en su defecto, volver a contactarle a través de un llamado telefónico, para luego completar la información faltante y/o corregir o corroborar la información que se detectó como inconsistente o fuera de rango, detallando las justificaciones necesarias para su uso por parte del equipo encargado del piloto de la VII EME.

Si la responsabilidad recae en el SDT, se deben ejecutar los procedimientos de corroboración y precisión de la información según lo establecido en el proceso de Diseño y planificación (**ESPECIFICAR**).

Como resultado se obtendrá un reporte de valores observados con respuestas obtenidas en el subproceso Ejecutar, supervisar y finalizar la recolección.

4.1.7 Evaluar respuesta

En base al reporte de valores observados obtenido en la actividad anterior, se procederá a evaluar las respuestas obtenidas. Para ello, se debe revisar la respuesta otorgada en la actividad Recuperar información. Posteriormente, se ha de verificar que todos los casos observados y enviados para recuperar información determinada cuenten con un estado de respuesta adecuado, es decir, que sean catalogados como dato recuperado o dato no recuperado, acompañando una observación que justifique tal categorización.

Luego, se han de revisar y evaluar tales respuestas, a modo de determinar decisiones respecto de cómo proceder frente a los datos. Puede haber dos escenarios distintos. En primer lugar, la recuperación puede catalogarse como **válida o satisfactoria** en aquellos casos donde los datos recuperados son validados e incorporados en la base de datos con estado validado. Ello puede darse cuando el dato original es correcto, por lo cual se mantiene, justificando tal decisión en el reporte; o cuando el dato original es incorrecto, por lo cual se

registra la respuesta y se deriva al subproceso **5.4 Editar e imputar**. En segundo lugar, la recuperación puede determinar **no válida o insatisfactoria** en aquellos casos en que la respuesta otorgada señala que no se logró la recuperación, o que esta no cumple con los criterios de calidad necesarios. En tales casos, el dato se registra como **inválido**, recibiendo un tratamiento automático en R dependiendo de los procedimientos de edición e imputación relativos a su tipo de dato.

Posteriormente, si la recuperación fue validada, los datos que lo requieran deben ser neviados hacia el subproceso **5.4 Editar e imputar**. Si, por el contrario, la respuesta de recuperación no es satisfactoria al no responder a la consultar ni registrar una justificación de la información recolectada, se debe volver a enviar la solicitud al subproceso *Ejecutar, supervisar y finalizar la recolección*. Es fundamental definir los plazos en que se requiere la información recuperada, a modo de cumplir con la planificación determinada. Por último, el resultado de la evaluación para cada respuesta ha de registrarse en un documento localizado en el repositorio de GitHub eme-ine/procesamiento.

Como resultado de esta actividad se obtendrá un reporte de los valores observados detectados con su respectiva validación y categorización como dato recuperado o no recuperado, realizada por la SDO. Además, se elaborará una nueva solicitud para corroborar la información, e caso de necesitarse. Por último, se generará la solicitud hacia el subproceso 5.4 Editar e imputar para los datos que lo requieran.

RECORDAR revisión valores atípicos post-creación de ponderadores

Considerar validación x encuestador/a

4.1.8 Revisar datos agrupados

Esta actividad parte de la base de datos revisada a nivel de microdatos, el reporte de valores observados, y el manual de procesamiento con las especificaciones para los procedimientos de revisión y validación de datos. Es importante considerar que este proceso contempla la expansión de datos a nivel poblacional, a partir de los ponderadores generados en el subproceso **5.6 Calcular ponderadores**.

Esta actividad busca realizar una primera revisión y validación de los datos a nivel agregado, a modo de asegurar los criterios de calidad correspondientes. La primera tarea es volver a revisar los outliers según lo especificado en la actividad Identificar valores atípicos, incorporando los criterios de rangos de aceptación, la revisión de variables comunes en términos de variación, incidencias, valores atípicos y anómalos, y variaciones atípicas. Luego, se debe identificar a todos los microdatos que presenten un comportamiento atípico para, por último, elaborar un reporte que constate los valores atípicos detectados.

