Estudio empírico de estandarización de reglas de validación en la producción estadística oficial

TRABAJO FIN DE GRADO

Curso 2019/2020

CONVOCATORIA: Junio

UNIVERSIDAD COMPLUTENSE
MADRID

FACULTAD DE CIENCIAS MATEMÁTICAS

DOBLE GRADO EN ECONOMÍA – MATEMÁTICAS Y ESTADÍSTICA

Nombre del estudiante:

Carlos Álvarez Aparicio

Nombre de los tutores

David Salgado Fernández

Elena Rosa Pérez

Madrid, 10 de julio de 2020

ÍNDICE

1. RESUMEN

La fase de depuración de datos en la producción estadística oficial requiere hasta un 40% de los recursos dedicados a la producción de una encuesta (U.S. Federal Committee on Statistical Methodology, 1990). Un elemento clave en esta fase es la construcción y contrastación de reglas de validación que permiten detectar los errores ajenos al muestreo para su posterior tratamiento (T. de Waal, J. Pannekoek et all, 2011). Existen estándares de producción internacionales cuya implementación se está llevando a cabo paulatinamente en las diversas oficinas de estadísticas nacionales e internacionales (Conference of European Statisticians. Geneva, 14-16 June, 2011), (UNECE. Generic Statistical Business Process Model v5.0. UNECE, 2013), (UNECE. Generic Statistical Information Model v1.1. UNECE, 2013), (UNECE. Common Statistical Production Arquitecture v1.5. UNECE, 2013), (UNECE. Generic Activity Model for Statistical Organizations v1.0. UNECE, 2015), (UNECE. Generic Statistical Data Editing Models v1.0. UNECE, 2015). En este trabajo se ofrece un estudio empírico aplicado a operaciones estadísticas del Plan Estadístico Nacional[1] sujetas también a reglamentación europea conjugando la metodología estadística más reciente con herramientas en R desarrolladas en el Sistema Estadístico Europeo[2] (Statistic Software, 2019).

El objetivo principal de este trabajo es poder demostrar empíricamente que diferentes encuestas elaboradas por el Instituto Nacional de Estadística[3] (INE) pueden seguir un mismo proceso de validación. Este proceso se construirá usando la herramienta R, y en concreto el paquete *validate* (van der Loo M. and de Jonge E., 2019).

1 https://ine.es/ss/Satellite?c=Page&cid=1254735995577&pagename=INE%2FINELayout&L=0

2 https://ine.es/ss/Satellite?L=es_ES&c=Page&cid=1254735905268&p=1254735905268&pagename=INE%2FINELayout

3 https://ine.es/index.htm

2. INTRODUCCIÓN

La validación de datos es un proceso que asegura la entrega de datos limpios y claros a los programas, aplicaciones y servicios que lo utilizan. Comprueba la integridad y validez de los datos que se están introduciendo en diferentes softwares y sus componentes. La validación de los datos garantiza que los datos enviados a las aplicaciones conectadas sean completos, precisos, seguros y consistentes. Esto se logra a través de controles de validación de datos y reglas que rutinariamente comprueban la validez de los datos (Lebied.M., 2018).

Desde un punto de vista muy general, según cómo sean los medios de recogida de la información se pueden definir, fundamentalmente, dos tipos de validaciones de datos:

- Manual: la recogida de datos se da a través del cumplimiento de unas preguntas previamente definidas, que darán lugar a las variables. Estas se pueden dividir en tres tipos a su vez, que son, cuestionarios, análisis de escritos y entrevistas. Son encuestas que cumplimentan personas, por lo que el error más frecuente será de tipo humano. (W. Gilley. Jerry, 1990)
- Automáticos: la recopilación de datos es recogida por un mecanismo que se dedica específicamente a ello y estos procesos tienen la ventaja que se pueden programar los propios automatismos para que no recojan datos erróneos. (Lethbridge, T.C., Sim, S.E et al., 2005)

Desde el punto de vista de la producción estadística oficial, se distinguen diversos modos de depuración (T. de Waal, J. Pannekoek et all, 2011), siendo la depuración interactiva o manual la que acarrea un mayor coste de producción. En particular, en este trabajo nos centraremos en la formulación y aplicación de reglas de depuración sobre los datos (van der Loo M. and de Jonge E., 2018).

Trabajaremos empíricamente sobre dos encuestas claramente diferenciadas, la Operacón Estadística, Indicadores de Actividad del Sector Servicios[4] (IASS, a partir de ahora) (INE

[4]https://ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176863&menu=ultiDatos&idp=1254735576778

IASS) y la Encuesta Nacional de Salud de España[5][6] (ENSE, a partir de ahora) (INE ENSE), la cual a su vez contiene información tanto a nivel individual< como a nivel de hogar. Ambas encuestas presentan diversos métodos de recogida (CAPI, CAWI, PAPI, fundamentalmente), pero en todos los casos los datos son sometidos a diversas reglas de validación a lo largo de la estrategia de depuración de cada operación estadística. En particular, los controles a aplicar pertenecerán a algunos de los siguientes tipos de control (Técnicas de Validación de Datos):

Cuadro 1. Tipos de control.

Método de Validación	Descripción
Requerimiento	Comprueba que el usuario ha contestado a la cuestión evitando que accidentalmente lo deje vacío
Formato	Asegura que los datos siguen un patrón establecido
Coherencia	Comprueba que dos o más preguntas que guardan relación no se contradicen.
Rango	Comprueba que los datos se encuentran entre un valor aceptable superior e inferior, dentro de un cierto rango
Longitud	Comprueba que la cantidad de caracteres cumple con las expectativas.
Cuadro Desplegable	Asegura que el usuario solo puede elegir una opción predefinida de una lista, reduciendo las posibilidades de errores ortográficos o respuestas no deseadas

[5] Esta encuesta será dividida en dos partes, la encuesta a nivel adulto y a nivel hogar, no se trata la parte de menores.

[6] https://ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176783&menu=ultiDatos&idp=1254735573175

3. MATERIALES Y MÉTODOS

3.1 MATERIALES

Para poder realizar empíricamente las validaciones pertinentes a cada una de las dos encuestas, se necesitan, las preguntas de los cuestionarios correspondientes (donde se especifiquen las preguntas y el tipo de respuesta que se espera tener en dichas cuestiones) y la muestra recogida.

3.1.1 ENCUESTAS

IASS: los Indicadores de Actividad del Sector Servicios[7] miden la evolución mensual de la actividad de las empresas cuya actividad económica son los Servicios de mercado no financieros a través de dos variables: la cifra de negocios y el personal ocupado. La cifra de negocios comprende los importes facturados por la empresa por la prestación de servicios y la venta de bienes. El personal ocupado incluye tanto el personal asalariado como el no remunerado.

Para su obtención se realiza una encuesta continua que investiga todos los meses más de 28.000 empresas que operan en este sector. Los resultados se presentan en forma de índices con el objetivo de medir variaciones respecto del año base 2015. (INE IASS).

Se debe tener en cuenta, que, debido a la ley de protección de datos, los microdatos de encuestas económicas no son accesibles, por lo que los datos que hemos empleado han sido simulados en función a los valores reales y que se emplean en algunas asignaturas del máster EMOS. Por tanto, no son datos de empresas reales.

A continuación, se detallan en profundidad las características principales de cada variable:

- ID: identificador de empresa. Está anonimizado, pero asigna a cada empresa un valor numérico diferente al resto.
- b1: cifra de negocios. Es la repuesta que otorga el encargado de rellenar la encuesta a la pregunta ‘A. Ingresos (valorado sin incluir IVA). Importe de la cifra de negocios: (euros)’.
- emp: número total de empleados.
- c121: remunerados fijos. Número de empleados que poseen un contrato indefinido con la empresa.

[7] https://ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176863&menu=metodologia&idp=1254735576778

- c122: remunerados eventuales. Número de empleados que poseen un contrato temporal, sin pertenecer a una Empresa de Trabajo Temporal (ETT).
- c11: personal no remunerado. Número de personas que trabajan en nombre de la empresa, pero no reciben contraparte monetaria por esta labor.
- división: clasificación económica de la empresa de acuerdo con la Clasificación Nacional de Actividades Económicas (CNAE-2009)[8].
- rama: clasificación interna dentro del sector en el que opera la empresa.
- factor: peso de la empresa dentro del sector asignada externamente a la encuesta.
- existencias: cantidad de existencia acumuladas de la empresa.
- exist: *flag* que indica si la empresa posee existencias acumuladas o no.
- mes: fecha de la extracción de los datos. Tiene le formato 'MM'+MMYYYY.

ENSE: la Encuesta Nacional de Salud de España es una investigación dirigida a la población que reside en viviendas familiares cuya finalidad principal es obtener datos sobre el estado de salud y los factores determinantes del mismo desde la perspectiva de los ciudadanos.

En la encuesta realizada durante 2016-2017 se investigaron aproximadamente 37.500 viviendas distribuidas en 2.500 secciones censales. Su periodicidad es quinquenal, alternándose cada dos años y medio con la Encuesta Europea de salud, con la que comparte un grupo de variables armonizadas.

Se debe tener en cuenta, que, debido a la ley de protección de datos, los microdatos de encuestas de salud están anonimizados, por lo que los datos que hemos empleado son reales y han sido publicados de por el INE de forma que ya estaban anonimizados.

Para conocer en detalle las variables que estas encuestas tratan se debe acudir a la sección *3.1.2 DATOS* para una visión general de estos o si se quiere conocer en profundidad se debe acudir a los *ANEXOS IV* y *VIII*.

3.1.2 DATOS

Los microdatos para el caso de IASS empleados en este trabajo han sido generados artificialmente empleando la misma estructura de metadatos de la encuesta real para un conjunto muy reducido de 50 unidades. Este conjunto de microdatos se emplea en algunas asignaturas del máster EMOS para estudiar esta fase de la producción.

Los resultados no deben tomarse como una inferencia real sobre las encuestas, ya que al modificar las variables se puede incurrir en fallos que los datos originales no poseen, incluso puede llegar a haber fallos que han sido generados aposta para realizar comprobaciones.

[8] https://ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177032&menu=ultiDatos&idp=1254735976614

Los microdatos para el caso de ENSE se han tomado junto con su diseño de registro de la página web del INE, donde se están disponibles de modo anonimizado en ficheros con formato fwf (INE ENSE).

Las características de los datos se muestran a continuación para cada encuesta:

IASS: La encuesta IASS tiene 50 unidades estadísticas para la fecha enero de 2018, pero para realizar los controles de dicha encuesta se ha tenido que generar también un año de simulaciones mes a mes, donde cada mes se tienen las correspondientes mismas 50 unidades estadísticas.

Nombre	Descripción	Carácterísticas
ID	Identificador de cliente	Numérica
b1	Cifra de negocios	Numérica
emp	Número total de empleados	Numérica
c121	Número de empleados remunerados fijos	Numérica
c122	Número de empleados remunerados eventuales	Numérica
c11	Número de empleados no remunerados	Numérica
division	Identificador de la actividad económica	Carácter
rama	Clasificación dentro del sector	Carácter
factor	Peso de la empresa en el sector	Numérica
existencias	Cantidad de existencias	Numérica
exist	Acumulación o no de existencias	Lógica
mes	Fecha del	Carácter

ENSE (adulto): La encuesta ENSE para adultos tratada posee 23089 observaciones para un gran número de variables, que se agrupan por grupos de significado y relación entre las preguntas. Se puede observar que todas las variables presentan variable de tipo carácter, a pesar de tener significados numéricos o lógicos entre sus posibles respuestas[9].

Nombre	Descripción	Carácterísticas
CCAA	Código numérico de Comunidad Autónoma	Carácter
IDENTHOGAR	Variable identificadora del hogar	Carácter
A7_2a	Número de orden del adulto seleccionado.	Carácter
SEXOa	Variable binaria de sexo	Carácter
EDADa	Variable numérica de edad	Carácter
ACTIVa	Variable binaria de actividad	Carácter

[9] Para mayor información sobre las variables de esta encuesta acudir a *ANEXO IV*

PROXY_0, PROXY_1, PROXY_2, PROXY_2b, PROXY_3b, PROXY_4, PROXY_5	Especificaciones sobre la persona que informa	Carácter
E1_1, E2_1a, E2_1b, E2_1c, E2_1d, E3, E4, E4b	Características demográficas sobre la persona adulta seleccionada	Carácter
NIVEST	Nivel de estudios del adulto seleccionado	Carácter
F6, F7, F8_2, F9_2, F10, F11, F12, F13, F14a, F14b, F15, F16, F17, F18a_2, F18b_2, F19b_2, F20	Relación entre el adulto y la actividad económica desempeñada.	Carácter
G21, G22, G23, G24, G25a_1, G25b_1, G25c_1, G25a_2, G25b_2, G25c_2, G25a_3, G25b_3, G25c_3, G25a_4, G25b_4, G25c_4, G25a_5, G25b_5, G25c_5, G25a_6, G25b_6, G25c_6, G25a_7, G25b_7, G25c_7, G25a_8, G25b_8, G25c_8, G25a_9, G25b_9, G25c_9, G25a_10, G25b_10, G25c_10, G25a_11, G25b_11, G25c_11, G25a_12, G25b_12, G25c_12, G25a_13, G25b_13, G25c_13, G25a_14, G25b_14, G25c_14, G25a_15, G25b_15, G25c_15, G25a_16, G25b_16, G25c_16, G25a_17, G25b_17, G25c_17, G25a_18, G25b_18, G25c_18, G25a_19, G25b_19, G25c_19, G25a_20, G25b_20, G25c_20, G25a_21, G25b_21, G25c_21, G25a_22, G25b_22, G25c_22, G25a_23, G25b_23, G25c_23, G25a_24, G25b_24, G25c_24, G25a_25, G25b_25, G25c_25, G25a_26, G25b_26, G25c_26, G25a_27, G25b_27, G25c_27, G25a_28, G25b_28, G25c_28, G25a_29, G25b_29, G25c_29, G25a_30, G25b_30, G25c_30, G25a_31, G25b_31, G25c_31, G25a_32, G25b_32, G25c_32	Estado de salud	Carácter
H26_1, H26_2, H26_3, H27	Accidentalidad	Carácter
I28_1, I28_2, I29_1, I29_2	Restricción de la actividad	Carácter
K32, K33, K34, K35, K36, K37, K38, K38a	Limitaciones físicas y sensoriales	Carácter
L39_1, L39_2, L39_3, L39_4, L39_5, L40, L41, L42_1, L42_2, L42_3, L42_4, L42_5, L42_6, L42_7, L43, L44, L45, L46,	Limitaciones para la realización de las actividades de la vida cotidiana	Carácter
M47_1, M47_2, M47_3, M47_4, M47_5, M47_6, M47_7, M47_8, M47_9, M47_10, M47_11, M47_12, M47a, M47b	Salud mental	Carácter

