Departamento de Economía

TESIS DOCTORAL

Métodos estadísticos de depuración e imputación de datos

Pedro Revilla Novella

2013

DEPARTAMENTO DE ECONOMÍA
Plaza de la Victoria
28802 Alcalá de Henares (Madrid)
Teléfonos :91 8854201/ 4202/ 5154
dpto.economia@uah.es

Prof. Dr. D. José Miguel Casas Sánchez, Profesor de la Universidad de Alcalá, como director de la tesis titulada "Métodos estadísticos de depuración e imputación de datos" realizada por D. Pedro Revilla Novella dentro del programa de Economía Aplicada del antiguo departamento de Estadística, Estructura Económica y Organización Económica Internacional de la Universidad de Alcalá.

EXPONE: que la citada tesis doctoral, en mi opinión, reúne todas las condiciones necesarias para el inicio de los tramites destinados a su defensa pública por parte del doctorando. Constituye un trabajo de investigación de ámbito académico, por lo que

AUTORIZA la presentación de la citada tesis doctoral para efectuar los trámites necesarios conducentes a la defensa pública de la misma.

En Alcalá de Henares, a dieciséis de julio de dos mil trece

Firmado: Prof. Dr. D. José Miguel Casas Sánchez

DEPARTAMENTO DE ECONOMÍA
Plaza de la Victoria
28802 Alcalá de Henares (Madrid)
Teléfonos :91 8854201/ 4202/ 5154
dpto.economia@uah.es

D. Antonio García Tabuenca, director del departamento de Economía de la Universidad de Alcalá autoriza, después de su aprobación por el departamento en su reunión de 17 de julio de 2013 la presentación de la tesis titulada "Métodos estadísticos de depuración e imputación de datos" realizada por D. Pedro Revilla Novella dentro del programa de Economía Aplicada del antiguo departamento de Estadística, Estructura Económica y Organización Económica Internacional de la Universidad de Alcalá, una vez efectuados los trámites oportunos, su defensa ante el tribunal correspondiente.

En Alcalá de henares, a diecisiete de julio de dos mil trece.

Fdo.: Dr. Antonio García Tabuenca

UNIVERSIDAD DE ALCALÁ

FACULTAD DE CIENCIAS ECONÓMICAS, EMPRESARIALES Y TURISMO

DEPARTAMENTO DE ECONOMÍA

TESIS DOCTORAL

Métodos estadísticos de depuración e imputación de datos

Pedro Revilla Novella

Director : Prof. Dr. D. José Miguel Casas Sánchez

Alcalá de Henares, Noviembre de 2013

A mis padres

Agradecimientos

En primer lugar, quiero manifestar mi agradecimiento al profesor Dr. D. José Miguel Casas Sánchez por su generosidad al aceptar, desde un primer momento, dirigir mi tesis con su sólida experiencia y prestigio profesional, en un campo de investigación todavía no demasiado estudiado en el ámbito académico como es la depuración e imputación de datos estadísticos, y por brindarme su continuo apoyo académico y personal.

Quería también expresar mi deuda con mis compañeros de investigaciones, de los que aprendo todos los días y además es un placer trabajar, en particular Ignacio Arbués, Pilar Rey y David Salgado.

Asimismo, me gustaría mencionar a los colegas del grupo de depuración de las Naciones Unidas, que me han mostrado muchas buenas ideas a lo largo de estos años, en especial a Leopold Granquist y John Kovar.

Índice

página

“Que errar es humano no sólo significa que hemos de luchar constantemente contra el error, sino también que, aún cuando hayamos puesto el máximo cuidado, no podemos estar totalmente seguros de no haber cometido un error”

Karl Popper

Capítulo 1

Introducción

Esta tesis consiste en un conjunto de investigaciones desarrolladas en el campo de la depuración e imputación de datos. En concreto, se centra en dos principales líneas de investigación interconectadas entre sí: la utilización de modelos estadísticos y de técnicas de optimización.

Las oficinas de estadística y otros organismos y empresas recogen y procesan grandes cantidades de datos con el objetivo de producir información estadística. Una parte importante de este proceso de producción consiste en chequear los datos, corregir los errores y dar un tratamiento adecuado a la falta de respuesta. La depuración e imputación de datos es una fase fundamental en el proceso de producción de información estadística. Sin esta fase, la calidad de las estimaciones finales podría reducirse significativamente, y la credibilidad de las encuestas verse seriamente dañada. Por otra parte, constituye una de las tareas más costosas en el proceso de producción estadística. Diversos estudios muestran que suponen por término medio el 20% del coste total de las encuestas a hogares y el 40% del de las encuestas a empresas.

