BANCO DE ESPAÑA
Eurosistema

BELab
BANCO DE ESPAÑA
Eurosistema

Dirección General de Economía y Estadística

08/03/2022

Documentación Operativa del BELab

Guía para el control del output

BELab. Banco de España
Departamento de Estadística

ÍNDICE

1 Introducción

Este documento forma parte de los manuales operativos de funcionamiento del BELab y establece una serie de normas, recomendaciones y mejores prácticas para asegurar la difusión segura de los resultados de los trabajos realizados por los investigadores externos con los microdatos del BELab.

El Laboratorio de datos del Banco de España (BELab), ofrece acceso a microdatos de alta calidad del Banco de España. Los datos proporcionados al BELab se tratan de forma responsable y respetando todas las normas legales e internas sobre el tratamiento y divulgación de datos.

Los investigadores externos que trabajan con los microdatos del Banco de España son responsables de garantizar que los resultados de sus cálculos no violen los requisitos de confidencialidad de los datos. Cuando los resultados de sus investigaciones se extraen por los procedimientos previstos del entorno seguro del BELab, deben asegurarse de que los resultados del cálculo obtenidos no contengan ningún dato que pueda ser rastreado hasta unidades de observación individuales o cualquiera de las unidades estadísticas que conforman el catálogo del BELab (empresas individuales, grupos consolidados, etc.), facilitando la información relevante suficiente para que el personal del BELab pueda hacer la verificación final, denominada “control del output”

El cumplimiento de estos requisitos de confidencialidad de los datos a extraer, es comprobado adicionalmente por el personal del BELab en el llamado proceso de Control del Output. Si no se cumplen estos requisitos, los resultados de los cálculos no se podrán extraer ni, por lo tanto, publicar.

Este documento pretende ayudar a los investigadores visitantes a cumplir más fácilmente los requisitos estipulados para la extracción de sus resultados, extracción que se realiza por el personal del BELab y se pone a disposición de los investigadores fuera ya del sistema controlado, y su uso en las publicaciones basadas en los mismos. Una mayor implicación de los investigadores en el proceso redundará en una mayor agilidad en la revisión de los datos y un mejor aprovechamiento de los recursos del BELab. Así mismo, este proceso sirve para aumentar la confianza de los productores y proveedores de los datos, que ven cómo su información se utiliza de forma segura y responsable. Dependiendo de la base de datos que se utilice, el control del output podría realizarse también por los proveedores de los datos, mediante una tercera capa de protección, por tanto (del investigador, del personal del BELab y de los proveedores de los datos)

El BELab se reserva el derecho de modificar, complementar o ampliar los siguientes principios y reglas durante el proceso de Control del Output, si así lo estimara necesario.

Los apartados de este documento contienen los principios y normas para el Control del Output y las normas relativas a las publicaciones.

2 Principios para el control del output

2.1 Principios de anonimización

Las siguientes normas tienen como objetivo facilitar el cumplimiento de la regulación relativa a la confidencialidad de los datos. Todos los resultados que se pretenda extraer del entorno seguro del BELab deben estar completamente anonimizados. Esto quiere decir que ningún agente individual puede ser identificado directa o indirectamente (por deducción), teniendo en cuenta las posibles formas en las que terceros usuarios pudieran reidentificar los microdatos. Los investigadores externos son responsables en todo momento de garantizar que sus resultados cumplen con los criterios que aseguran una anonimización completa. Dichos criterios son los siguientes:

1. **No extracción de identificadores:** Los identificadores no pueden estar incluidos ni en los resultados a extraer ni en los códigos que los generan.

2. **No extracción de microdatos:** Ningún resultado puede contener un microdato. Esto implica la no extracción de subconjuntos de datos, ni de tablas, gráficos, códigos o log files que contengan microdatos en sí mismos. En consecuencia, no está permitida la extracción de mínimos y máximos, ni de aquellos cuantiles que correspondan exactamente con un microdato o no cumpla con el criterio número 7.