N48, N49, N50, N51, N52, N53, N54, N55_1, N55_2, N55_3, N56_1, N56_2, N56_3, N57, N58_1, N58_2, N58_3, N59, N60_1, N60_2, N60_3, N60_4, N60a_1, N60a_2, N60a_3, N60a_4, N61_1, N61_2, N61_3, N61_4, N61_5, N62, N62b, N63_1, N63_2, N63_3, N63_4, N63_5, N63_6, N63_7, N63_8, N63_9, N63_10, N64, N65_1, N65_2, N65_3, N65_4, N65_5, N65_6, N65_7, N65_8,	Consultas médicas y otros servicios ambulatorios	Carácter
O66, O67, O69, O70, O71, O72, O73, O74, O75, O76, O77, O78, O79, O80_1, O80_2, O80_3, O81_1, O81_2, O81_3, O82_1, O82_2, O83, O84_1, O84_2, O84_3, O84_4, O84_5, O84_6, O84_7, O84_8, O84_9,	Hospitalizaciones, urgencias y seguro sanitario	Carácter
P85, P86, P87_1a, P87_1b, P87_2a, P87_2b, P87_3a, P87_3b, P87_4a, P87_4b, P87_5a, P87_5b, P87_6a, P87_6b, P87_7a, P87_7b, P87_8a, P87_8b, P87_9a, P87_9b, P87_10a, P87_10b, P87_11a, P87_11b, P87_12a, P87_12b, P87_13a, P87_13b, P87_14a, P87_14b, P87_15a, P87_15b, P87_16a, P87_16b, P87_17a, P87_17b, P87_18a, P87_18b, P87_19a, P87_19b, P87_20a, P87_20b, P87_21a, P87_21b, P87_22a, P87_22b, P87_23a, P87_23b,	Consumo de medicamentos	Carácter
Q88, Q89, Q90,Q91, Q92, Q93, Q94, Q95, Q96, Q97, Q98, Q99, Q100, Q101, Q102, Q103, Q104, Q105	Prácticas preventivas	Carácter
R106, R107, R108_1, R_108_2, R108_3, R108_4,	Necesidades de atención médicas no cubiertas	Carácter
S109, S110	Características físicas	Carácter
T111, T112, T113, T114_1, T114_2, T115, T116_1, T116_2, T117, T118_1, T118_2, T119_1, T119_2,	Actividad física	Carácter
U120_1, U120_1a, U120_2, U120_3,U120_4, U120_5, U120_6, U120_7, U120_7a, U120_8, U120_9, U120_10, U120_11, U120_12, U120_13, U120_14, U120_15,U120_15a, U120FZ, U120CANTFZ, U2_120F	Alimentación e higiene dental	Carácter
V121, V122, V123, V124, V125, V126,	Consumo de tabaco y exposición al humo	Carácter
W127, W128Cer, W128Cer_1, W128Cer_2, W128Cer_3, W128Cer_4, W128Cer_5, W128Cer_6, W128Cer_7, W128Vin, W128Vin_1, W128Vin_2,	Consumo de alcohol	Carácter

W128Vin_3, W128Vin_4, W128Vin_5, W128Vin_6, W128Vin_7, W128Vermut, W128Vermut_1, W128Cer_2, W128Cer_3, W128Vermut_4, W128Vermut_5, W128Vermut_6, W128Vermut_7, W1Lic28Lic, W128Lic_1, W128Lic_2, W128Lic_3, W128Lic_4, W128Lic_5, W128Lic_6, W128Lic_7, , W128Comb, W128Comb_1, W128Comb_2, W128Comb_3, W128Comb_4, W128Comb_5, W128Comb_6, W128Comb_7, , W128Sidra, W128Sidra_1, W128Sidra_2, W128Sidra_3, W128Sidra_4, W128Sidra_5, W128Sidra_6, W128Sidra_7, W129		
X130_1, X130_2, X130_3, X130_4, X130_5, X130_6, X130_7, X130_8, X130_9, X130_10, X130_11	Apoyo afectivo y personal	Carácter
Y133, Y134, Y135	Cuidados a otras personas con problemas de salud	Carácter
FACTORADULTO	Factor de elevación del adulto	Carácter
CLASE_PR	Clase social basada en la ocupación de la persona de referencia	Carácter
IMCa	Índice de masa corporal del adulto	Carácter

ENSE (Hogar): La encuesta ENSE (hogar) tratada posee 60143 unidades estadísticas, las cuales se corresponden con hogares diferentes. Las variables que compartan significado y características en la respuesta se agrupan[10].

Nombre	Descripción	Carácterísticas
CCAA	Código numérico de Comunidad Autónoma	Carácter
IDENTHOGAR	Variable identificadora del hogar	Carácter
ESTRATO	Variable que indica un rango del número de habitantes que tiene el municipio del hogar	Carácter
SEXO_i	Variable binaria de sexo	Carácter
EDAD_i	Variable numérica de edad	Numérico
NORDEN_Ai, NORDEN_Mi	Orden de adultos y menores de edad en el hogar	Carácter
NADULTOS, NMENORES	Número de adultos y menores de edad en el hogar	Carácter
NORDEN_Pref	Número de orden de la persona de referencia	Carácter

[10] Para mayor información sobre las variables de esta encuesta acudir a *ANEXO VIII*

A7_1_i, A7_2a, A7_2m	Persona indicada: de referencia, adulto seleccionado y menor seleccionado.	Numérico
A8_1_i, A8_2_i	Adulto seleccionado o relación con él.	Numérico
NORINF	Número de orden del informante dentro del hogar.	Carácter
A9_otra	Relación con el hogar del informante del Cuestionario del Hogar.	Carácter
A10_i	Nivel de estudios.	Carácter
A11_i	Situación laboral.	Carácter
A12	Tipo de hogar en función de las personas que habitan.	Carácter
B13	Pensión contributiva. Persona de referencia.	Carácter
B14	Trabajo anterior. Persona de referencia.	Carácter
B15_2	Actividad de la empresa que genera la pensión contributiva.	Carácter
B16_2	Ocupación, profesión u oficio de la persona que genera la pensión.	Carácter
B17	Situación profesional en la ocupación que desempeñó.	Carácter
B18	Ha trabajado antes. Persona de referencia.	Carácter
B19a_2, B19b_2	Actividad de la empresa donde trabajó o trabaja. Persona de referencia.	Carácter
B20a_2, B20b_2	Ocupación, profesión u oficio donde trabaja o trabajó. Persona de referencia.	Carácter
B21a, B21b	Situación profesional del último trabajo. Persona de referencia.	Carácter
C22	Número de dormitorios del hogar.	Numérico
C23	Número de metros cuadrados del hogar.	Numérico
C24_1, C24_2, C24_3, C24_4, C24_5, C24_6, C24_7, C24_8, C24_9	Problemas con el entorno de la vivienda.	Carácter
D26_1, D26_2, D26_3, D26_4, D26_5, D26_6, D26_7, D26_8, D26_9, D26_10, D26_11	Ingresos monetarios de los miembros del hogar.	Carácter
D27	Principal tipo de ingreso.	Carácter
D29	Ingreso mensual neto.	Carácter
FACTORHOGAR	Factor de elevación del hogar.	Numérico
CLASE_PR	Clase social derivada de la ocupación. Persona de referencia.	Carácter

3.1.3 CONTROLES

Los controles de las diferentes encuestas en producción se definen habitualmente por expertos en la materia que conocen en profundidad el significado de las variables, su posible interrelación y los rangos de valores permitidos.

En este caso, para la encuesta del IASS nos hemos inspirado en algunos controles de depuración empleados en el máster EMOS para esta fase de la producción, proporcionando aquí una formulación minuciosa en lenguaje .YAML.

Para las encuestas del ENSE los controles han sido creados a partir del significado de las variables y las posibles contradicciones a las que se puede llegar. Los controles se proporcionan aquí también en lenguaje .YAML. han sido diseñados en un fichero EXCEL que se encuentra en los *ANEXOS V* y *IX*.

Adviértase que existirán muchas discrepancias con los controles oficiales empleados en la producción estadística oficial de estas encuestas, pero en todo caso la formulación y aplicación de los controles, independientemente de su contenido semántico, será igual tanto en estos dos ejemplos como en la producción estadística oficial.

3.2 MÉTODOS

Para realizar la validación además de necesitar conocer las variables proporcionadas por las encuestas, los datos y la creación de los controles; se requiere unificar la herramienta informática con la que se aplican los controles a los datos de las diferentes encuestas. Para realizar estas tareas, en consonancia con la tendencia en muchas oficinas del Sistema Estadístico Europeo, se utilizará el entorno de computación estadística R. En concreto, desarrollaremos un *script con una estructura común para los tres conjuntos de datos*. Este *script* hará uso central del paquete *validate [REF]* y un fichero .YAML, donde se incluyen los *edits (*reglas de validación*)* minuciosamente descritos.

3.2.1 SCRIPT

3.2.1.1 PARTE COMUN

Los *scripts* se desarrollan en archivos de texto plano en lenguaje R que permiten ser leídos por esta herramienta y ejecutados por la consola que la propia aplicación lleva incorporada. (R Development Core Team, 2000).

Se elaboran tres *scripts* diferentes, uno para cada conjunto de datos**, la composición de los tres es fundamentalmente la misma**, con unas ligeras diferencias y adaptaciones propias para cada base de datos, que se detallarán tras explicar las características comunes a los tres.

La parte común del *script* y cuyo objetivo es que sea extrapolable a cualquier encuesta es la siguiente:

- En primer lugar, se deben cargar los paquetes predeterminados de R, que serán necesarios en la interpretación y el tratamiento de los datos y los *edits*[11]:

```
# Load packages
devtools::install_github('david-salgado/fastReadfwf')
library(data.table)
library(validate)
library(ggplot2)
library(stringr)
library(dplyr)
```

1. *FastReadfwf*: Paquete que se ha descargar e instalar de la librería de paquetes llamada *github* que se encuentra virtualmente en la red. Este paquete es encargado de leer ficheros de datos con formato fwf en poco tiempo (R documentation FastReadFWF) a partir de un diseño de registro estándar. Este paquete está guardado en la siguiente ruta virtual: **https://github.com/david-salgado/fastReadfwf**
2. *data.table*: Paquete útil para agregación rápida de datos grandes (por ejemplo, 100 GB en RAM), uniones ordenadas rápidas, adición / modificación / eliminación rápida de columnas por grupo sin ninguna copia, listas de columnas, lectura / escritura amigable y rápida de valores separados por caracteres. Ofrece una sintaxis natural y flexible para un desarrollo más rápido (R documentation data.table).
3. *validate*[12]*:* Este paquete permite declarar reglas de validación de datos e indicadores de calidad de datos; confrontar datos con ellos y analizar o visualizar los resultados. El paquete admite reglas que son por campo, en registro, registro cruzado o conjunto de datos cruzados. Las reglas se pueden analizar automáticamente para el tipo de regla y la conectividad (R documentation validate).
4. *ggplot2*: Un sistema para crear gráficos 'declarativamente', basado en "La gramática de los gráficos". Se proporcionan los datos, se asignan variables a la estética, se indican qué primitivas gráficas usar y el software se ocupa de los detalles (R documentation ggplot2).
5. *stringr*: Un conjunto consistente, simple y fácil de usar de *wrappers* alrededor del paquete *stringi*. Todos los nombres de funciones y argumentos (y posiciones) son consistentes, todas las funciones tratan con "NA" y los vectores de longitud cero de la misma manera, y la salida de una función es fácil de alimentar a la entrada de otra (R documentation stringr).

5.1.stringi: Procesamiento de cadena / texto de caracteres rápido, correcto, consistente, portátil y conveniente en cada configuración regional y cualquier

[11] No todos los paquetes son necesarios para los tres *Scripts,* y es posible que para otra encuesta sean necesario o útiles otros paquetes no considerados en el presente trabajo.
[12] Este paquete es fundamental para el propósito de este trabajo y será desarrollado en profundidad en el apartado 3.2.3 VALIDATE.

codificación nativa. Debido al uso de la biblioteca 'ICU' (Componentes internacionales para Unicode), el paquete proporciona a los usuarios R funciones independientes de la plataforma conocidas por *Java*, *Perl*, *Python*, *PHP* y *Ruby*. Las características disponibles incluyen: búsqueda de patrones, generación de cadenas aleatorias, mapeo de casos, transliteración de cadenas, concatenación, normalización Unicode, formato y análisis de fecha y hora y muchos más (R documentation stringi).

6. *dplyr*: Una herramienta rápida y consistente para trabajar con marcos de datos como objetos, tanto en memoria como sin memoria (R documentation dplyr).

- Se define la librería donde se guarda toda la información relativa a cada encuesta[13] (código a modo de ejemplo de una de las tres encuestas):

```
# Change for user's path
path <- 'C:\\Users\\X541\\Documents\\TFG_MATES\\EMPRESAS\\Duda
1'
```

- Se cargan las bases de datos. Primero se cargan los datos y se guardan bajo el nombre *dataFile_hogar* (por ejemplo), después se carga la descripción de las variables y por último se obtienen los datos en formato de *data table* (*dt*), lo que permite que los datos sean leídos y tratados posteriormente por la herramienta. Para esta tarea se usa la función *fread_fwf* del paquete *FastReadfwf*, previamente definido. (Código a modo de ejemplo de una de las encuestas ENSE):

```
# Load data sets
dataFile_hogar <- file.path(path, 'MICRODAT.CH.txt')
schemaFileName <- file.path(path, 'stENSE2017Hogar_Schema.xlsx')
stSchema_hogar <- fastReadfwf::xlsxToSchema(schemaFileName,
sheetname = 'stSchema', lang = 'en')
data_hogar.dt <- fastReadfwf::fread_fwf(dataFile_hogar,
stSchema_hogar, outFormat = 'data.table')
```

La función *readRDS* se usa en el caso de la encuesta IASS, al haber sido los microdatos creados directamente en R y almacenados en este formato:

```
# Load data sets
dataHistory <- readRDS(file.path(path, 'dataHistory.rds'))
```

- Se cargan los controles, que previamente han sido creados en formato .YAML.[14] El método empleado usa la función *validator* del paquete *validate*. La forma de cargarlos es la siguiente (código a modo de ejemplo de una de las encuestas):

```
# Load edit rules from .YAML file
validator <- validator(.file = file.path(path,
'EditsIASS_corr..YAML'))
```

[13] Para este trabajo, se han creado en local, para poder ser ejecutado hay que adaptarlo a la ruta local donde se guarden los archivos.

[14] Se explican las cualidades y las diferencias entre cada fichero de *Edits* en el apartado 3.2.2 EDITS

donde *path* es el fichero de cargado y guardado de tablas definido anteriormente y 'EditHogIndv.YAML' es el nombre del fichero donde están creados los *Edits*.

- Se realiza la validación. Para este paso se cruzan los datos cargados en la *data table* llamada anteriormente, con los controles del archivo que contiene los *Edits*. Se utiliza la función *confront* del paquete *Validate*. La forma de cruzarlos es la siguiente (código a modo de ejemplo de una de las encuestas):

```
# Apply rules to data set
cf <- confront(MM022018, validator, ref = FL_MM022020.dt)
```

3.2.1.2 PARTE ESPECÍFICA

El *script* de la encuesta del IASS posee características particulares, las cuales se detallan a continuación:

- Para la encuesta IASS se ha de modificar el formato en el que se dan los datos, y es que esta base de datos requiere información histórica y se ha de unificar en una sola tabla creando una variable llamada 'mes' que indicará el mes de referencia de los datos:

```
dataHistory_MM022018.dt <- rbindlist(
  lapply(names(dataHistory), function(fecha){

  DT <- dataHistory[[fecha]][
    , mes := fecha]
  return(DT)
})
)
```

- Al tener los controles definidos y una base histórica consistente, se requieren hacer controles sobre el *dataset* verificando que el valor esté entre unos límites. Para esta tarea el paquete *Validate* ofrece la posibilidad de que cuando se ejecute la confrontación (*confront*) se pueda introducir otra tabla con la que poder hacer dichas comparaciones. El código es el siguiente:

```
FL_MM022020.dt <- readRDS(file.path(path,
'dataMM022020_Intervalos.dt'))[
  , b1_anterior := dataHistory_MM022018.dt[mes == 'MM012018',
b1]]

# Apply rules to data set
cf <- confront(MM022018, validator, ref = FL_MM022020.dt)
```

donde *FL_MM022020.dt* es la tabla con la que se va a comparar definida previamente.

3.2.2 EDITS

Para establecer opciones y metadatos para las reglas, se utiliza el conocido formato .YAML. (.YAML ain't markup language). .YAML es una forma legible para definir estructuras (anidadas). La estructura es la siguiente:

```
rules:
-
  expr:
  name:
  label:
  description:
```

donde *expr* define la regla, *name* le otorga un nombre a la regla, *label* es una breve explicación de lo que hace la regla y *description* es la definición extendida de la regla.

Se debe tener en cuenta que para escribir los Edits en formato .YAML se deben cumplir una serie de requisitos (YAML):

- La definición de un conjunto de reglas con metadatos es directa.
- Después de cada guion empieza una regla.
- La sangría es importante. La sangría debe ser de un espacio, y si se necesita cambiar de línea, se usa '|' ó '>' y se agrega una espacio más.
- Cuando una regla comience con '!' se debe poner entre comillas, ya que el símbolo de exclamación tiene un significado especial en .YAML.