La depuración e imputación se centra en la detección y tratamiento de los errores ajenos al muestreo. Estos incluyen un amplio conjunto, como errores de no-respuesta, de medida, sistemáticos, aleatorios, influyentes, outliers, inliers, de unidad de medida, de redondeo, etc. A su vez, esta diversidad ha dado lugar a diferentes técnicas y algoritmos, tales como microdepuración, macrodepuración, depuración selectiva, depuración interactiva, o depuración automática. Una revisión reciente puede verse en de Waal et al. (2011). La importancia de este tema se refleja en la organización por parte de Naciones Unidas de un seminario periódico (*Work Session on Statistical Data Editing)*. El mismo organismo, ha publicado manuales donde se recogen contribuciones en este campo (*Statistical Data Editing*, 1994, 1997 y 2006).

El paradigma de Fellegi-Holt (1976) proporciona la base para la construcción de sistemas generalizados. En estos sistemas las reglas están estandarizadas, programadas y contrastadas y pueden aplicarse a encuestas diferentes. Habitualmente, realizan las siguientes funciones: análisis de los edits para chequear su consistencia, detección de errores, identificación de las variables a imputar e imputación. Sin embargo, no resuelven la especificación de los edits, tarea que se considera un input del sistema y que debe ser realizada en cada caso por los usuarios. Tradicionalmente, los usuarios establecen los edits de acuerdo a su experiencia práctica, sin que exista un marco teórico para los mismos. En este trabajo se propone un marco conceptual para tratar la especificación de los edits y se diseñan herramientas concretas de depuración e imputación de datos, basándose en modelos estadísticos. Al mismo tiempo, en busca de una mayor eficciencia en la depuración, se investiga la forma de resolver el problema de la depuración selectiva, mediante la utilización de técnicas de optimización.

Esta tesis se estructura en cuatro capítulos. En el capítulo 1 se realiza la introducción. El capítulo 2, “El problema de la depuración e imputación. Panorámica actual de la

depuración e imputación", plantea el contexto donde se encuadran las investigaciones y presenta el estado del arte en este campo.

El capítulo 3, "Planteamiento teórico y desarrollo empírico de nuevas técnicas de depuración e imputación basadas en modelos estadísticos y en técnicas de optimización", constituye la parte central de esta tesis y en él se describen las investigaciones llevadas a cabo y sus principales resultados y aportaciones. En el subcapítulo 3.1 "Depuración e imputación basada en modelos de series temporales", están presentes ya las principales líneas de investigación de esta tesis, restringidas al caso de modelos de series temporales y encuestas continuas con datos cuantitativos. Los modelos utilizados, tanto para los microdatos como para los macrodatos son modelos RegARIMA. En las encuestas continuas la principal información para efectuar la depuración e imputación son los datos pasados de la misma encuesta. La depuración e imputación histórica tradicional (tasas de variación, etc.) puede mejorarse si se utilizan modelos de series temporales. A partir de estos modelos pueden determinarse los edits, basados en intervalos de confianza, tanto para la microdepuración como para la macrodepuración. Los datos sospechosos de error son aquellos que se salen de los intervalos. Para llevar a cabo la depuración selectiva se han desarrollado un conjunto de herramientas (que hemos denominado sorpresa, sorpresa estándar, sorpresa estándar ponderada e influencias) basadas en la función de predicción un periodo por delante del modelo. En algunos casos estas herramientas constituyen una función score, utilizadas ampliamente en la literatura (Berthelot y Latouche, 1993). Una aportación de esta tesis consiste en que la función score propuesta se obtiene a partir de los modelos y no por razones de tipo heurístico. Adicionalmente, a partir de los modelos anteriormente mencionados, se ha obtenido información de utilidad para llevar a cabo la macrodepuración, siguiendo la línea del análisis exploratorio de datos (Tukey, 1977).