3. **Número mínimo de observaciones:** Todos los resultados que se vayan a extraer deben estar basados en al menos 3 observaciones diferentes. Esto aplica tanto a resultados agregados (medias, medianas, etc.) como a gráficos y tablas (mínimo 3 observaciones por celda/ nodo de información). La forma más sencilla de demostrar el cumplimiento de esta norma es generar siempre la tabla de frecuencias asociada a cada resultado.

4. **Grados de libertad:** Los modelos de regresión deben estar calculados como mínimo con 10 observaciones, y sus grados de libertad deben ser también como mínimo 10.

5. **Los grados de libertad se calculan como:** Número de observaciones – número de parámetros estimados (variables) - otras restricciones del modelo.

6. **Regla de dominancia (p%):** Para garantizar que no es posible identificar ninguna entidad, incluso cuando se está cumpliendo con el número mínimo de 3 observaciones, es necesario asegurar que la observación más grande no supera el 85% del peso total del valor analizado, o de cualquier otra ponderación que se use. Así se evita la identificación indirecta de las observaciones con mayor peso.

   Ejemplo: Para el cálculo las ventas totales en un determinado sector en un determinado año tenemos solo 3 empresas. El volumen total es de 100 millones de euros, cuya composición es la siguiente: 90 millones de la mayor empresa, y 5 de cada una de las otras. En este caso la empresa más grande es potencialmente identificable por su contribución al valor total.

7. **Confidencialidad en tablas múltiples, control de diferencias:** Si se calculan los resultados sobre una población G, pero posteriormente se recalculan para un subconjunto X de G, las normas explicadas arriba deben cumplirse para las observaciones de la diferencia. En otro caso, las observaciones individuales podrían identificarse en base a la diferenciación.

   Ejemplo: Tenemos una tabla con todas las empresas de un determinado sector, y otra con las empresas de ese sector que superan un volumen X de ventas. Tendríamos que crear una tercera tabla con las empresas que no llegan a ese volumen X y comprobar que se cumplen los criterios de confidencialidad en ella, porque en otro caso podrían identificarse sus empresas por diferenciación.

8. **Variables categóricas dicotómicas de 0-1 (dummies):** Si se calculan las medias de estas variables, debe haber al menos 3 observaciones con cada categoría (3 observaciones con 0 y 3 con 1).

9. **Tratamiento de ceros y valores missing:** Los ceros están permitidos en regresiones y análisis estadístico descriptivo siempre que no representen valores missing en variables dicotómicas y categóricas. En estadísticos descriptivos, los valores missing no se tendrán en cuenta para determinar el número de observaciones diferentes empleadas. En caso de realizar imputaciones a los valores faltantes, deberán reportarse el número de observaciones imputadas y observadas.

2.2 Principio de verificabilidad

El proceso de Control del Output implica un tiempo y esfuerzos considerables por parte del Equipo del BELab. Para optimizar la utilización de los recursos del Laboratorio de Datos y minimizar el tiempo de espera desde que solicitan la extracción de los datos, los investigadores externos deben cumplir las siguientes normas:

1. **Master File:** Se debe crear un archivo maestro que contenga toda la información relevante sobre el proyecto de investigación y que llame a todos los subprogramas utilizados.

2. **Log file:** La función de log (*log file*) debe activarse para cada código de programa. El registro debe comenzar antes de la descripción del contenido del proyecto de investigación y antes de la primera línea de código de cálculo.

3. **Orden y estructura dentro del código:** El código debe estar estructurado de forma visualmente clara, de modo que los bloques individuales del código (cabecera, etapas analíticas individuales, etc.) se distingan visualmente. Los bucles deben tener sangrías. Los programas largos o los pasos analíticos deben dividirse en archivos de código más pequeños, por ejemplo, (“0_master.do”, “1_data_preparation.do”, “2_descriptive_analysis.do” y “3_regressions.do”)

4. **Comentarios en los códigos de los programas:** El código del programa debe incluir suficientes comentarios para que, incluso las personas que no estén familiarizadas con el proyecto, sean capaces de entenderlo en un tiempo razonable.