Cada archivo de *Edits* tiene controles que son característicos de la propia encuesta, es por ello que serán tratados como independientes:

- IASS: Los controles de esta encuesta están perfectamente definidos por expertos en la materia. En este caso la creación de los *Edits* consiste en que su programación sea adecuada[15].
- ENSE: Los controles de estas dos encuestas han sido creados mediante asunciones lógicas, lo cual implica que algunas reglas pueden ser innecesarias y/o insuficientes, no obstante, para el propósito del presente trabajo serán suficientes[16].

3.2.3 VALIDATE

El paquete *validate*, el cual es fundamental para desempeñar la tarea que este trabajo quiere llevar a cabo, es útil para declarar reglas de validación e indicadores de calidad;

[15] Para consultar el fichero donde se crean los *edits* de la encuesta IASS consultar *ANEXO II*.
[16] Para consultar el fichero donde se crean los e*dits* de las encuestas ENSE adulto y ENSE hogar consultar *ANEXO VI* y *ANEXO X*, respectivamente; y para su definición recurrir a *ANEXO V* y *ANEXO IX*.

confrontar datos con ellos y analizar o visualizar los resultados. El paquete admite reglas que son por campo, en registro, en registro cruzado o en conjunto de datos cruzados. Las reglas se pueden analizar automáticamente por tipo y registro (R documentation validate).

Como la carga de los *edits* y la confrontación de los datos con las reglas ya han sido comentados, se procede a explicar la metodología que ha sido usada para mostrar resultados:

- En primer lugar, se crea una matriz que marque en binario para todos los registros y para todos los controles si la regla es pasada o fallida. El código común para todas las encuestas se muestra a continuación:

```
# Flag por unidad y edit
flag.matrix <- values(cf)
```

- Se cuenta el número de registros que pasan y que fallan en cada regla, también el número de registros que están desinformados. El código común para todas las encuestas se muestra a continuación:

```
npass <- colSums(flag.matrix)
nfail <- dim(flag.matrix)[1] - npass
nNA <- apply(flag.matrix, 1, function(x){sum(is.na(x))})
```

- Se relativiza en función del total, para ver el impacto que tiene el error sobre la muestra extraída. El código común para todas las encuestas se muestra a continuación:

```
rel.pass <- npass / dim(flag.matrix)[1]
rel.fail <- nfail / dim(flag.matrix)[1]
rel.NA <- nNA / dim(flag.matrix)[1]
```

- Se aúna toda la información anterior en una única matriz. El código común para todas las encuestas se muestra a continuación:
`relPass_edit.df <- as.data.frame(aggregate(cf))`

- Se representa gráficamente (a través de un histograma) el número de registros que satisface cada *edit*, el número de registros que fallan y el número de variables desinformadas que posee cada *edit* para que sea fácilmente interpretable. El código común para todas las encuestas se muestra a continuación:

```
ggplot(relPass_edit.df, aes(x = editNames, y = rel.pass)) +
  geom_col() +
  labs(x = 'Edit', y = '', title = 'Tasa de unidades que
satisface cada edit') +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90),
        plot.title = element_text(hjust = 0.5))
```

```
ggplot(relPass_edit.df, aes(x = editNames, y = rel.fail)) +
  geom_col() +
  labs(x = 'Edit', y = '', title = 'Tasa de unidades fallidas de
cada edit') +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90),
        plot.title = element_text(hjust = 0.5))

ggplot(relPass_edit.df, aes(x = editNames, y = rel.NA)) +
  geom_col() +
  labs(x = 'Edit', y = '', title = 'Tasa de unidades
desinformadas de cada edit') +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90),
        plot.title = element_text(hjust = 0.5))
```

donde 'relPass_edit.df' denota los datos de entrada a representar, 'aes(x= , y=)' indica los datos que se representarán en el eje de ordenadas y en el eje de abscisas; 'geom_col()' indica el tipo de gráfico que se requiere siendo el que posee este nombre el histograma, 'labs(x=, y=, title)' indican el nombre del eje horizontal, del vertical y el título del gráfico; y por último, el formato de los títulos y de los elementos de los ejes 'theme(axis.text.x = element_text(angle =), plot.title = element_text(hjust=))'.

- Se pasa la información por unidad estadística en vez de por *edit*, es decir, se mostrará la tasa de acierto que tiene cada registro sobre el total de los *edits*. El código común para todas las encuestas se muestra a continuación:

```
# La misma informaciÃ³n por unidad estadística:
relPass_record.df <- as.data.frame(aggregate(cf, by = 'record'))
(relPass_record.df$ID <- str_pad(as.character(1:50), 2, 'left',
'0'))

ggplot(relPass_record.df, aes(x = ID, y = rel.pass)) +
  geom_col() +
  labs(x = 'Unidad', y = '', title = 'Tasa de edits satisfechos
por cada unidad') +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90),
        plot.title = element_text(hjust = 0.5))

ggplot(relPass_record.df, aes(x = ID, y = rel.fail)) +
  geom_col() +
  labs(x = 'Unidad', y = '', title = 'Tasa de edits fallidos por
cada registro') +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90),
        plot.title = element_text(hjust = 0.5))

ggplot(relPass_record.df, aes(x = ID, y = rel.NA)) +
```

```
  geom_col() +
  labs(x = 'Unidad', y = '', title = 'Tasa de edits desinformado
s por cada registro') +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90),
        plot.title = element_text(hjust = 0.5))
```

4. RESULTADOS

En este apartado se procede a exponer y analizar las respuestas que otorga el programa previamente indicado en el apartado *validate* para cada encuesta. No se debe olvidar que los datos en la encuesta IASS son simulaciones, y no hay que tener en cuenta el resultado en sí, pues puede que la incidencia detectada sea un error generado por la simulación.

4.1 IASS

- *Flag* por unidad y *edit.* Esta parte es muy importante porque es la matriz que indica para cada observación si el *edit* ha sido superado con éxito o no, y es la tabla base de donde parten todos los resultados. Se exponen únicamente 5 unidades del registro a modo de ejemplo, y se omiten las restantes:

```
# Flag por unidad y edit
flag.matrix <- values(cf)


##       Req3 Req4 Req5 Req6 Req7 Rango5 Rango6 Rango7 Rango8 Ra
ngo14 Nulo1 Nulo2
##  [1,] TRUE TRUE TRUE TRUE TRUE   TRUE   TRUE   TRUE   TRUE
TRUE  TRUE  TRUE
##  [2,] TRUE TRUE TRUE TRUE TRUE   TRUE   TRUE   TRUE   TRUE
TRUE  TRUE  TRUE
##  [3,] TRUE TRUE TRUE TRUE TRUE   TRUE   TRUE   TRUE   TRUE
TRUE  TRUE  TRUE
##  [4,] TRUE TRUE TRUE TRUE TRUE   TRUE   TRUE   TRUE   TRUE
TRUE  TRUE  TRUE
##  [5,] TRUE TRUE TRUE TRUE TRUE   TRUE   TRUE   TRUE   TRUE
TRUE  TRUE  TRUE

##       CNAnt Emp_W_1 EmpNoRem_W_1 EmpRemFij_W_1 EmpRemEve_W_1
lCN_W_1 lCN_W_3
##  [1,]  TRUE    TRUE         TRUE          TRUE          TRUE
TRUE    TRUE
##  [2,]  TRUE    TRUE         TRUE          TRUE          TRUE
TRUE    TRUE
##  [3,]  TRUE    TRUE         TRUE          TRUE          TRUE
TRUE    TRUE
##  [4,]  TRUE    TRUE         TRUE          TRUE          TRUE
TRUE   FALSE
##  [5,]  TRUE    TRUE         TRUE          TRUE          TRUE
TRUE    TRUE


##       lCN_noW_1 lrCNPR_W_1 lrCNPR_W_3 lrCNPR_noW_3 lEX_W_1 lE
X_W_3 lEX_noW_3
##  [1,]      TRUE       TRUE       TRUE         TRUE    TRUE
```

```
TRUE      TRUE
##  [2,]      TRUE      FALSE       TRUE         TRUE    TRUE
TRUE      TRUE
##  [3,]      TRUE       TRUE       TRUE         TRUE    TRUE
TRUE      TRUE
##  [4,]     FALSE       TRUE      FALSE         TRUE    TRUE
TRUE      TRUE
##  [5,]      TRUE       TRUE       TRUE         TRUE    TRUE
TRUE      TRUE
```

- Número de unidades satisfechas por *edit:*

```
npass <- colSums(flag.matrix)


##          Req3          Req4          Req5          Req6
Req7
##            50            50            50            50
50
##        Rango5        Rango6        Rango7        Rango8
Rango14
##            50            50            50            50
50
##         Nulo1         Nulo2         CNAnt       Emp_W_1  EmpN
oRem_W_1
##            50            50            49            50
50
## EmpRemFij_W_1 EmpRemEve_W_1       lCN_W_1       lCN_W_3     l
CN_noW_1
##            49            50            35            40
38
##    lrCNPR_W_1    lrCNPR_W_3  lrCNPR_noW_3       lEX_W_1
lEX_W_3
##            34            46            47            47
47
##     lEX_noW_3
##            42
```

- Número de unidades fallidas por e*dit:*

```
nfail <- dim(flag.matrix)[1] - npass


##          Req3          Req4          Req5          Req6
Req7
##             0             0             0             0
0
##        Rango5        Rango6        Rango7        Rango8
Rango14
##             0             0             0             0
0
##         Nulo1         Nulo2         CNAnt       Emp_W_1  EmpN
oRem_W_1
```

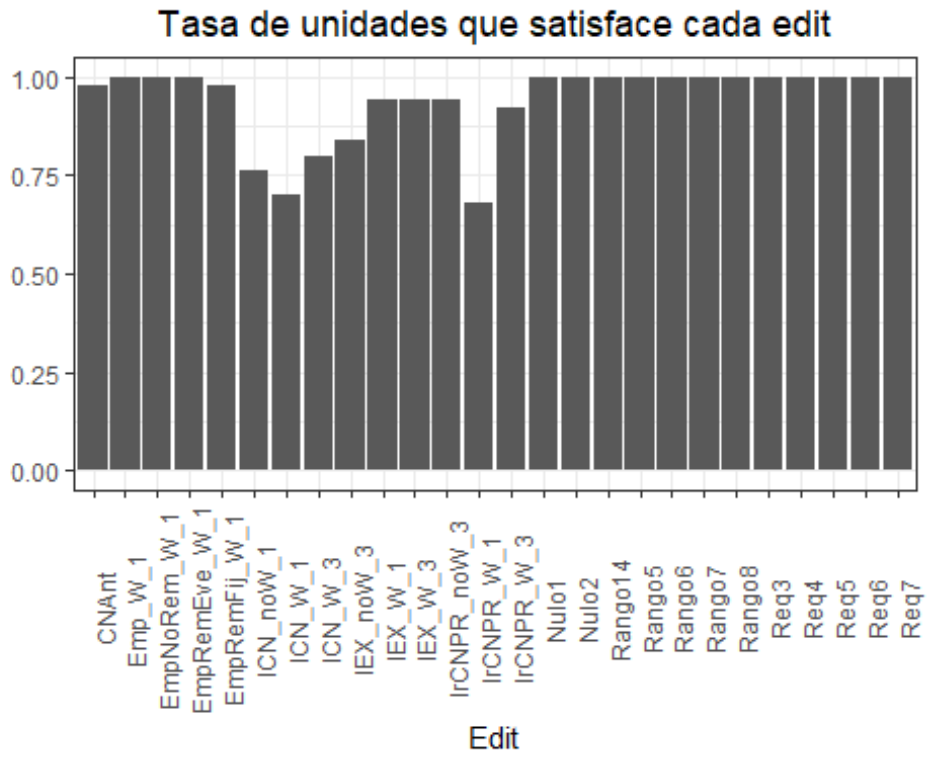
```
##             0             0             1             0
0
## EmpRemFij_W_1 EmpRemEve_W_1       lCN_W_1       lCN_W_3     l
CN_noW_1
##             1             0            15            10
12
##    lrCNPR_W_1    lrCNPR_W_3  lrCNPR_noW_3       lEX_W_1
lEX_W_3
##            16             4             3             3
3
##     lEX_noW_3
##             8
```

- Matriz que aúna la información anterior:

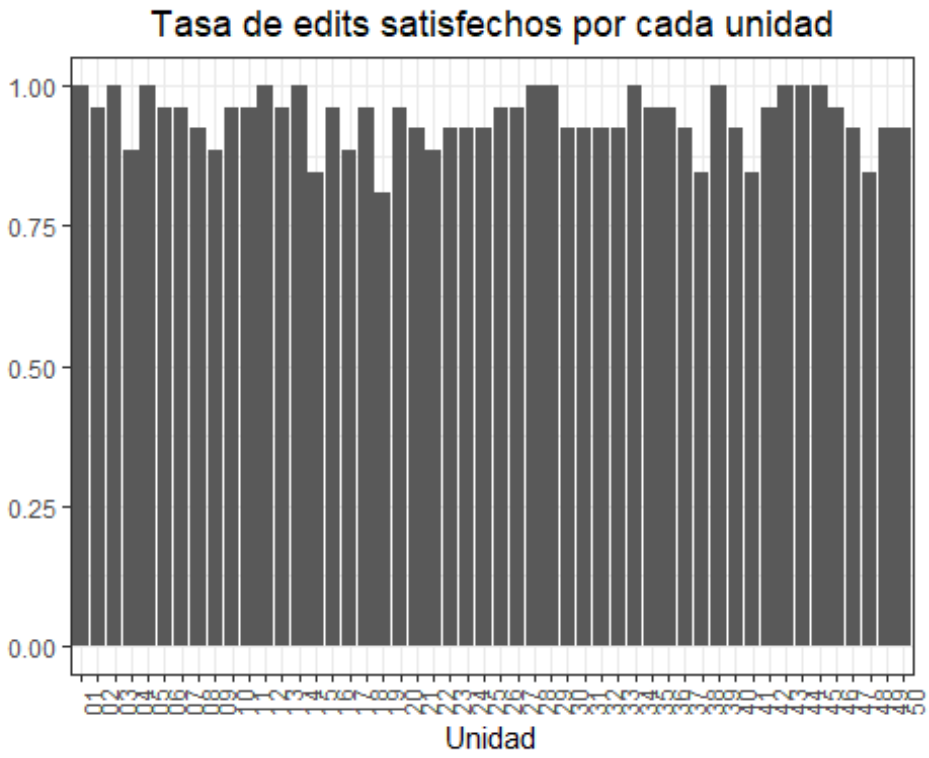
```
relPass_edit.df <- as.data.frame(aggregate(cf))
```

```
##                   npass nfail nNA rel.pass rel.fail rel.NA      editNames
## Req3                 50     0   0     1.00     0.00      0           Req3
## Req4                 50     0   0     1.00     0.00      0           Req4
## Req5                 50     0   0     1.00     0.00      0           Req5
## Req6                 50     0   0     1.00     0.00      0           Req6
## Req7                 50     0   0     1.00     0.00      0           Req7
## Rango5               50     0   0     1.00     0.00      0         Rango5
## Rango6               50     0   0     1.00     0.00      0         Rango6
## Rango7               50     0   0     1.00     0.00      0         Rango7
## Rango8               50     0   0     1.00     0.00      0         Rango8
## Rango14              50     0   0     1.00     0.00      0        Rango14
## Nulo1                50     0   0     1.00     0.00      0          Nulo1
## Nulo2                50     0   0     1.00     0.00      0          Nulo2
## CNAnt                49     1   0     0.98     0.02      0          CNAnt
## Emp_W_1              50     0   0     1.00     0.00      0        Emp_W_1
## EmpNoRem_W_1         50     0   0     1.00     0.00      0  EmpNoRem_W_1
## EmpRemFij_W_1        49     1   0     0.98     0.02      0 EmpRemFij_W_1
## EmpRemEve_W_1        50     0   0     1.00     0.00      0 EmpRemEve_W_1
## lCN_W_1              35    15   0     0.70     0.30      0        lCN_W_1
## lCN_W_3              40    10   0     0.80     0.20      0        lCN_W_3
## lCN_noW_1            38    12   0     0.76     0.24      0      lCN_noW_1
## lrCNPR_W_1           34    16   0     0.68     0.32      0     lrCNPR_W_1
## lrCNPR_W_3           46     4   0     0.92     0.08      0     lrCNPR_W_3
## lrCNPR_noW_3         47     3   0     0.94     0.06      0  lrCNPR_noW_3
## lEX_W_1              47     3   0     0.94     0.06      0        lEX_W_1
## lEX_W_3              47     3   0     0.94     0.06      0        lEX_W_3
## lEX_noW_3            42     8   0     0.84     0.16      0     lEX_noW_3
```