El subcapítulo 3.2, "La depuración selectiva como un problema de optimización estocástica", se centra en el enfoque de la depuración selectiva. Disponer de métodos eficientes de depuración es un objetivo fundamental, ya que la depuración manual exhaustiva es considerada poco eficiente, puesto que la mayor parte del trabajo de depuración no tiene consecuencias a nivel agregado e incluso puede llegar a deteriorar la calidad de los datos. Los métodos de depuración selectiva son estrategias para seleccionar un subconjunto de los cuestionarios recogidos en una encuesta para someterlos a una depuración minuciosa. Una razón por la que es conveniente seleccionar algunos cuestionarios es que depurando ciertas unidades es más probable que mejore la calidad que si se depuran otras. Esto puede ser debido a que algunas son más sospechosas de contener un error o a que, en caso de tenerlo, probablemente tenga más impacto en los agregados. Una buena estrategia de depuración selectiva persigue compatibilizar dos objetivos: buena calidad de las estimaciones agregadas y reducido trabajo de depuración. El enfoque actualmente predominante es el de calcular algún tipo de función score (FS) que asigne una puntuación a cada unidad, priorizando la depuración de las unidades con mayor puntuación. Cuando para cada unidad se recogen varias variables, se pueden calcular diferentes FS locales y combinarlas en una global. Así, las unidades cuya FS excede un cierto umbral, son depuradas manualmente. Hasta ahora, estas cuestiones han sido tratadas de manera empírica. En Lawrence y McKenzie (2000) se proponen algunas directrices, pero en esencia se depende del criterio del experto. La mayor aportación de la tesis en este terreno es la resolución del problema de depuración selectiva a través de técnicas de optimización estocástica. En este subcapítulo, se plantea y se resuelve formalmente el problema. Posteriormente, se

propone un método para calcular ciertos momentos condicionales que se requieren para obtener la solución. También se presentan los resultados de una aplicación práctica.

Finalmente, en el subcapítulo 3.3 “Desarrollo de un marco teórico de depuración e imputación basado en modelos y optimización”, se intenta alcanzar un mayor grado de formalización y abstracción y se desarrolla un marco teórico general que pueda ayudar a resolver distintos problemas. Hoy en día se acepta ampliamente que ninguna técnica o algoritmo puede hacer frente a todo tipo de errores. Por lo tanto, deben ser convenientemente combinados en una estrategia global de depuración e imputación. Dentro de este contexto general, este trabajo se centra en dos aspectos principales: el uso de modelos estadísticos para resolver los problemas de la especificación de los edits y aportar un método de imputación, y la utilización de técnicas de optimización para resolver los problemas que presenta la depuración selectiva. Los sistemas generalizados no resuelven la especificación de los edits, y habitualmente los usuarios establecen los edits de acuerdo a su experiencia práctica, sin que exista un marco teórico para los mismos. En este trabajo se aborda la forma en que pueden introducirse modelos estadísticos que relacionan los valores verdaderos, los observados y los depurados construidos a partir de distintos tipos de información disponible.

La depuración selectiva se centra en los errores influyentes. En las últimas dos décadas, esta modalidad de depuración ha sido reconocida como un elemento clave en las estrategias de depuración e imputación. Sin embargo, hasta la fecha, se ha tratado de forma heurística. En el subcapítulo anterior se ha intentado dar un marco formal a la depuración selectiva, mediante técnicas de optimización estocástica. En este, se pretende ampliar el marco teórico a un problema de optimización general, del que pueden deducirse dos versiones: un problema de optimización combinatoria y el mencionado problema de optimización estocástica. Se parte de dos principios generales para abordar una depuración selectiva: se debe minimizar la cantidad de recursos utilizados y garantizar la calidad de los datos. Se propone una traducción matemática de estos principios a un problema general de optimización, cuya solución es la selección de unidades a depurar. En nuestra formulación, los recursos utilizados son equivalentes al número de cuestionarios seleccionados para depuración, mientras que la calidad de los datos se refleja en la precisión de los estimadores. Por tanto se toma como objetivo minimizar el número de unidades seleccionadas para depuración, sujetas a restricciones sobre el valor de las funciones de pérdida. Estas funciones se pueden dirigir al sesgo, al error cuadrático medio, a la varianza o a cualquier otra medida de incertidumbre en la estimación, lo que supone una ampliación respecto al subcapítulo anterior, donde sólo se consideraba el error cuadrático. Pueden ser de naturaleza heurística, tales como las medidas relacionadas con el llamado pseudo-sesgo utilizadas tradicionalmente para las funciones score, o pueden derivarse explícitamente a partir de modelos de medición del error.