Como producto de la actividad se obtendrá una base de datos revisada a nivel de grupo, en la cual se identifican valores atípicos que, a su vez, serán reportados. También se elaborará un reporte para consultas Corroboración de microdatos.

4.1.9 Revisar base de datos

La entrada de la actividad es la base de datos revisada con identificación de valores atípicos; el reporte de valores atípicos detectados, y el reporte de respuestas validadas en la actividad Evaluar respuesta.

El objetivo es asegurar el cumplimiento de los criterios de calidad estipulados en los lineamientos de la operación estadística (**INCORPORAR CUALES SON ESOS CRITERIOS**). Para lograrlo, el responsable SDT ha de verificar que los procedimientos de revisar y validar se realizaron de manera adecuada, verificando el cumplimiento de los siguientes tareas:

- a) Revisar que la base de datos de la VII EME es aquella recibida en la actividad Revisar insumos
- b) Verificar que las revisiones fueron realizadas de manera correcta, registrando las respuestas del procedimiento de corroboración, precisión y recuperación de datos.
- c) Si se identificasen valores atípicos en el procedimiento de validación, se debe volver a realizar la actividad Revisar a nivel de microdatos, informando a través de un correo electrónico al funcionario responsable de tal revisión sobre el error ejecutado.

Como resultado se obtendrá una base de datos revisada, validada y almacenada en el repositorio de GitHub eme-ine/procesamiento.

4.1.10 Disponer base

A partir de los datos verificados, se procederá a compartir la base de datos validada para iniciar el siguiente subproceso **5.4 Editar e imputar**. Para ello, se ha de formalizar la disposición de la base, incorporándola al repositorio de GitHub eme-ine/procesamiento en los plazos estipulados. Posteriormente, se informará por correo electrónico el cierre del subproceso, para iniciar el siguiente.

Como resultado, se obtendrá una base de datos revisada, validada y almacenada en el repositorio de GitHub eme-ine/procesamiento, junto con un reporte de errores que explique el procedimiento aplicado, y un reporte de recuperación validado.

4.2 Indicadores

Tasa de errores reales (p. 21)

Busca identificar los datos incorrectos (errores reales) en la etapa de procesamiento, incluyendo entradas faltantes, no válidas o inconsistentes, o tipos de datos que se han corroborado como errores. la fórmula de calculo es la siguiente:

$$\text{Tasa de errores reales} = \frac{N_{\text{de errores reales}}}{N_{\text{de observaciones efectivas}}} * 100$$

5 Edición e imputación

El subproceso de edición e imputación consta de dos etapas: Revisión, y Editar e imputar datos. Este subproceso sólo será aplicado a las bases de datos FULL y de Empleo.

5.1 Revisión

5.1.1 Revisar insumos

Datos + reglas y documentación de validaciones

5.1.2 Identificar inconsistencias adicionales

tanto a base full como a base empleo

Casos fuera de marco revisar si se encuestó a casos no elegibles (no independientes, no residentes habituales de viviendas seleccionadas, o empleadores de unidad de más de 10 trabajadores).

Inconsistencias Horas trabajadas Revisar horas de trabajo > 112 hrs, o respuesta parcial no sabe (888). En esos casos, considerar info de la ENE.

Periodo de funcionamiento del negocio Revisar valores nulos. Revisar valores en preguntas c7 y c7a, y plantear estrategia de edición

Remuneraciones Gastos y remuneraciones No presentar gastos en remuneraciones, pese a declarar trabajadores remunerados Remuneracion de socios Socios que no reportan monto bruto total pagado.

Impuestos Boleta de honorarios e IVA Declaracion de IVA e inicio de actividades como trabajador independiente con boleta a honorarios

Propiedad de herramientas y tipo de inscripción Informantes independientes o no inscritos en el SII con herramientas a nombre del negocio.

Gastos e impuestos a la renta Informantes que indicaron inicio de actividades como trab. indepe. con boleta de honorarios, en un tramo imponible, pero sin gastos por impuestos, permisos o patentes.