- Histograma de los registros pasados por cada regla y de la tasa de *edits* pasados por cada registro:

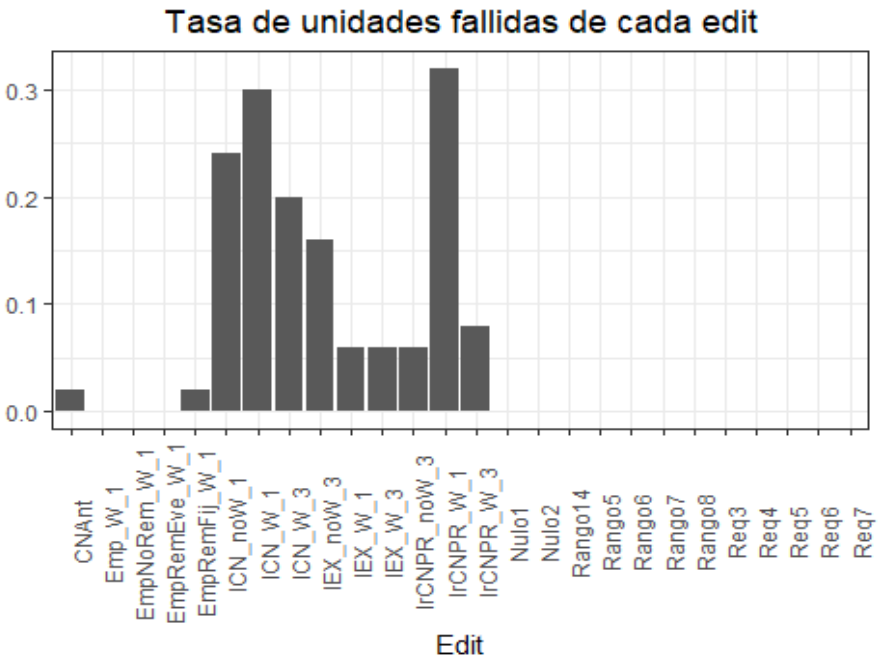


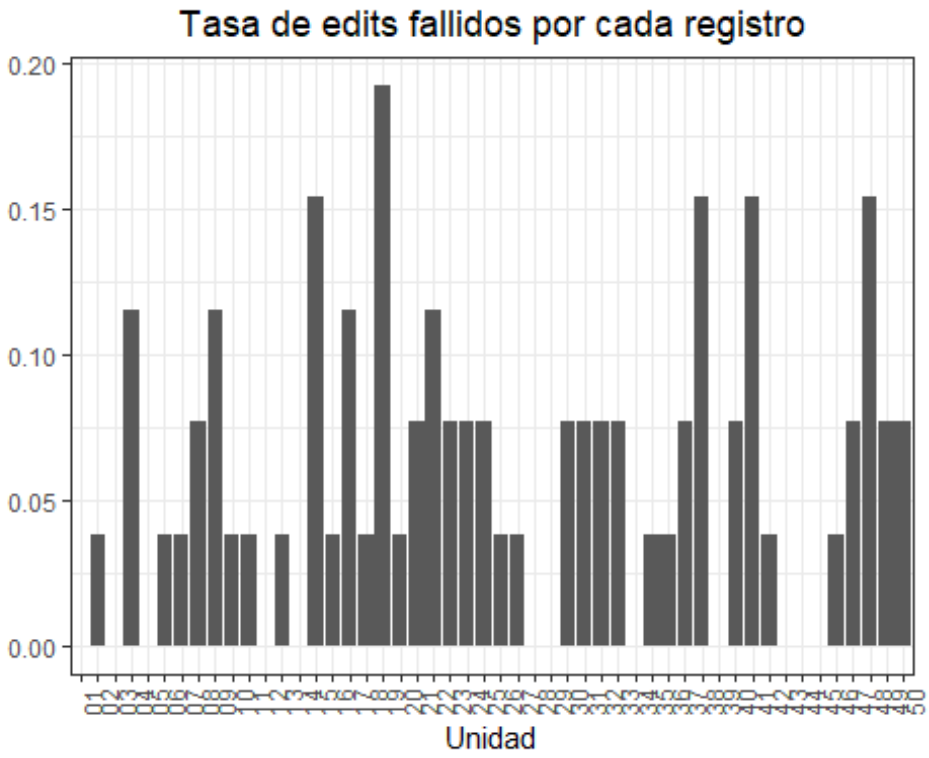
El gráfico presenta la tasa de respuestas correctas de cada *Edit* siendo 1 el máximo. La mayoría de los controles pasan con la tasa máxima lo cual quiere decir que no se incumple ese posible error en ningun caso. Sin embargo, hay controles que no llegan al 75% de acierto, y habría que saber por qué se responde con alta tasa de error. Estos campos *lCN_noW_1, lNC_W_1* y *lrCNPR_W_1* tienen errores que, o bien deben ser justificados, como que la cifra de negocios sea inusual, o bien debe corregirse la pregunta para que no genere tan alta tasa de error, o bien debe adaptarse los límites para el intervalo de confianza.



El presente histograma muestra que no hay ningún registro que esté especialmente peor informado que la media, por lo cual se puede concluir que los fallos que se producen pueden ser provocados por la constitución de la encuesta, y no por error específico del usuario al rellenarla.

- Histograma de los registros fallidos por cada regla y de la tasa de *edits* fallidos por cada registro:





La información que proporcionan los dos últimos gráficos presentados pretende complementar la información presentada anteriormente en los gráficos que muestran los registros y los *edits* que han pasado los controles, llegando así, a las mismas conclusiones.

4.2 ENSE Adulto

- *Flag* por unidad y *Edit.* Se exponen únicamente 3 unidades del registro a modo de ejemplo, y se omiten las restantes:

```
##      Req1  Req2  Req3  Req4 Req5 Req6 Req7 Req8 Req9 Req10 Nacionalidad1
## [1,] TRUE FALSE FALSE FALSE TRUE TRUE TRUE TRUE TRUE  TRUE          TRUE
## [2,] TRUE FALSE FALSE FALSE TRUE TRUE TRUE TRUE TRUE  TRUE          TRUE
## [3,] TRUE FALSE FALSE FALSE TRUE TRUE TRUE TRUE TRUE  TRUE          TRUE

##      Nacionalidad2 Años_Residencia Duplicada1 Duplicada2 Empresario1
## [1,]          TRUE            TRUE       TRUE      FALSE        TRUE
## [2,]          TRUE            TRUE       TRUE       TRUE        TRUE
## [3,]          TRUE            TRUE       TRUE      FALSE        TRUE
##      Empresario2 Asalariado1 Asalariado2 Empresario3 DuracionContr1
## [1,]        TRUE        TRUE        TRUE        TRUE           TRUE
## [2,]        TRUE        TRUE        TRUE        TRUE           TRUE
## [3,]        TRUE        TRUE        TRUE        TRUE           TRUE

##      DuracionContr2 EnfermedadCronica1 EnfermedadCronica3 EnfermedadCronica4
## [1,]           TRUE               TRUE               TRUE             TRUE
## [2,]           TRUE               TRUE               TRUE             TRUE
## [3,]           TRUE               TRUE               TRUE             TRUE
```

```
##      EnfermedadCronica5 EnfermedadCronica6 EnfermedadCronica7
## [1,]               TRUE               TRUE               TRUE
## [2,]               TRUE               TRUE               TRUE
## [3,]               TRUE               TRUE               TRUE

##      EnfermedadCronica8 EnfermedadCronica9 EnfermedadCronica10
## [1,]               TRUE               TRUE                TRUE
## [2,]               TRUE               TRUE                TRUE
## [3,]               TRUE               TRUE                TRUE

##      EnfermedadCronica11 EnfermedadCronica12 EnfermedadCronica13
## [1,]                TRUE                TRUE                TRUE
## [2,]                TRUE                TRUE                TRUE
## [3,]                TRUE                TRUE                TRUE

##      EnfermedadCronica14 EnfermedadCronica15 EnfermedadCronica16
## [1,]               TRUE                TRUE               TRUE
## [2,]               TRUE                TRUE               TRUE
## [3,]               TRUE                TRUE               TRUE

##      EnfermedadCronica17 EnfermedadCronica18 EnfermedadCronica19
## [1,]               TRUE                TRUE               TRUE
## [2,]               TRUE                TRUE               TRUE
## [3,]               TRUE                TRUE               TRUE

##      EnfermedadCronica20 EnfermedadCronica21 EnfermedadCronica22
## [1,]               TRUE                TRUE               TRUE
## [2,]               TRUE                TRUE               TRUE
## [3,]               TRUE                TRUE               TRUE

##      EnfermedadCronica23 EnfermedadCronica24 EnfermedadCronica25
## [1,]               TRUE                TRUE               TRUE
## [2,]               TRUE                TRUE               TRUE
## [3,]               TRUE                TRUE               TRUE

##      EnfermedadCronica26 EnfermedadCronica27 EnfermedadCronica28
## [1,]               TRUE                TRUE               TRUE
## [2,]               TRUE                TRUE               TRUE
## [3,]               TRUE                TRUE               TRUE
##      EnfermedadCronica29 EnfermedadCronica30 EnfermedadCronica31
## [1,]               TRUE                TRUE               TRUE
## [2,]               TRUE                TRUE               TRUE
## [3,]               TRUE                TRUE               TRUE
##      EnfermedadCronica32 EnfermedadCronica33 EnfermedadCronica34 Prostata
## [1,]               TRUE                TRUE                TRUE     TRUE
## [2,]               TRUE                TRUE                TRUE     TRUE
## [3,]               TRUE                TRUE                TRUE     TRUE

##      Menopausia Vision Audicion Dependiente1 Duplicada3 Dependiente2
## [1,]       TRUE   TRUE     TRUE          TRUE      FALSE         TRUE
## [2,]       TRUE   TRUE     TRUE          TRUE       TRUE         TRUE
## [3,]       TRUE   TRUE     TRUE          TRUE       TRUE         TRUE
##      VisitaMedico VisitaMedico2 VisitaMedico3 LimiteTemp1 LimiteTemp2
## [1,]         TRUE           TRUE           TRUE        TRUE        TRUE
## [2,]         TRUE           TRUE           TRUE        TRUE        TRUE
## [3,]         TRUE           TRUE           TRUE        TRUE        TRUE
##      LimiteTemp3 LimiteTemp4 LimiteTemp5 Privacidad1 Privacidad2
PruebaMedica1
## [1,]        TRUE        TRUE        TRUE        TRUE        TRUE
```

```
TRUE
## [2,]        TRUE        TRUE        TRUE        TRUE        TRUE
TRUE
## [3,]        TRUE        TRUE        TRUE        TRUE        TRUE
TRUE
##      Dentista1 Dentista2 Dentista3 IngresoHospitalario1 IngresoHospitalario2
## [1,]      TRUE      TRUE     FALSE                 TRUE                 TRUE
## [2,]      TRUE      TRUE     FALSE                 TRUE                 TRUE
## [3,]      TRUE      TRUE     FALSE                 TRUE                 TRUE
##      IngresoHospitalario3 IngresoHospitalario4 IngresoHospitalario5
## [1,]                 TRUE                 TRUE                 TRUE
## [2,]                 TRUE                 TRUE                 TRUE
## [3,]                 TRUE                 TRUE                 TRUE
##      IngresoHospitalario6 IngresoHospitalario7 IngresoHospitalario8
## [1,]                 TRUE                 TRUE                 TRUE
## [2,]                 TRUE                 TRUE                 TRUE
## [3,]                 TRUE                 TRUE                 TRUE
##      IngresoHospitalario9 TipoSegSan Colonoscopia Mamografia CitologiaV
## [1,]                 TRUE       TRUE         TRUE       TRUE       TRUE
## [2,]                 TRUE       TRUE         TRUE       TRUE       TRUE
## [3,]                 TRUE       TRUE         TRUE       TRUE       TRUE
##      CitologiaHombres Altura Peso DiasActividadFisica minutosActividadFisica
## [1,]             TRUE   TRUE TRUE                TRUE                  TRUE
## [2,]             TRUE   TRUE TRUE                TRUE                  TRUE
## [3,]             TRUE   TRUE TRUE                TRUE                  TRUE
##      DiasActividadFisica2 HorasActividadFisica2 minutosActividadFisica2
## [1,]                 TRUE                  TRUE                    TRUE
## [2,]                 TRUE                  TRUE                    TRUE
## [3,]                 TRUE                  TRUE                    TRUE
##      DiasActividadFisica3 HorasActividadFisica3 minutosActividadFisica3
## [1,]                 TRUE                  TRUE                    TRUE
## [2,]                 TRUE                  TRUE                    TRUE
## [3,]                 TRUE                  TRUE                    TRUE
##      HorasActividadFisica4 minutosActividadFisica4 Fumadores FumadoresDia
## [1,]                  TRUE                   TRUE      TRUE         TRUE
## [2,]                  TRUE                   TRUE      TRUE         TRUE
## [3,]                  TRUE                   TRUE      TRUE         TRUE
##      edadfumador
## [1,]        TRUE
## [2,]        TRUE
## [3,]        TRUE
## [4,]        TRUE
## [5,]        TRUE
## [6,]        TRUE
```

- Número de unidades satisfechas por *edit:*

```
npass <- colSums(flag.matrix)


##                    Req1                    Req2                    Req3
##                   23089                     467                     218
##                    Req4                    Req5                    Req6
##                     218                   23089                   23089
##                    Req7                    Req8                    Req9
##                   23089                   23089                   23089
##                   Req10           Nacionalidad1           Nacionalida2
##                   23089                   23089                   23089
##         Años_Residencia              Duplicada1              Duplicada2
```

```
##                  23089                 19133                 12452
##             Empresario1           Empresario2          Asalariado1
##                  23089                 23089                 23089
##             Asalariado2           Empresario3       DuracionContr1
##                  23089                 23089                 23089
##          DuracionContr2      EnfermedadCronica1  EnfermedadCronica3
##                  23089                 23013                 22973
##      EnfermedadCronica4      EnfermedadCronica5  EnfermedadCronica6
##                  23073                 23077                 23064
##      EnfermedadCronica7      EnfermedadCronica8  EnfermedadCronica9
##                  22434                 22864                 22826
##     EnfermedadCronica10     EnfermedadCronica11 EnfermedadCronica12
##                  22777                 22914                 23051
##     EnfermedadCronica13     EnfermedadCronica14 EnfermedadCronica15
##                  23057                 23080                 23031
##     EnfermedadCronica16     EnfermedadCronica17 EnfermedadCronica18
##                  22981                 23021                 22960
##     EnfermedadCronica19     EnfermedadCronica20 EnfermedadCronica21
##                  23029                 22978                 23079
##     EnfermedadCronica22     EnfermedadCronica23 EnfermedadCronica24
##                  22881                 22890                 23081
##     EnfermedadCronica25     EnfermedadCronica26 EnfermedadCronica27
##                  23067                 22805                 22843
##     EnfermedadCronica28     EnfermedadCronica29 EnfermedadCronica30
##                  23049                 23052                 23058
##     EnfermedadCronica31     EnfermedadCronica32 EnfermedadCronica33
##                  23040                 23059                 23001
##     EnfermedadCronica34               Prostata          Menopausia
##                  23039                 23089                 23089
##                 Vision               Audicion        Dependiente1
##                  23089                 23089                 22565
##              Duplicada3           Dependiente2        VisitaMedico
##                  20638                 22677                 23089
##           VisitaMedico2         VisitaMedico3         LimiteTemp1
##                  23089                 23089                 22940
##             LimiteTemp2           LimiteTemp3         LimiteTemp4
##                  22940                 22940                 22807
##             LimiteTemp5           Privacidad1         Privacidad2
##                  22807                 22972                 23089
##           PruebaMedica1             Dentista1           Dentista2
##                  23087                 18292                 22850
##               Dentista3 IngresoHospitalario1 IngresoHospitalario2
##                  16090                 22872                 23089
##    IngresoHospitalario3 IngresoHospitalario4 IngresoHospitalario5
##                  23089                 23089                 23076
##    IngresoHospitalario6 IngresoHospitalario7 IngresoHospitalario8
##                  23089                 21119                 22709
##    IngresoHospitalario9           TipoSegSan        Colonoscopia
##                  23089                 23089                 23076
##              Mamografia           CitologiaV    CitologiaHombres
##                  23055                 23029                 23089
##                 Altura                  Peso DiasActividadFisica
##                  23089                 23089                 23071
##  minutosActividadFisica DiasActividadFisica2 HorasActividadFisica2
```

```
##                   23076                   23068                   23072
## minutosActividadFisica2  DiasActividadFisica3 HorasActividadFisica3
##                   23072                   23054                   22953
##minutosActividadFisica3 HorasActividadFisica4  minutosActividadFisica4
##                   22953                   22711                   22711
##               Fumadores            FumadoresDia             edadfumador
##                   23089                   23089                   23089
```

