

Unidad Gestión de la Calidad

Asunto: Informe sobre la protección de los microdatos de la Encuesta de Actividades de Innovación en Empresas 2016-2018 del INE-ANII.

Fecha: 2020-08-26

Descripción de la actividad: analizar el riesgo de revelar información confidencial del archivo de microdatos de la Encuesta de Actividades de Innovación en Empresas 2016-2018. Se trata de una encuesta cuya muestra ha sido seleccionada a partir del Directorio de Empresas y Establecimiento 2018 del INE.

Escenario de divulgación: archivo de microdatos en formato SPSS a ser publicado en el sitio web del INE y la ANII. Los usuarios malintencionados podrían disponer de otras fuentes de información para intentar hacer matching con el archivo de microdatos y de esta forma re-identificar a la empresa informante. El archivo a entregar contiene todas las variables del cuestionario de la encuesta. El usuario podría utilizar la combinación de variables clave, e identificar la empresa informante, existe cierta probabilidad de ocurrencia.

Variables clave: a continuación se listan las variables clave que podrían determinar combinaciones de celdas únicas o “raras” y que serían utilizadas por el usuario para hacer el matching con otro archivo que contenga variables identificatorias: Departamento, CIUCodigo (Código de clase de actividad), departamento (Departamento), PTO_4.1.2016 (Total de personal ocupado año 2016), PTO_4.1.2017 (Total de personal ocupado año 2017), PTO_4.1.2018 (Total de personal ocupado año 2018), DE_5.1_1_2016 (Ingresos por venta de bienes o servicios 2016), DE_5.1_1_2017 (Ingresos por venta de bienes o servicios 2017), DE_5.1_1_2018 (Ingresos por venta de bienes o servicios 2018), DE_5.1_1_2016 (Total ingresos 2016), DE_5.1_Tot_2017 (Total ingresos 2017), DE_5.1_Tot_2018 (Total ingresos 2018).

De acuerdo con en el diseño muestral se ha considerado a los efectos de estimar el riesgo de re-identificación, la variable “peso.cs” para la ponderación de los casos.

El escenario de divulgación y las variables clave disponibles para el usuario han sido analizados y se llegó a la conclusión que dichas variables son las que tendrían la más alta probabilidad de ser utilizadas para intentar re-identificar a los informantes.

El análisis de riesgo de re-identificación y por tanto de divulgar información confidencial se basó en la combinación de las variables clave que se han establecido, tomando en cuenta los ponderadores de la muestra. Se utilizó el paquete de “R” sdcMicro para determinar las frecuencias de observaciones en riesgo de ser re-identificadas que arrojó los resultados que se indican a continuación.

Unidad Gestión de la Calidad

Resultados del análisis de riesgo:

Se utilizó el archivo de microdatos sin las siguientes variables identificatorias: RUT, NroINE, RazSocial. Las variables IB_3.6.loc1, IB_3.6.loc2, AI_B.2_Esp, FAI_D.1_6_Esp, FAI_D.2_2_7_Esp, FAI_D.2_2_11_Esp, FAI_D.3_Otros, FAI_D.5, RAI_E.2, FI_F.1_Esp, VS_G.1_1_Agen, VS_G.1_2_Agen, VS_G.1_3_Agen, VS_G.2_2_Otros, FOD_H.1_Otros, FOD_H.2_1_1, FOD_H.2_2_1, FOD_H.2_3_1, FOD_H.2_4_1, OPT_J.Obs se eliminaron del archivo de datos anonimizado ya que el texto de las mismas posibilitaban la identificación de empresas.

Se realizó el análisis para cada uno de los 3 años por separado, tomando las siguientes combinaciones de variables clave (CIUCodigo, departamento, PTO_4.1.2016, DE_5.1_1_2016, DE_5.1_Tot_2016), (CIUCodigo, departamento, PTO_4.1.2017, DE_5.1_1_2017, DE_5.1_Tot_2017), (CIUCodigo, departamento, PTO_4.1.2018, DE_5.1_1_2018, DE_5.1_Tot_2018).

El reporte del package sdcMicro dio como resultado que en el archivo de microdatos original hay 752 casos que son únicos. El riesgo de re-identificación en la población, tomando en cuenta el ponderador, no es demasiado significativo. Además, no se publican otras variables que se han utilizado para determinar los ponderadores, por lo tanto se ha minimizado el riesgo de re-identificación utilizando dichas variables.

Técnicas de protección aplicadas:

Se recodificó la variable CIUCodigo a 2 dígitos con el nombre CIU2digitos.

Se eliminó la variable de identificación geográfica “departamento” ya que con la inclusión de la misma en la base permite la identificación de casos únicos a nivel de la combinación CIU2digitos - Departamento.

En las ramas de actividad que por contener un número muy pequeño de casos daban lugar a la posible identificación de la empresa se procedió a la unión de las mismas y recodificación que se indica a continuación:

Variable CIU2digitos original	Variable CIU2digitos recodificada
13	C1
12	
19	C2
20	
36	E1
38	

Unidad Gestión de la Calidad

Se realizó micro agregación para las variables DE_5.1_Tot_2016, DE_5.1_Tot_2017, DE_5.1_Tot_2018, DE_5.1_1_2016, DE_5.1_1_2017, DE_5.1_1_2018

Resultados luego de la protección:

En el archivo de datos original el riesgo medido en base a las re-identificaciones esperadas es de 414,84 (13.93%), luego de la aplicación de las técnicas de protección antes mencionadas, en el archivo de datos anonimizado se logró reducir la cantidad de “casos únicos” a 0 y el número de re-identificaciones esperadas a 21,94 (0.74%), lo cual no es significativo.

Evaluación de la utilidad de los microdatos anonimizados:

El diseño muestral y por tanto los ponderadores no han sido estimados tomando en cuenta el departamento donde se ubica la empresa. Por lo cual los indicadores que se generen a partir de los microdatos tendrán la precisión.

Conclusiones:

Siempre existe cierto riesgo de re-identificación de los informantes, pero se ha considerado que luego de aplicadas las técnicas de “anonimización”, el riesgo residual no es significativo y por lo tanto es baja la probabilidad de re-identificación.