Gastos e IVA informantes que señalan declarar IVA habiendo iniciado actividades antes del 2021, sin indicar gastos por pago de impuestos, permisos o patentes

Inscripción en SII e impuesto a la renta Informante declara estar inscrito en SII, comenzando actividades antes de 2018, estando en un tramo imponible y sin declarar impuesto a la renta.

Herramientas Valorización de equipos de otros El informante no puede valorizar equipos que no son de su propiedad

Lugar de trabajo y herramientas Informantes que trabajan en vehículo propio pagado, pero no declaran esa herramienta en módulo h (ni en la pregunta de medios de transporte ni

de maquinaria específica). Otra inconsistencia es declarar trabajar en vehículo arrendado o prestado, sin declarar tal herramienta en modulo h.

Socios y trabajadores Socios Informantes que respondían tener socios trabajadores que no residen en su vivienda, pero de quienes tampoco indican ganancias. Revisar si hay socios o no en la grilla f2.

Ganancias socios vs del negocio informantes con socios no residentes del hogar donde la ganancia declarada no era la del negocio, sino el porcentaje de utilidades que corresponden al informante en su calidad de socio. Esto es identificable cuando las ganancias del negocio son iguales o menores a la de los socios.

Ajustes F1 y F3 en relación con información de f2 Informante declara contar con trabajadores en f2, pero respondió no tener trabajadores empleados o contratados el mes pasado en f1. También puede ocurrir a la inversa.

Incorporación y eliminación de trabajadores en base empleo Revisar autodeclaración de informante como trabajador, la consideración de trabajadores externos (como servicios profesionales), la no declaración de socios trabajadores del hogar, entre otras.

Cuotas Problemas de flujo por no solicitud de crédito

Cuotas por pagar Número de cuotas declaradas supera el monto posible por tipos de préstamos, lo cual tiende a suceder por declarar créditos hipotecarios

Formas de pago Revisar casos en que los informantes no declaran aceptar ninguna forma de pago

Uso de internet informantes declaran usar internet, pero no declaran ningún uso en i3.

Depuración estimación de gastos, ingresos y ganancias (54) Corrección en diversos casos, depurando la estimación de ganancias en razón de los datos en gastos e ingresos.

5.1.3 Elaborar estrategias de edición e imputación

Una vez identificadas las variables que requieren un tratamiento especial en lo que respecta a edición e imputación, es necesario elaborar una estrategia de edición e imputación para hacerse cargo de los problemas detectados. Ello debe quedar documentado en un archivo que especifique los procedimientos a seguir para cada una de las variables. Tal documento debe ser depositado en el repositorio eme-ine/procesamiento de GitHub, para luego ser revisado por parte del equipo de procesamiento, solucionando los problemas o ausencias constatadas en la estrategia de edición e imputación.

5.2 Editar e imputar datos

5.2.1 Implementar estrategias de edición e imputación

Una vez revisadas las estrategias de edición e imputación elaboradas, ha de proceder a implementarlas. Ello se realizará en R, siguiendo las pautas estipuladas en el documento.

5.2.2 Revisión y validación de ediciones e imputaciones

Una vez implementadas las ediciones e imputaciones planificadas, el equipo de procesamiento debe revisar los datos, verificando que todos los casos identificados como inconsistentes en la primera etapa del subproceso de Editar e imputar hayan sido solucionados según las pautas planteadas.

5.2.3 Disponer base de datos editada e imputada

Una vez se haya validado la implementación de las ediciones e imputaciones planificadas, y habiendo asegurado de que cada uno de los casos con inconsistencias y falta de información hubiesen sido tratados según lo planificado en el documento de estrategia de edición e imputación, se debe proceder a disponer la base de datos editada e imputada en la carpeta input/data del repositorio eme-ine/procesamiento de GitHub.

6 Derivación de nuevas variables y unidades

7 Cálculo de ponderadores

8 Cálculo de agregados

9 Finalización de los archivos de datos

10 Bibliografía