- Número de unidades fallidas por *edit:*

```
##                Req1                Req2                Req3
##                   0               22622               22871
##                Req4                Req5                Req6
##               22871                   0                   0
##                Req7                Req8                Req9
##                   0                   0                   0
##               Req10        Nacionalidad1        Nacionalidad2
##                   0                   0                   0
##     Años_Residencia           Duplicada1           Duplicada2
##                   0                3956               10637
##         Empresario1          Empresario2          Asalariado1
##                   0                   0                   0
##         Asalariado2          Empresario3       DuracionContr1
##                   0                   0                   0
##      DuracionContr2   EnfermedadCronica1   EnfermedadCronica3
##                   0                  76                 116
##  EnfermedadCronica4   EnfermedadCronica5   EnfermedadCronica6
##                  16                  12                  25
##  EnfermedadCronica7   EnfermedadCronica8   EnfermedadCronica9
##                 655                 225                 263
## EnfermedadCronica10  EnfermedadCronica11  EnfermedadCronica12
##                 312                 175                  38
## EnfermedadCronica13  EnfermedadCronica14  EnfermedadCronica15
##                  32                   9                  58
## EnfermedadCronica16  EnfermedadCronica17  EnfermedadCronica18
##                 108                  68                 129
## EnfermedadCronica19  EnfermedadCronica20  EnfermedadCronica21
##                  60                 111                  10
## EnfermedadCronica22  EnfermedadCronica23  EnfermedadCronica24
##                 208                 199                   8
## EnfermedadCronica25  EnfermedadCronica26  EnfermedadCronica27
##                  22                 284                 246
## EnfermedadCronica28  EnfermedadCronica29  EnfermedadCronica30
##                  40                  37                  31
## EnfermedadCronica31  EnfermedadCronica32  EnfermedadCronica33
##                  49                  30                  88
## EnfermedadCronica34            Prostata           Menopausia
##                  50                   0                   0
##              Vision             Audicion         Dependiente1
##                   0                   0                 524
##          Duplicada3         Dependiente2         VisitaMedico
##                2451                 412                   0
```

```
##       VisitaMedico2        VisitaMedico3          LimiteTemp1
##                   0                    0                  149
##         LimiteTemp2          LimiteTemp3          LimiteTemp4
##                 149                  149                  282
##         LimiteTemp5          Privacidad1          Privacidad2
##                 282                  117                    0
##       PruebaMedica1            Dentista1            Dentista2
##                   2                 4797                  239
##           Dentista3 IngresoHospitalario1 IngresoHospitalario2
##                6999                  217                    0
##IngresoHospitalario3 IngresoHospitalario4 IngresoHospitalario5
##                   0                    0                   13
##IngresoHospitalario6 IngresoHospitalario7 IngresoHospitalario8
##                   0                 1970                  380
## IngresoHospitalario9          TipoSegSan         Colonoscopia
##                   0                    0                   13
##            Mamografia          CitologiaV     CitologiaHombres
##                  34                   60                    0
##               Altura                 Peso  DiasActividadFisica
##                   0                    0                   18
##minutosActividadFisica   DiasActividadFisica2  HorasActividadFisica2
##                  13                   21                   17
##minutosActividadFisica2  DiasActividadFisica3  HorasActividadFisica3
##                  17                   35                   136
## minutosActividadFisica3    HorasActividadFisica4  minutosActividadFisica4
##                  136                  378                  378
##            Fumadores         FumadoresDia          edadfumador
##                   0                    0                    0
```

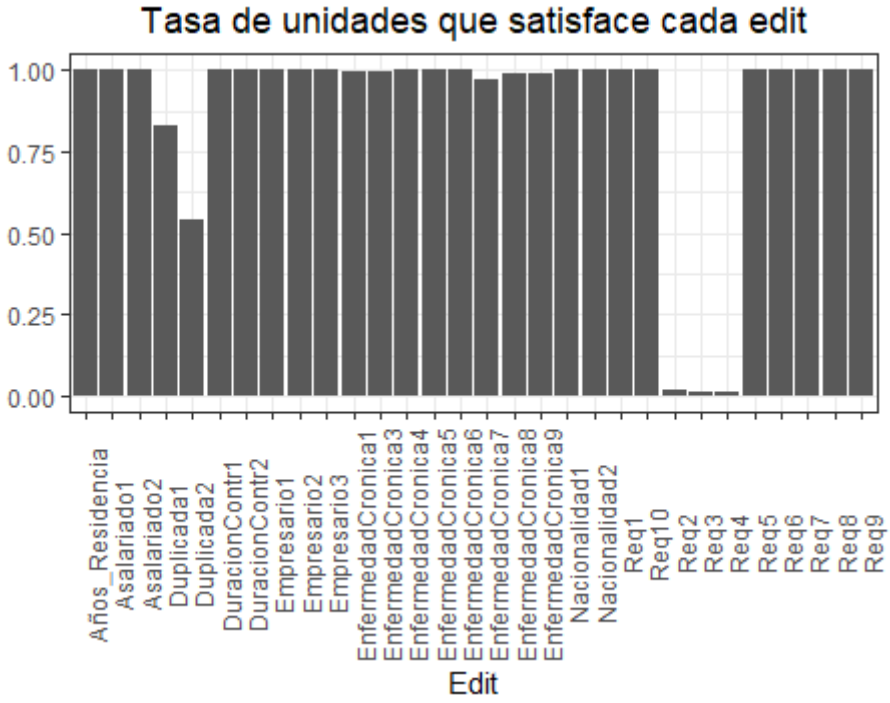
- Matriz que aúna la información anterior:

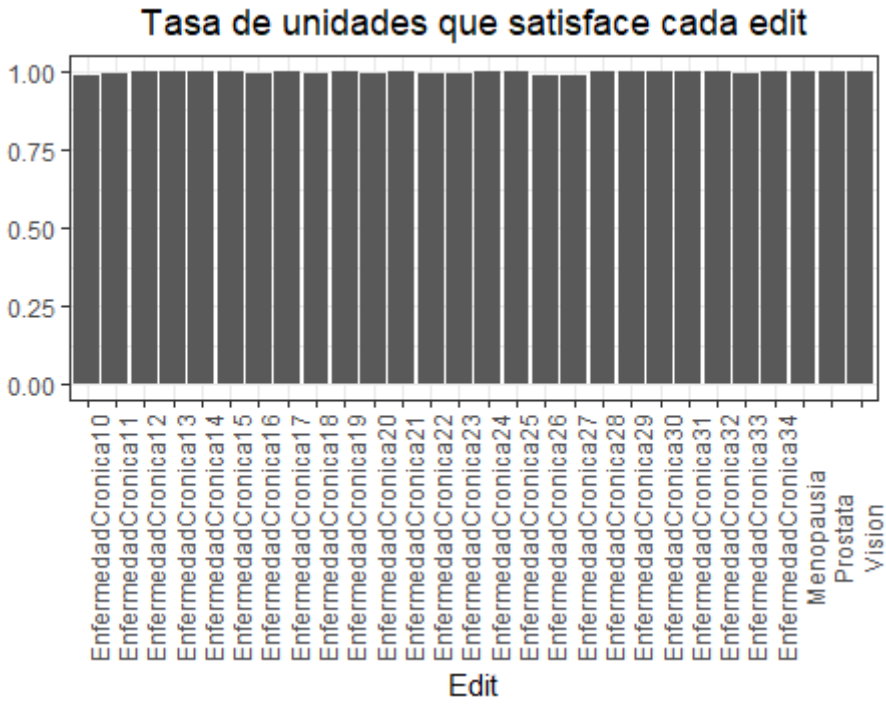
```
##            npass   nfail   nNA     rel.pass       rel.fail rel.NA
## Req1       23089       0     0  1.000000000   0.000000e+00      0
## Req2         467   22622     0  0.020226082   9.797739e-01      0
## Req3         218   22871     0  0.009441725   9.905583e-01      0
## Req4         218   22871     0  0.009441725   9.905583e-01      0
## Req5       23089       0     0  1.000000000   0.000000e+00      0
## Req6       23089       0     0  1.000000000   0.000000e+00      0
## Req7       23089       0     0  1.000000000   0.000000e+00      0
## Req8       23089       0     0  1.000000000   0.000000e+00      0
## Req9       23089       0     0  1.000000000   0.000000e+00      0
## Req10      23089       0     0  1.000000000   0.000000e+00      0
## Nacionalidad1 23089    0     0  1.000000000   0.000000e+00      0
## Nacionalidad2 23089    0     0  1.000000000   0.000000e+00      0
## Años_Residencia 2308   0     0  1.000000000   0.000000e+00      0
## Duplicada1 19133    3956     0  0.828663000   1.713370e-01      0
## Duplicada2 12452   10637     0  0.539304431   4.606956e-01      0
## Empresario1 23089      0     0  1.000000000   0.000000e+00      0
## Empresario2 23089      0     0  1.000000000   0.000000e+00      0
## Asalariado1 23089      0     0  1.000000000   0.000000e+00      0
## Asalariado2 23089      0     0  1.000000000   0.000000e+00      0
## Empresario3 23089      0     0  1.000000000   0.000000e+00      0
## DuracionContr1 23089   0     0  1.000000000   0.000000e+00      0
## DuracionContr2 23089   0     0  1.000000000   0.000000e+00      0
```

```
##EnfermedadCronica1  23013  76 0  0.996708389  3.291611e-03      0
##EnfermedadCronica3  22973 116 0  0.994975963  5.024037e-03      0
##EnfermedadCronica4  23073  16 0  0.999307029  6.929707e-04      0
##EnfermedadCronica5  23077  12 0  0.999480272  5.197280e-04      0
##EnfermedadCronica6  23064  25 0  0.998917233  1.082767e-03      0
##EnfermedadCronica7  22434 655 0  0.971631513  2.836849e-02      0
##EnfermedadCronica8  22864 225 0  0.990255100  9.744900e-03      0
##EnfermedadCronica9  22826 263 0  0.988609294  1.139071e-02      0
##EnfermedadCronica10 22777 312 0  0.986487072  1.351293e-02      0
##EnfermedadCronica11 22914 175 0  0.992420633  7.579367e-03      0
##EnfermedadCronica12 23051  38 0  0.998354195  1.645805e-03      0
##EnfermedadCronica13 23057  32 0  0.998614059  1.385941e-03      0
##EnfermedadCronica14 23080   9 0  0.999610204  3.897960e-04      0
##EnfermedadCronica15 23031  58 0  0.997487981  2.512019e-03      0
##EnfermedadCronica16 22981 108 0  0.995322448  4.677552e-03      0
##EnfermedadCronica17 23021  68 0  0.997054875  2.945125e-03      0
##EnfermedadCronica18 22960 129 0  0.994412924  5.587076e-03      0
##EnfermedadCronica19 23029  60 0  0.997401360  2.598640e-03      0
##EnfermedadCronica20 22978 111 0  0.995192516  4.807484e-03      0
##EnfermedadCronica21 23079  10 0  0.999566893  4.331067e-04      0
##EnfermedadCronica22 22881 208 0  0.990991381  9.008619e-03      0
##EnfermedadCronica23 22890 199 0  0.991381177  8.618823e-03      0
##EnfermedadCronica24 23081   8 0  0.999653515  3.464853e-04      0
##EnfermedadCronica25 23067  22 0  0.999047165  9.528347e-04      0
##EnfermedadCronica26 22805 284 0  0.987699770  1.230023e-02      0
##EnfermedadCronica27 22843 246 0  0.989345576  1.065442e-02      0
##EnfermedadCronica28 23049  40 0  0.998267573  1.732427e-03      0
##EnfermedadCronica29 23052  37 0  0.998397505  1.602495e-03      0
##EnfermedadCronica30 23058  31 0  0.998657369  1.342631e-03      0
##EnfermedadCronica31 23040  49 0  0.997877777  2.122223e-03      0
##EnfermedadCronica32 23059  30 0  0.998700680  1.299320e-03      0
##EnfermedadCronica33 23001  88 0  0.996188661  3.811339e-03      0
##EnfermedadCronica34 23039  50 0  0.997834467  2.165533e-03      0
## Prostata           23089   0 0  1.000000000  0.000000e+00      0
## Menopausia         23089   0 0  1.000000000  0.000000e+00      0
## Vision             23089   0 0  1.000000000  0.000000e+00      0
## Audicion           23089   0 0  1.000000000  0.000000e+00      0
## Dependiente1       22565 524 0  0.977305210  2.269479e-02      0
## Duplicada3         20638 2451 0  0.893845554 1.061544e-01      0
## Dependiente2       22677 412 0  0.982156005  1.784399e-02      0
## VisitaMedico       23089   0 0  1.000000000  0.000000e+00      0
## VisitaMedico2      23089   0 0  1.000000000  0.000000e+00      0
## VisitaMedico3      23089   0 0  1.000000000  0.000000e+00      0
## LimiteTemp1        22940 149 0  0.993546711  6.453289e-03      0
## LimiteTemp2        22940 149 0  0.993546711  6.453289e-03      0
## LimiteTemp3        22940 149 0  0.993546711  6.453289e-03      0
## LimiteTemp4        22807 282 0  0.987786392  1.221361e-02      0
## LimiteTemp5        22807 282 0  0.987786392  1.221361e-02      0
## Privacidad1        22972 117 0  0.994932652  5.067348e-03      0
## Privacidad2        23089   0 0  1.000000000  0.000000e+00      0
## PruebaMedica1      23087   2 0  0.999913379  8.662133e-05      0
## Dentista1          18292 4797 0 0.792238728  2.077613e-01      0
## Dentista2          22850 239 0  0.989648750  1.035125e-02      0
## Dentista3          16090 6999 0 0.696868639  3.031314e-01      0
```