5. **Nombres claros de los archivos de salida**: Los nombres de todos los archivos de salida deben comenzar con el mismo nombre que el programa utilizado para generar el archivo y deben estar numerados de manera lógica.

6. **Nombres claros de las variables:** Todos los nombres asignados deben ser lo más informativos posible y utilizarse de forma coherente. Las etiquetas de las variables y las descripciones breves de las mismas deben proporcionarse para todos los datos generados por el propio usuario, así como para todos los datos de origen externo. Si se crean o modifican variables (categóricas), es necesario asignar las correspondientes etiquetas de valor a estos valores.

7. **Especificación del cumplimiento de las normas de anonimización:** El investigador debe incluir código que justifique el cumplimiento de los requisitos de anonimización descritos más arriba. Así, por ejemplo, proveerá de tablas de frecuencias, descripción de los modelos, o cualquier otro elemento que muestre el cumplimiento de las normas del output que solicita extraer.

8. **Re-evaluación de output:** En caso de solicitar la extracción de un código ya revisado previamente, pero sobre el que se han realizado pequeños cambios, estos deben quedar específicamente reflejados en la nueva solicitud. Siempre que sea posible, los investigadores deberán presentar a evaluación únicamente los elementos del programa que han modificado.

2.3 Principio de reproducibilidad

Para verificar el cumplimiento de los requisitos de anonimización especificados en el principio 3, todos los resultados de los cálculos presentados para su revisión y extracción deben ser reproducibles. Los investigadores deben cumplir las siguientes normas:

1. **Reproducibilidad de los códigos de los programas:** Todos los resultados de los cálculos deben ser generados sin problemas por un programa denominado "0_master.do" que pueda ejecutarse sin errores y que debe contener todos los programas de análisis utilizados a lo largo del proyecto. Este programa debe comenzar por la carga de los datos originales proporcionados por el BELab. El programa debe ejecutar siempre los mismos pasos y producir exactamente los mismos resultados que los ya presentados para su revisión. En el archivo "0_master.do", cada llamada de ejecución a otro sub-programa debe ir seguida de una breve descripción del contenido de este.

2. **Reproducibilidad del software:** Todo software informático utilizado para generar los resultados de los cálculos debe ser claramente descrito al principio del archivo "0_master.do" (nombre y número de versión). Junto con el número de versión del software de análisis, se incluirán también los nombres de todos los paquetes (por ejemplo, R, Python, Octave) utilizados.

3. **Reproducibilidad de los datos:** Todos los conjuntos de datos del BELab utilizados para generar los resultados de los cálculos deben ser claramente descritos al principio del archivo "0_master.do" (DOI si está disponible, variables y año). Cualquier conjunto de datos utilizado que proceda de proveedores de datos externos debe describirse también en el fichero "0_master.do".

4. **Reproducibilidad del output a publicar:** Todos los resultados de los cálculos que se quieran extraer deben poder encontrarse de forma rápida, fácil y clara en los resultados producidos por los programas de análisis. Para ello, se debe relacionar de manera clara el código utilizado para generar los resultados susceptibles de ser extraídos y publicados. Para ello, el BELab recomienda crear un documento archivo "master_yyymmdd_publication.do" que contenga únicamente los últimos pasos para generar los cálculos, tablas y gráficos que compongan el output que se solicita extraer.

2.4 Principio de uso razonable de los recursos

Por regla general, los elementos que se solicite extraer del BELab serán únicamente para ser directamente utilizados en una publicación. Por esta razón, los investigadores visitantes deben guiarse por el principio de utilizar los recursos de forma razonable, especialmente a la hora de decidir qué resultados solicitan extraer. El número de elementos a presentar deberá estar en consonancia con lo que normalmente se espera en el ámbito de un artículo científico empírico.