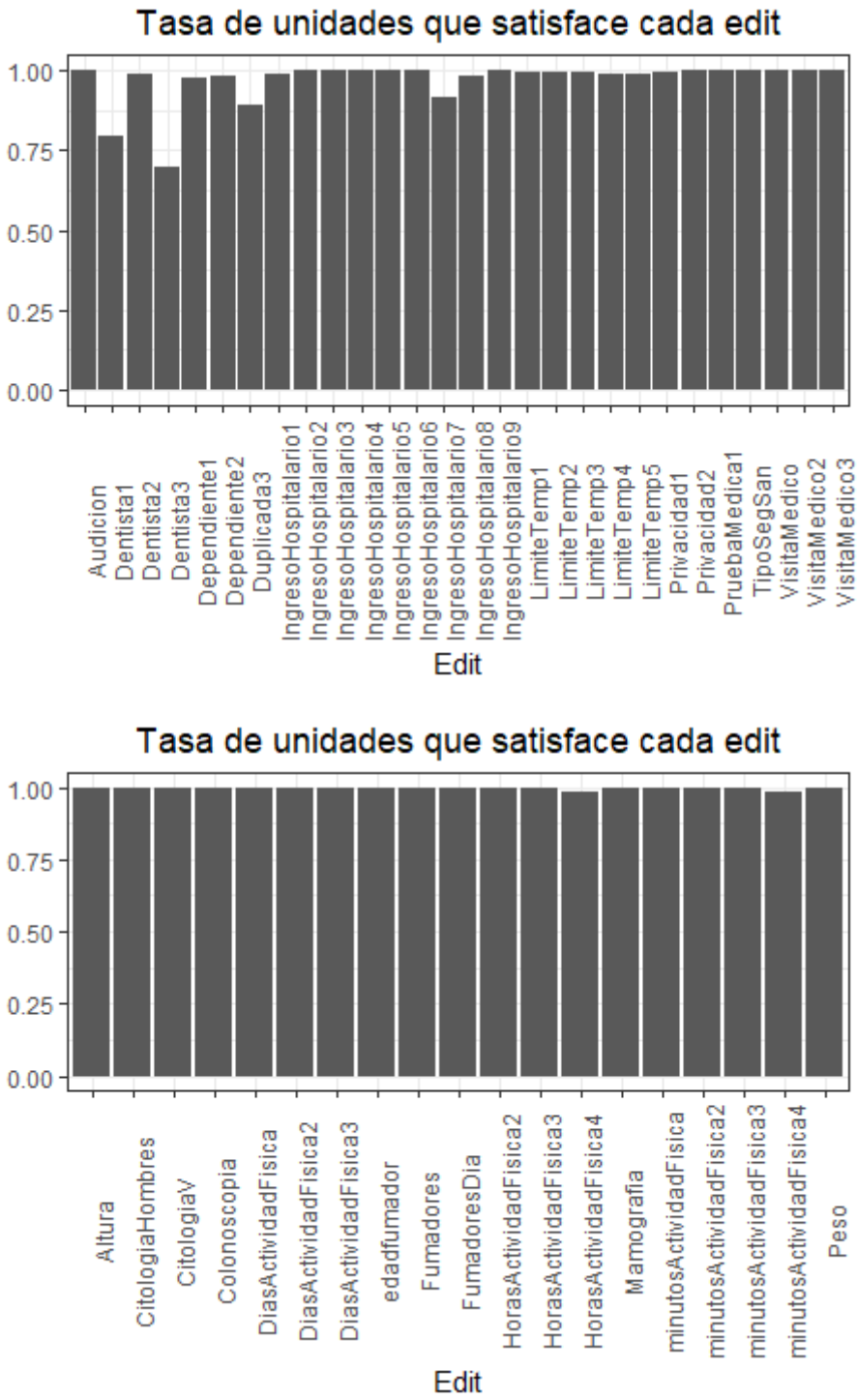
```
## IngresoHospitalario1 22872 217 0  0.990601585 9.398415e-03     0
## IngresoHospitalario2 23089   0 0  1.000000000 0.000000e+00     0
## IngresoHospitalario3 23089   0 0  1.000000000 0.000000e+00     0
## IngresoHospitalario4 23089   0 0  1.000000000 0.000000e+00     0
## IngresoHospitalario5 23076  13 0  0.999436961 5.630387e-04     0
## IngresoHospitalario6 23089   0 0  1.000000000 0.000000e+00     0
## IngresoHospitalario7 21119 1970 0 0.914677985 8.532201e-02     0
## IngresoHospitalario8 22709  380 0 0.983541946 1.645805e-02     0
## IngresoHospitalario9 23089    0 0 1.000000000 0.000000e+00     0
## TipoSegSan           23089    0 0 1.000000000 0.000000e+00     0
## Colonoscopia         23076   13 0 0.999436961 5.630387e-04     0
## Mamografia           23055   34 0 0.998527437 1.472563e-03     0
## CitologiaV           23029   60 0 0.997401360 2.598640e-03     0
## CitologiaHombres     23089    0 0 1.000000000 0.000000e+00     0
## Altura               23089    0 0 1.000000000 0.000000e+00     0
## Peso                 23089    0 0 1.000000000 0.000000e+00     0
## DiasActividadFisica  23071   18 0 0.999220408 7.795920e-04     0
## minutosActividadFisica 23076 13 0 0.999436961 5.630387e-04     0
## DiasActividadFisica2 23068   21 0 0.999090476 9.095240e-04     0
## HorasActividadFisica2 23072  17 0 0.999263719 7.362813e-04     0
## minutosActividadFisica2 23072 17 0 0.999263719 7.362813e-04    0
## DiasActividadFisica3  23054  35 0 0.998484127 1.515873e-03     0
## HorasActividadFisica3 22953 136 0 0.994109749 5.890251e-03     0
## minutosActividadFisica3 22953 1360 0.994109749 5.890251e-03    0
## HorasActividadFisica4 22711 378  0.983628568 1.637143e-02      0
## minutosActividadFisica4 22711 378 0 0.983628568 1.637143e-02   0
## Fumadores             23089   0  0  1.000000000 0.000000e+00   0
## FumadoresDia          23089   0  0  1.000000000 0.000000e+00   0
## edadfumador           23089   0  0  1.000000000 0.000000e+00   0
```

- Histograma de los registros pasados por cada regla y de la tasa de *Edits* pasados por cada registro[17]:





[17] Se ha dividido el histograma del número de registros pasados y fallidos por *Edit* en 4 gráficos para hacerlo inteligible.

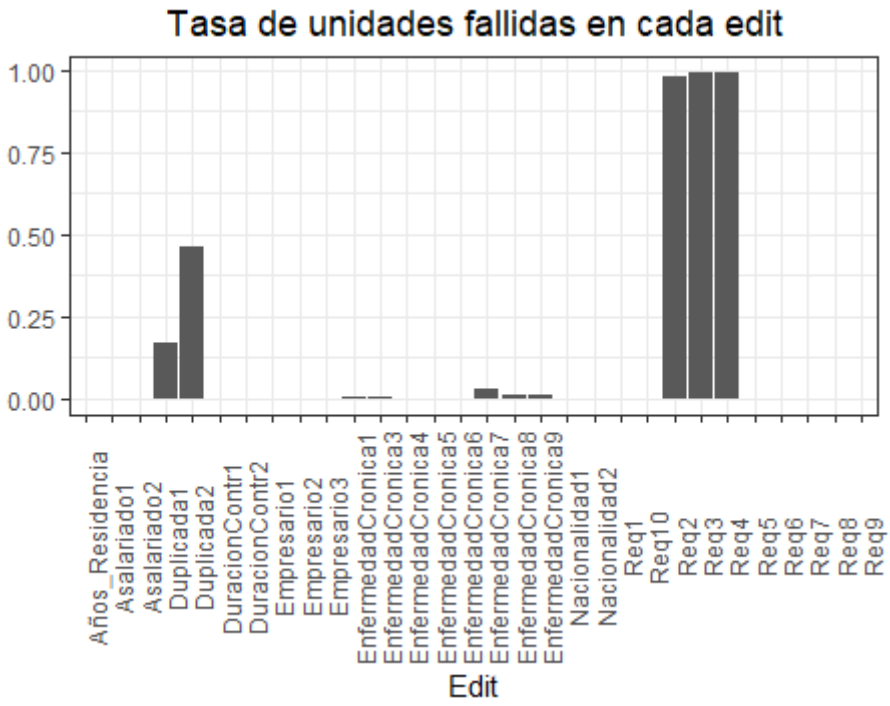


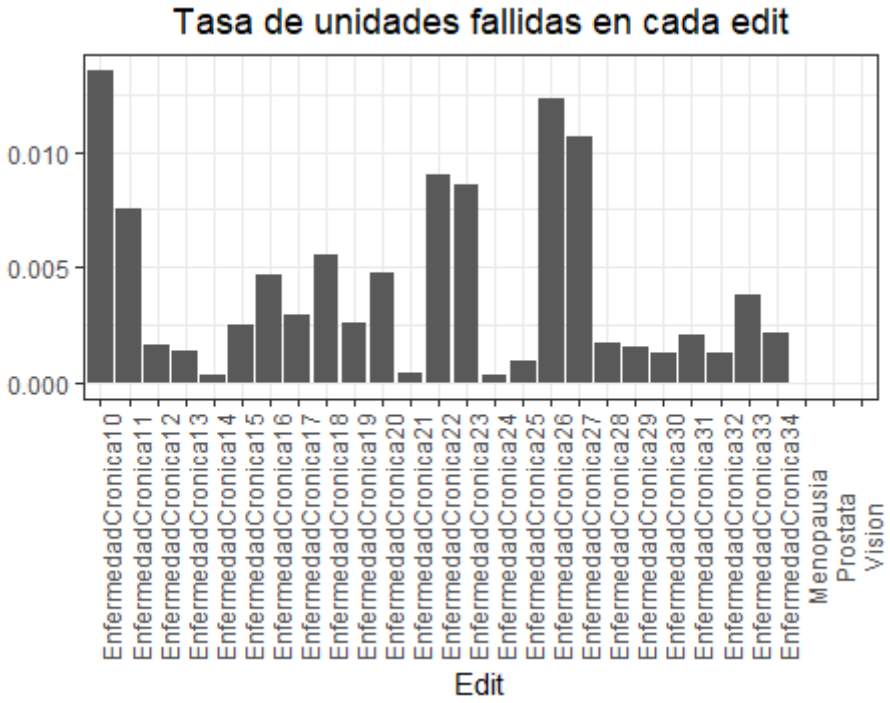
El gráfico presenta la tasa de respuestas correctas de cada *Edit* siendo 1 el máximo. La mayoría de los controles pasan con la tasa máxima, lo cual quiere decir que no se incumple ese posible error en ningun caso. Sin embargo, hay controles que no llegan al 75% de acierto, y habría que saber por qué se responde con alta tasa de error en esos campos. Estos campos *Req2, Req3* y *Req4* tienen tasas de acierto muy bajas, lo cual quiere decir que las variables *PROXY_2b*, *PROXY_3b* y *PROXY_4* que deberían estar contestadas y no es así. Las variables *Dentista1* y

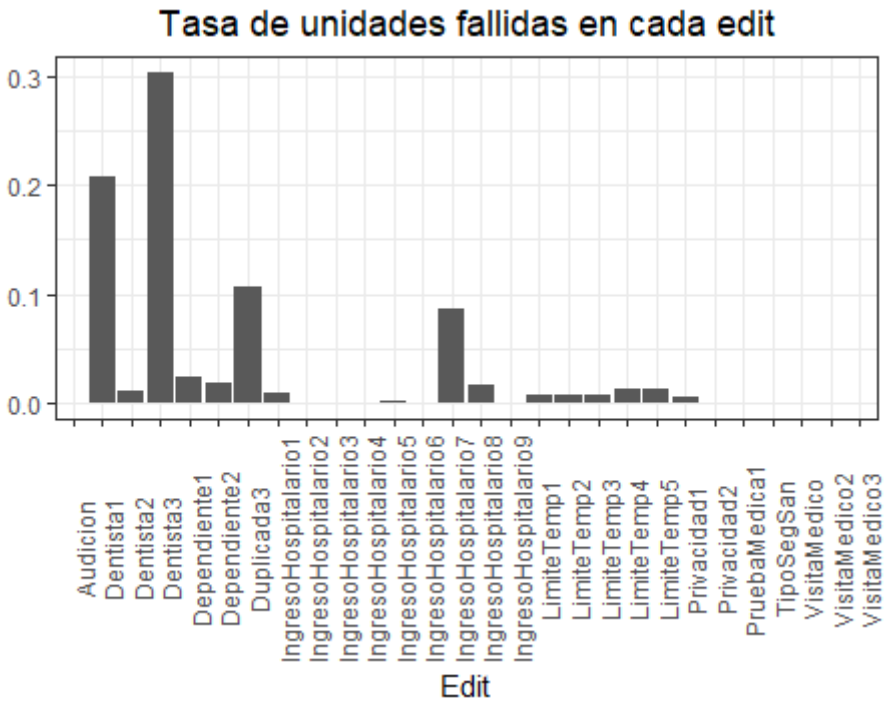
Dentista3 también tienen una tasa de acierto menor a la apropiada, esto quiere decir que no hay coherencia entre las preguntas de los empastes y sobre las fundas y prótesis. Otra cuestión a resolver es que las variables F10 y F20 que son la misma pregunta presentan resultados diferentes.

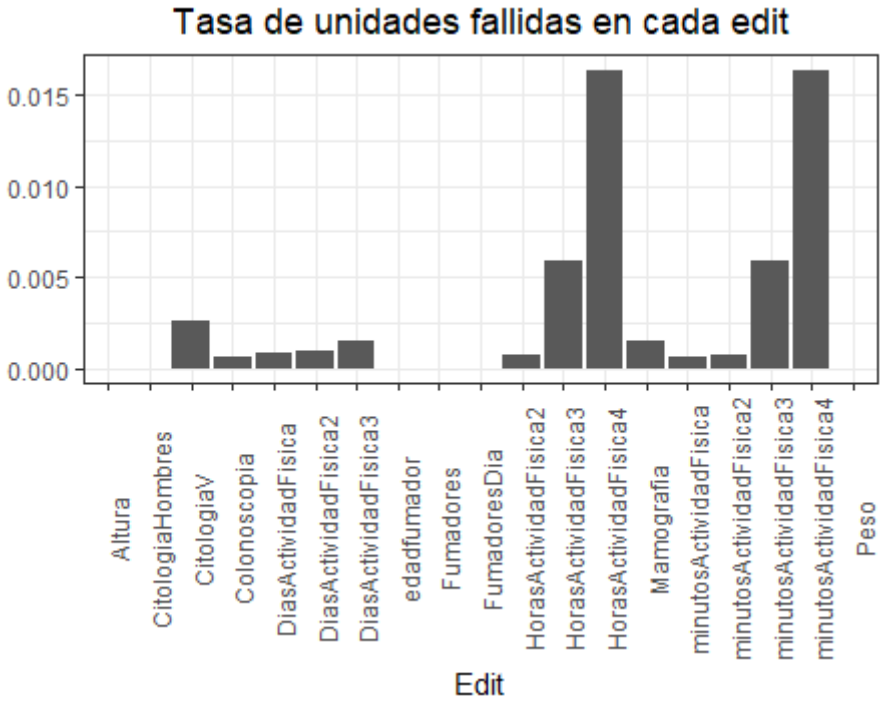
En este caso se puede apreciar como hay muestras que presenta tasas de acierto nulas o prácticamente nulas, lo cual indica que hay usuarios que responden inapropiadamente al test en todas o casi todas sus preguntas o que se realiza un mal registro de dichas repuestas. Esto debe ser tenido en cuenta pues adultera de una manera incontrolable la tasa de error de las preguntas sin que los encargados de los test puedan hacer nada al respecto.

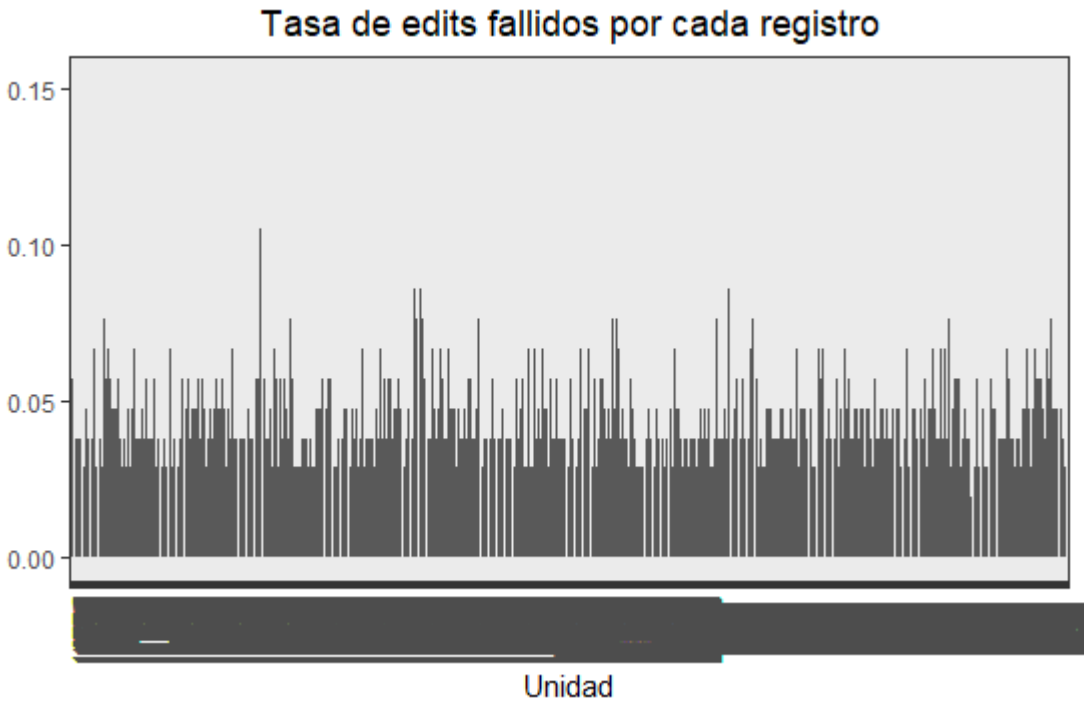
- Histograma de los registros fallidos por cada regla y de la tasa de *edits* fallidos por cada registro:


Tasa de unidades fallidas en cada edit
1.00
0.75
0.50
0.25
0.00
Años_Residencia
Asalariado1
Asalariado2
Duplicada1
Duplicada2
DuracionContr1
DuracionContr2
Empresario1
Empresario2
Empresario3
EnfermedadCronica1
EnfermedadCronica3
EnfermedadCronica4
EnfermedadCronica5
EnfermedadCronica6
EnfermedadCronica7
EnfermedadCronica8
EnfermedadCronica9
Nacionalidad1
Nacionalidad2
Req1
Req10
Req2
Req3
Req4
Req5
Req6
Req7
Req8
Req9
Edit


Tasa de unidades fallidas en cada edit
0.010
0.005
0.000
EnfermedadCronica10
EnfermedadCronica11
EnfermedadCronica12
EnfermedadCronica13
EnfermedadCronica14
EnfermedadCronica15
EnfermedadCronica16
EnfermedadCronica17
EnfermedadCronica18
EnfermedadCronica19
EnfermedadCronica20
EnfermedadCronica21
EnfermedadCronica22
EnfermedadCronica23
EnfermedadCronica24
EnfermedadCronica25
EnfermedadCronica26
EnfermedadCronica27
EnfermedadCronica28
EnfermedadCronica29
EnfermedadCronica30
EnfermedadCronica31
EnfermedadCronica32
EnfermedadCronica33
EnfermedadCronica34
Menopausia
Prostata
Vision
Edit


Tasa de unidades fallidas en cada edit
0.3
0.2
0.1
0.0
Audicion
Dentista1
Dentista2
Dentista3
Dependiente1
Dependiente2
Duplicada3
IngresoHospitalario1
IngresoHospitalario2
IngresoHospitalario3
IngresoHospitalario4
IngresoHospitalario5
IngresoHospitalario6
IngresoHospitalario7
IngresoHospitalario8
IngresoHospitalario9
LimiteTemp1
LimiteTemp2
LimiteTemp3
LimiteTemp4
LimiteTemp5
Privacidad1
Privacidad2
PruebaMedica1
TipoSegSan
VisitaMedico
VisitaMedico2
VisitaMedico3
Edit


Tasa de unidades fallidas en cada edit
0.015
0.010
0.005
0.000
Altura
CitologiaHombres
CitologiaV
Colonoscopia
DiasActividadFisica
DiasActividadFisica2
DiasActividadFisica3
edadfumador
Fumadores
FumadoresDia
HorasActividadFisica2
HorasActividadFisica3
HorasActividadFisica4
Mamografia
minutosActividadFisica
minutosActividadFisica2
minutosActividadFisica3
minutosActividadFisica4
Peso
Edit



4.3 ENSE Hogares

- *Flag* por unidad y *edit*. Se exponen únicamente 5 unidades del registro a modo de ejemplo, y se omiten las restantes:

```
##      PersRef2 LimNumPer PersRef Req1 CoheEdad AdultSelec CoheEdad2 CoheEdad3
## [1,]    TRUE     TRUE   TRUE TRUE    TRUE      TRUE     TRUE      TRUE
## [2,]    TRUE     TRUE   TRUE TRUE    TRUE      TRUE     TRUE      TRUE
## [3,]    TRUE     TRUE   TRUE TRUE    TRUE      TRUE     TRUE      TRUE
## [4,]    TRUE     TRUE   TRUE TRUE    TRUE      TRUE     TRUE      TRUE
## [5,]    TRUE     TRUE   TRUE TRUE    TRUE      TRUE     TRUE      TRUE

##      CoheEdad4 CoheEdad5  Req2  HogarSinMenor  TrabajoMenor   HogUni1     HogUni2
## [1,]     TRUE     TRUE   TRUE          TRUE         TRUE     TRUE      FALSE
## [2,]     TRUE     TRUE   TRUE          TRUE         TRUE     TRUE       TRUE
## [3,]     TRUE     TRUE   TRUE          TRUE         TRUE     TRUE       TRUE
## [4,]     TRUE     TRUE   TRUE          TRUE         TRUE     TRUE       TRUE
## [5,]     TRUE     TRUE   TRUE          TRUE         TRUE     TRUE       TRUE

##      HogBiper NoTrab NoTrab2 NoTrab3 CoherPreg1 CoherPreg2 CoherPreg3 CoheEdad6
## [1,]    TRUE   TRUE    TRUE   TRUE      TRUE     TRUE      TRUE      TRUE
## [2,]    TRUE   TRUE    TRUE   TRUE      TRUE     TRUE      TRUE      TRUE
## [3,]    TRUE   TRUE    TRUE   TRUE      TRUE     TRUE      TRUE      TRUE
## [4,]    TRUE   TRUE    TRUE   TRUE      TRUE     TRUE      TRUE      TRUE
## [5,]    TRUE   TRUE    TRUE   TRUE      TRUE     TRUE      TRUE      TRUE

##      CoheEdad7 PadreHijo AdultoNieto PadreHijo2 Ingresos
```

```
## [1,]     FALSE      TRUE        TRUE       TRUE    FALSE
## [2,]     FALSE      TRUE        TRUE       TRUE     TRUE
## [3,]     FALSE      TRUE        TRUE       TRUE     TRUE
## [4,]     FALSE      TRUE        TRUE       TRUE     TRUE
## [5,]     FALSE      TRUE        TRUE       TRUE     TRUE
```

- Número de unidades satisfechas por *edit:*

```
##      PersRef2     LimNumPer       PersRef          Req1    CoheEdad
##         60143         60143         60129         60143       60143
##    AdultSelec     CoheEdad2     CoheEdad3     CoheEdad4   CoheEdad5
##         60143         60143         60059         59438       60143
##          Req2 HogarSinMenor  TrabajoMenor       HogUni1     HogUni2
##         60143         60143         59653         60143       54414
##      HogBiper        NoTrab       NoTrab2       NoTrab3  CoherPreg1
##         49045         60143         60143         60143       60143
##    CoherPreg2    CoherPreg3     CoheEdad6     CoheEdad7   PadreHijo
##         60143         60120         58117         32017       60143
##   AdultoNieto    PadreHijo2      Ingresos
##         60143         60143         40582
```

- Número de unidades fallidas por *Edit:*

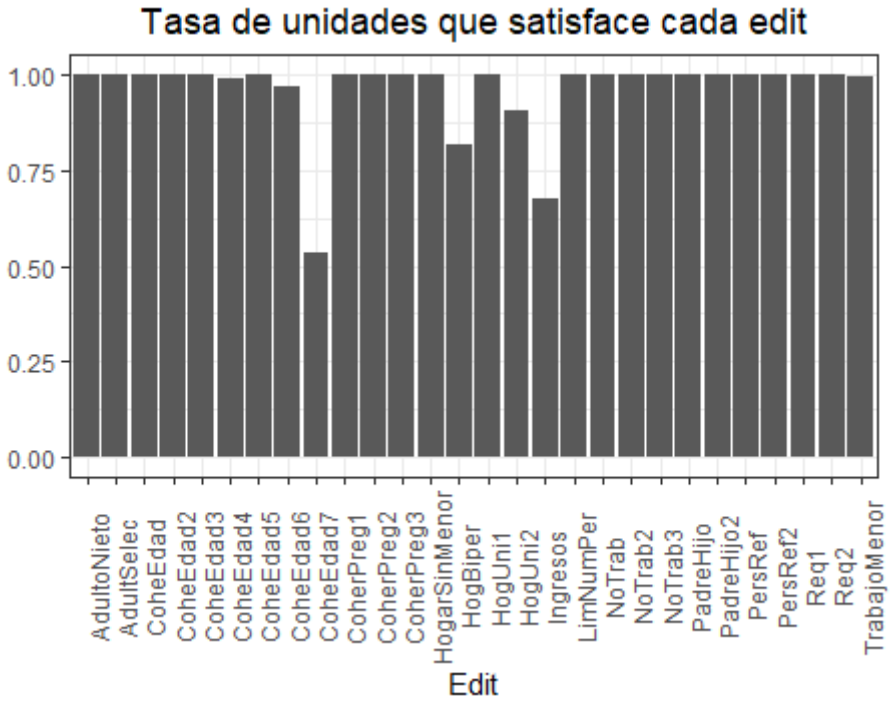
```
##      PersRef2     LimNumPer       PersRef          Req1    CoheEdad
##             0             0            14             0           0
##    AdultSelec     CoheEdad2     CoheEdad3     CoheEdad4   CoheEdad5
##             0             0            84           705           0
##          Req2 HogarSinMenor  TrabajoMenor       HogUni1     HogUni2
##             0             0           490             0        5729
##      HogBiper        NoTrab       NoTrab2       NoTrab3  CoherPreg1
##         11098             0             0             0           0
##    CoherPreg2    CoherPreg3     CoheEdad6     CoheEdad7   PadreHijo
##             0            23          2026         28126           0
##   AdultoNieto    PadreHijo2      Ingresos
##             0             0         19561
```

- Matriz que aúna la información anterior:

```
##                npass nfail nNA     rel.pass         rel.fail  rel.NA    editNames
## PersRef2       60143     0   0    1.0000000   0.0000000000     0      PersRef2
## LimNumPer      60143     0   0    1.0000000   0.0000000000     0     LimNumPer
## PersRef        60129    14   0    0.9997672   0.0002327785     0       PersRef
## Req1           60143     0   0    1.0000000   0.0000000000     0          Req1
## CoheEdad       60143     0   0    1.0000000   0.0000000000     0      CoheEdad
## AdultSelec     60143     0   0    1.0000000   0.0000000000     0    AdultSelec
## CoheEdad2      60143     0   0    1.0000000   0.0000000000     0     CoheEdad2
## CoheEdad3      60059    84   0    0.9986033   0.0013966713     0     CoheEdad3
## CoheEdad4      59438   705   0    0.9882779   0.0117220624     0     CoheEdad4
## CoheEdad5      60143     0   0    1.0000000   0.0000000000     0     CoheEdad5
## Req2           60143     0   0    1.0000000   0.0000000000     0          Req2
## HogarSinMenor 60143   0   0    1.0000000   0.0000000000     0  HogarSinMenor
## TrabajoMenor 59653 490   0    0.9918528   0.0081472491     0   TrabajoMenor
## HogUni1        60143     0   0    1.0000000   0.0000000000     0       HogUni1
```

```
## HogUni2      54414 5729  0  0.9047437  0.0952563058    0     HogUni2
## HogBiper     49045 11098  0  0.8154731  0.1845268776     0  HogBiper
## NoTrab       60143     0  0  1.0000000  0.0000000000     0    NoTrab
## NoTrab2      60143     0  0  1.0000000  0.0000000000     0   NoTrab2
## NoTrab3      60143     0  0  1.0000000  0.0000000000     0   NoTrab3
## CoherPreg1   60143     0  0  1.0000000  0.0000000000     0  CoherPreg1
## CoherPreg2   60143     0  0  1.0000000  0.0000000000     0  CoherPreg2
## CoherPreg3   60120    23  0  0.9996176  0.0003824219     0  CoherPreg3
## CoheEdad6    58117  2026  0  0.9663136  0.0336863808     0  CoheEdad6
## CoheEdad7    32017  28126 0  0.5323479  0.4676520958     0  CoheEdad7
## PadreHijo    60143     0  0  1.0000000  0.0000000000     0 PadreHijo
## AdultoNieto 60143     0  0  1.0000000  0.0000000000     0 AdultoNieto
## PadreHijo2   60143     0  0  1.0000000  0.0000000000     0  PadreHijo2
## Ingresos     40582 19561  0  0.6747585  0.3252415077     0  Ingresos
```

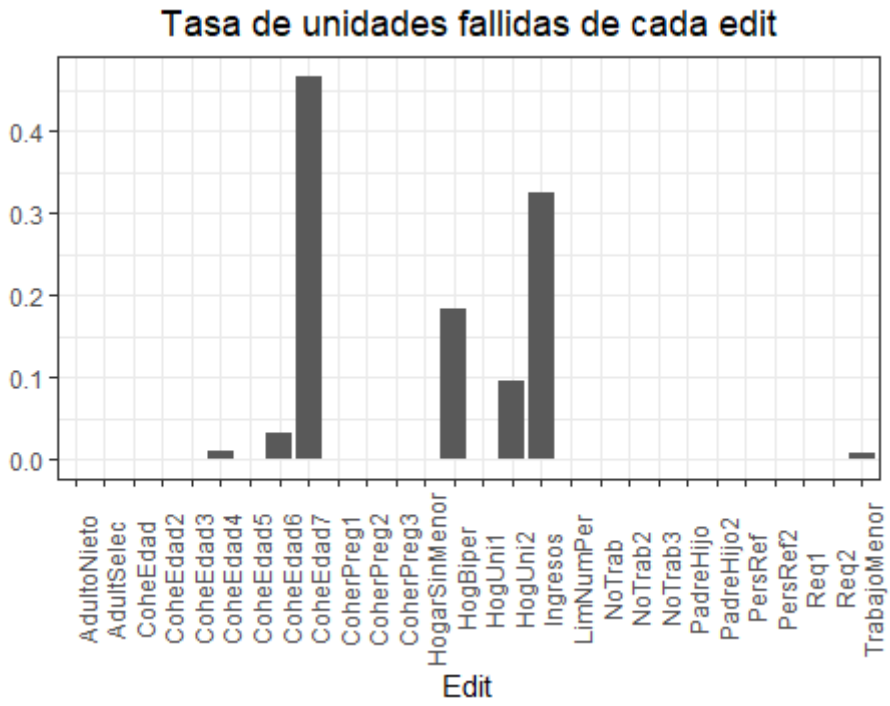
- Histograma de los registros pasados por cada regla y de la tasa de *edits* pasados por cada registro:

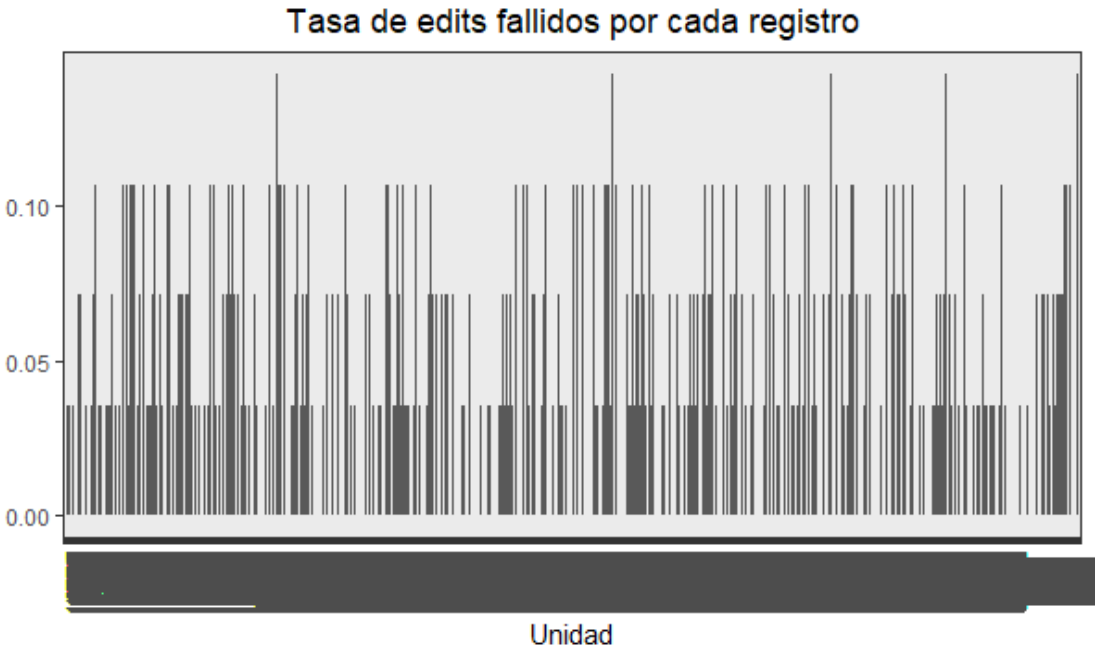


El gráfico presenta la tasa de respuestas correctas de cada *Edit* siendo 1 el máximo. La mayoría de los controles pasan con la tasa máxima lo cual quiere decir que no se incumple ese posible error en ningun caso. Sin embargo, hay controles que no llegan al 75% de acierto, y habría que saber por qué se responden con alta tasa de error esos campos. Los campos *CoheEdad7* e *Ingesos* tienen tasas de acierto muy bajas, lo cual quiere decir que las respuestas entre la situación laboral y la edad no son coherentes y que se indica la cantidad de ingresos o el tipo pero no ambos.