En general, los investigadores visitantes deben tener en cuenta las siguientes normas:

1. **El análisis exploratorio de los datos no puede formar parte del output a extraer.** Sólo se pueden presentar para su revisión los análisis que puedan ser directamente publicados. La tarea de seleccionar los resultados dignos de ser llevados a la publicación forma parte de los trabajos a realizar por el investigador dentro del BELab.

2. **Número máximo de líneas en el output.** El BELab no establece, a priori, un número máximo de líneas de código a revisar durante el proceso de Control del Output, pero se reserva la posibilidad de hacerlo si los investigadores no utilizan los recursos del Laboratorio de manera razonable, que implica ser prudente en la cantidad de output solicitado, usos de los programas, tiempos de ejecución, uso de las sesiones, etc.

2.5 Principio de responsabilidad

Los investigadores visitantes son responsables de garantizar el cumplimiento de todos los principios y normas establecidos en este documento. El incumplimiento de las normas dará lugar a que el BELab se niegue a entregarle los resultados de los cálculos. Los investigadores deben respetar las siguientes normas:

1. **Comprobación de todos los resultados de los cálculos para su publicación:** Antes de que los investigadores soliciten al BELab la revisión de un output que

desean extraer, ellos mismos deben comprobar que han aplicado los principios de Control de Output. Una vez revisado, avisarán al equipo del BELab para que proceda a revisar sus datos. Colocarán el output solicitado, en la carpeta /Out/Output de su proyecto e indicarán, tal y como se detalla en esta guía, de los elementos necesarios para realizar su revisión.

2. **Funcionamiento del código:** Si el código del programa contiene errores de sintaxis o de otro tipo, éstos serán dejados sin corregir por el BELab y se pedirá a los investigadores que corrijan su código.

3. **Formato de control de output:** Los resultados y los códigos de los programas sólo se aceptarán para el control de salida si son editables y se presentan como archivos de texto sin formato o .csv. Los gráficos deben tener un formato de sólo lectura (estáticos) y presentarse en formato .jpeg o .png.

3 Principios de control de publicaciones

Las siguientes normas tienen por objeto ayudar a los investigadores a cumplir más fácilmente las normas de control de las publicaciones ("control de las publicaciones").

1. **Revisión de todas las publicaciones por parte del BELab:** Ninguna información relativa a los proyectos de investigación desarrollados en el BELab puede publicarse hasta que no se haya revisado y autorizado. La autorización puede ser retenida si los resultados que se pretenden publicar no cumplen con los criterios del BELab.

2. **Copia de los trabajos:** Es responsabilidad de los investigadores entregar una copia de los trabajos publicados que preparen y que contengan resultados de investigación de los análisis realizados durante su estancia en el BELab. El incumplimiento de esta norma inhabilita para el uso futuro del BELab hasta que se faciliten los trabajos publicados como fruto de investigaciones previas realizadas en BELab.

3. **Citación de la fuente:** el investigador se compromete a mencionar la fuente última de los datos en cualquier publicación que resulte de este estudio según se indique en la respectiva guía de cada base de datos.

4. **Citación de gráficos y tablas:** Todos los gráficos y tablas deben aparecer citados de la siguiente manera:" Fuente: BELab. Laboratorio de Datos del Banco de España, <nombre del conjunto de microdatos utilizado del catálogo del BELab (en su caso con la abreviatura común)>, <período durante el cual se utilizaron los microdatos>, cálculos propios."

5. **Declaración del tipo de acceso a los datos:** Cada publicación debe especificar el tipo de acceso a los datos que tuvo el investigador, es decir, Acceso Presencial desde una Dataroom (indicar si fue Madrid o Barcelona), acceso remoto o mixto.

6. **Especificación de los conjuntos de datos utilizados, uso del DOI:** Todos los conjuntos de datos utilizados en el proyecto de investigación deben citarse indicando el nombre y, si está disponible, el DOI.