En este caso se puede apreciar como hay muestras que presenta tasas de acierto nulas o prácticamente nulas, lo cual indica que hay usuarios que responden inapropiadamente al test en todas o casi todas sus preguntas o que se realiza un mal registro de dichas repuestas. Esto debe ser tenido en cuenta pues adultera de una manera incontrolable la tasa de error de las preguntas sin que los encargados de los test puedan hacer nada al respecto

- Histograma de los registros fallidos por cada regla y de la tasa de *Edits* fallidos por cada registro:





La información que proporcionan los dos últimos gráficos presentados pretende complementar la información presentada anteriormente en los gráficos que muestran los registros y los *edits* que han pasado los controles, llegando así, a las mismas conclusiones.

5. CONCLUSION

El problema tratado en el presente trabajo se basa en la gran cantidad de recursos que se emplean por parte de las organizaciones estadísticas pertinentes de cada país en el proceso de validación. En este proceso cada OE es tratada de forma independiente por el órgano pertinente, teniendo el INE un total de 140 OEs, cifra la cual se multiplica por 28 si el problema se traslada a nivel europeo.

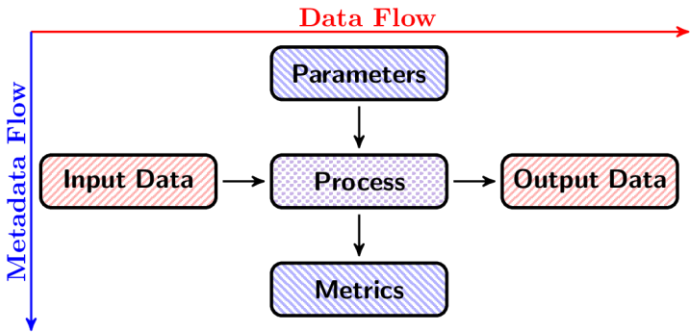
La solución propuesta en el presente trabajo agilizaría el proceso de validación, **estandarizando** el método para las diferentes OEs mediante la creación de un *script* único ligeramente modificable para los tratamientos específicos que cada proceso necesitase. Los controles en el fichero .YAML se crearían de cero la primera vez ocupando un despliegue de recursos mayor del que se instauraría después cuando ya se partiese de la base. Esto significaría a largo plazo una disminución cuantiosa de los recursos empleados, sobre todo en términos de coste de oportunidad medido en unidades temporales de los empleados que producen las estadísticas.

Esta solución también implica la **informatización** del proceso. La depuración es uno de las fases de la producción estadística oficial que aun se aborda de manera manual con una ayuda marginal de la informática en muchas operaciones estadísticas. La informatización de esta fase a través de una metodología sencilla y estándar, pero mas compleja que la manual permite ganar en trazabilidad y objetividad. Y la detección de fallos en los controles y en las respuestas permite una mejora continua de las preguntas y de los *Edits* generados para reducir el número de errores.

El fin de la informatización de los procesos debe ser la **automatización** del proceso. Esta lleva asociado un rendimiento muy alto a largo plazo en cuestión de ahorrar tiempo, aunque al principio puede llevar asociado un coste alto de recursos, pues a pesar de ser R una herramienta gratuita, se ha de incurrir en gastos de formación en R, de mantenimiento de las herramientas informáticas y su soporte, etc.

Partiendo de estos requisitos, el presente trabajo tenía como objetivo mostrar que es posible crear un programa común que sirviese para realizar la validación de datos pertinente a encuestas de diferente ámbito. Las OEs que se han usado para mostrar eso han sido el IASS y el ENSE, el programa común ha sido el *Script*, que como se puede observar es muy simple y prácticamente igual para ambas encuestas. Para los controles, que dependen de las características propias de cada encuesta, se crean los *edits* en formato .YAML con la metodología que marca el paquete *validate* que es el mismo que nos permite analizar los resultados con unos comandos muy sencillos.

En definitiva, se ha mostrado la informatización de este proceso siguiendo los estándares internacionales representados en la siguiente figura:



Es por ello que, como conclusión y usando como plantilla cualquiera de las creadas en el presente trabajo se podría adaptar al resto de Operaciones Estadísticas, adaptando los controles pertinentes de cada uno.

Reflexión personal: Existe un debate importante respecto al perfil profesional del estadístico oficial ante la modernización de la producción. Se reconocen tres perfiles extremos puros: (a) metodólogo/estadístico, (b) computer scientist/ingeniero informático y (c) business manager/científico social. Lo ideal es que el personal que trabaja produciendo estadísticas oficiales tenga suficientes conocimientos de cada área tanto para mejorar la comunicación interna como para el diseño, implementación, ejecución y monitorización de todos los pasos de producción. Este TFG constituye un ejemplo donde se conjuga la metodología estadística (expresión de reglas de validación), la informática (creación de los ficheros yaml y los scripts, así como su uso) y las ciencias sociales (los profesionales con conocimientos en economía, demografía, sociología, etc. suelen ser los expertos en la materia de cada operación). He mostrado como un potencial subject matter expert (como economista) puede construir reglas de validación de acuerdo con la metodología estándar e implementarlas para su uso informáticamente.

6. BIBLIOGRAFÍA

CNAE, 2009. *https://www.cnae.com.es/*

CNO, 2011. *https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177033&menu=ultiDatos&idp=1254735976614*

Conference of European Statisticians. Geneva, 14-16 June, 2011. *High-Level Group for the Modernisation of Statistical Production. Strategic vision of the High-Level Group for strategic developments in business architecture in Statistics.*

INE IASS. https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176783&menu=resultados&idp=1254735573175#!tabs-1254736195650

INE ENSE. https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176863&menu=ultiDatos&idp=1254735576778

Lebied, M. 2018. *The Ultimate Guide to Modern Data Quality Management (DQM) For An Effective Data Quality Control Driven by The Right Metrics.*

Lethbridge, T.C., Sim, S.E. & Singer, J., 2005. *Studying Software Engineers: Data Collection Techniques for Software Field Studies.*

M. van der Loo and E. de Jonge (2018). *Statistical Data Cleaning with R. Wiley.*

R Development Core Team, 2000. *Introducción a R. Notas sobre R: Un entorno de programación para Análisis de Datos y Gráficos. Versión 1.0.1*

R documentation FastReadFWF. *https://www.rdocumentation.org/packages/sss/versions/0.1-0/topics/fast.read.fwf*

R documentation ggplot2. *https://www.rdocumentation.org/packages/ggplot2/versions/3.3.1*

R documentation validate. *https://www.rdocumentation.org/packages/validate/versions/0.9.3*

R documentation dyplr. *https://www.rdocumentation.org/packages/dplyr/versions/0.7.8*

R documentation stringr. *https://www.rdocumentation.org/packages/stringr/versions/1.4.0*

R documentation stringi. *https://www.rdocumentation.org/packages/stringr/versions/1.4.0*

R documentation data.table. *https://www.rdocumentation.org/packages/data.table/versions/1.12.8*

Statistic Software, 2019. https://github.com/SNStatComp/awesome-official-statistics-software

Técnicas de Validación de Datos. *https://www.tecnologias-informacion.com/validacion.html*

T. de Waal, J. Pannekoek, and S. Scholtus, 2011. *Handbook of statistical data editing and imputation. Wiley.*

UNECE. Generic Statistical Business Process Model v5.0. UNECE, 2013. *Available at http://www1.unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model (accessed on May 19, 2016).*

UNECE. Generic Statistical Information Model v1.1. UNECE, 2013. *Available at http://www1.unece.org/stat/platform/display/metis/Generic+Statistical+Information+Model (accessed on May 19, 2016).*

UNECE. Common Statistical Production Arquitecture v1.5. UNECE, 2013. *Available at http://www1.unece.org/stat/platform/display/CSPA/Common+Statistical+Production+Architecture+Home (accessed on May 19, 2016).*

UNECE. Generic Activity Model for Statistical Organizations v1.0. UNECE, 2015. *Available at http://www1.unece.org/stat/platform/display/GAMSO/GAMSO+Home (accessed on May 19, 2016).*

UNECE. Generic Statistical Data Editing Models v1.0. UNECE, 2015. *Available at http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=117774163 (accessed on May 19, 2016).*

U.S. Federal Committee on Statistical Methodology, 1990. *Data Editing in Federal Statistical Agencies. Technical report, Statistical Policy Office: U.S. Office ofManagement and Budget, Washington, D.C.*

W. Gilley. Jerry, 1990 *How to Collect Data.*

.YAML. *https://cran.r-project.org/web/packages/validate/vignettes/rule_files.html*

7. ANEXOS

ANEXO I. CUESTIONARIO DEL IASS

Este documento muestra en formato .pdf el cuestionario que rellenan las empresas de la muestra poblacional del IASS.

https://github.com/CarlosAA5/IASS_ENSE_DATAVAL/blob/master/IASS-17.pdf

ANEXO II. EDITS DEL IASS

Este documento muestra en formato ..YAML los controles anteriormente diseñados, en un formato leíble por la herramienta de programación R-Studio, concretamente por el paquete *validate* de este.

https://github.com/CarlosAA5/IASS_ENSE_DATAVAL/blob/master/EditsIASS_corr..YAML

ANEXO III. DATOS DEL IASS

Este documento contiene la simulación de los datos que se recogen en la encuesta IASS en formato .rds para ser leídos por R-Studio. Los datos están en tres archivos. El primero contiene los datos simulados desde febrero de 2017 a enero de 2018, el segundo los datos simulados de febrero de 2018 y el último el valor de los límites tanto inferior como superior para cada variable.

https://github.com/CarlosAA5/IASS_ENSE_DATAVAL/blob/master/dataHistory.rds

https://github.com/CarlosAA5/IASS_ENSE_DATAVAL/blob/master/dataMM022018.rds

https://github.com/CarlosAA5/IASS_ENSE_DATAVAL/blob/master/dataMM022020_Intervalos.dt

ANEXO IV. INFORMACIÓN SOBRE LA ENCUESTA ENSE ADULTO

Este documento en formato .xlsx contiene información acerca de las variables contenidas en el cuestionario ENSE de adulto.

- En la primera pestaña se encuentra un esquema con los nombres de las variables, la longitud de las respuestas de las variables, la posición inicial y final sobre el total, el tipo de variable y una breve descripción de la variable.
- En la segunda pestaña se muestra la misma información que en la primera, pero clasificadas por el tipo de pregunta y por quien la formula (INE o el Módulo Europeo de Estado de Salud).
- En la tercera pestaña se definen de nuevo las variables y se establecen las posibles respuestas a dichas preguntas.

- En la cuarta pestaña se muestra la Clasificación Nacional de Actividades Económicas (CNAE) del año 2009 (CNAE, 2009) que otorga un código a la actividad económica desarrollada por la persona de referencia.
- En la quinta pestaña se muestra la Clasificación Nacional de Ocupaciones (CNO, 2011) que otorga un código de 3 dígitos según la ocupación de la persona de referencia.

https://github.com/CarlosAA5/IASS_ENSE_DATAVAL/blob/master/stENSE2017Adulto_Schema.xlsx

ANEXO V. CONTROLES ENSE ADULTO

Este documento muestra en formato .xlsx los controles diseñados por el desarrollador del presente trabajo. Estos controles han sido diseñados siguiendo un criterio experto, definido por una persona no profesional de este sector.

https://github.com/CarlosAA5/IASS_ENSE_DATAVAL/blob/master/EditsPropuestos_Salud_ADULTO.xlsx

ANEXO VI. EDITS ENSE ADULTO

Este documento muestra en formato ..YAML los controles anteriormente diseñados, en un formato leíble por la herramienta de programación R-Studio, concretamente por el paquete *validate* de este.

https://github.com/CarlosAA5/IASS_ENSE_DATAVAL/blob/master/EditHogIndv..YAML

ANEXO VII. DATOS ENSE ADULTO

Este documento contiene la simulación de los datos que se recogen en la encuesta ENSE Adulto en formato .rds para ser leídos por R-Studio.

https://github.com/CarlosAA5/IASS_ENSE_DATAVAL/blob/master/MICRODAT.CA.txt

ANEXO VIII. INFORMACIÓN SOBRE LA ENCUESTA ENSE HOGAR

Este documento en formato .xlsx contiene información acerca de las variables contenidas en la encuesta ENSE de adulto.

- En la primera pestaña se encuentra un esquema con los nombres de las variables, la longitud de las respuestas de las variables, la posición inicial y final sobre el total, el tipo de variable y una breve descripción de la variable.

- En la segunda pestaña se muestra la misma información que en la primera, pero clasificadas por el tipo de pregunta y por quien la formula (INE o el Módulo Europeo de Estado de Salud).
- En la tercera pestaña se definen de nuevo las variables y se establecen las posibles respuestas a dichas preguntas.
- En la cuarta pestaña se muestra la Clasificación Nacional de Actividades Económicas (CNAE) del año 2009 (CNAE, 2009) que otorga un código a la actividad económica desarrollada por la persona de referencia.
- En la quinta pestaña se muestra el código CNO 2011 que otorga un código 3 dígitos según la ocupación de la persona de referencia (CNO, 2011).

https://github.com/CarlosAA5/IASS_ENSE_DATAVAL/blob/master/EditsPropuestos_Salud_HOGAR.xlsb.xlsx

ANEXO IX. CONTROLES ENSE HOGAR

Este documento muestra en formato .xlsx los controles diseñados por el desarrollador del presente trabajo. Estos controles han sido diseñados siguiendo un criterio experto, definido por una persona no profesional de este sector.

https://github.com/CarlosAA5/IASS_ENSE_DATAVAL/blob/master/EditsPropuestos_Salud_HOGAR.xlsb.xlsx

ANEXO X. EDITS ENSE HOGAR

Este documento muestra en formato .YAML los controles anteriormente diseñados, en un formato leíble por la herramienta de programación R-Studio, concretamente por el paquete *validate* de este.

https://github.com/CarlosAA5/IASS_ENSE_DATAVAL/blob/master/EditHogIndv..YAML

ANEXO XI. DATOS ENSE HOGAR

Este documento contiene la simulación de los datos que se recogen en la encuesta ENSE Hogar en formato .rds para ser leídos por R-Studio.

https://github.com/CarlosAA5/IASS_ENSE_DATAVAL/blob/master/MICRODAT.CH.txt