Las dos versiones del problema de optimización anteriormente citadas corresponden a los dos escenarios típicos para la aplicación de la depuración selectiva. En el primer caso, la selección se lleva a cabo unidad por unidad, de tal manera que la selección de una unidad no depende de la de otras unidades. Este modo es adecuado cuando la selección puede hacerse en tiempo real a la llegada de cada cuestionario. Nos referimos a este caso como optimización estocástica, debido a que la obtención en tiempo real de la solución sólo puede establecerse con respecto a hipotéticas repeticiones del proceso de selección. En el segundo caso, la selección se lleva a cabo de manera conjunta para

todas (o un grupo de) las unidades. Este modo es adecuado en una etapa posterior de la recogida de datos, cuando ya se dispone de un número suficiente de unidades. Nos referimos a este caso como el problema de optimización combinatoria, donde la solución se puede establecer condicionada a las observaciones reales de la muestra bajo algún modelo especificado de medición del error. Se presenta la forma en que puede resolverse el problema de optimización, para lo que es necesario introducir un modelo multivariante. También se realiza una comparación con funciones score ampliamente utilizadas, usando datos reales

Finalmente, en el Capítulo 4, “Conclusiones”, se realizan comentarios finales y se perfilan algunas futuras líneas de investigación.

Capítulo 2

El problema de la depuración e imputación. Panorámica actual

2.1 El Problema de la depuración e imputación

Las oficinas de estadística y otros organismos y empresas recogen y procesan grandes cantidades de datos con el objetivo de producir información estadística. Una parte importante de este proceso de producción consiste en chequear los datos recogidos, corregir los errores encontrados y dar un tratamiento adecuado a la falta de respuesta. La depuración incide en varias dimensiones de la calidad de la encuesta como la precisión, la puntualidad, la carga de respuesta o la eficiencia. Comprende la detección y el tratamiento de los errores ajenos al muestreo, sobre todo los de falta de respuesta y de medición. .

Los errores de muestreo pueden ser más fácilmente controlados y evaluados Existe una abundante literatura que permite contar con un marco teórico abundante y sistemático, que posibilita controlar los errores de muestreo en la fase de diseño muestral y posteriormente evaluarlos una vez finalizada la encuesta. Entre las muchas referencias pueden citarse Cochran (1977), Särndal et al. (1992), Hidiroglou y Srinath (1993), Hidiroglou (1994), Valliant et al. (2000), etc. Contrariamente, los errores ajenos al muestreo encuentran una mayor dificultad en su tratamiento, por su propia definición como categoría residual. Para tratar de encarar esta complejidad, se ha tratado de clasificar los errores ajenos al muestreo en un determinado número de categorías, véase por ejemplo Cochran (1977), Biemer y Fecso (1995) o, más recientemente, Bethlehem (2009). Se ha desarrollado una tipología de errores, incluyendo errores sistemáticos, errores aleatorios, errores influyentes, outliers, inliers, o valores faltantes, por no hablar de los errores particulares dentro de estas clases como los errores de unidad de medida o errores de redondeo. Esta diversidad ha dado lugar a la aparición de diferentes técnicas y algoritmos para detectar y tratarlos, como la depuración interactiva, la depuración automática, la depuración selectiva, y la macrodepuración (véase De Waal et al. (2011) para una revisión completa). Hoy en día se acepta ampliamente que ninguna técnica puede hacer frente a todo tipo de errores. Por lo tanto deben ser convenientemente combinados en una estrategia de depuración e imputación (E & I en adelante),

Por otra parte, a pesar de los esfuerzos realizados para automatizarla, y de los instrumentos que proporcionan las nuevas tecnologías, sigue siendo una fase que consume tiempo e intensiva en intervención humana. Esta es una de las razones por las cuales la depuración e imputación se convierte en una de las fases más caras del proceso producción de estadísticas. El uso de ordenadores no ha reducido significativamente su coste (Granquist, 1995).

Por su parte, la falta de repuesta está también presente en las encuestas. Habitualmente se clasifica en falta de respuesta total (*unit nonresponse*), cuando se carece de información de todas las preguntas del cuestionario, y falta de respuesta parcial (*item*

nonresponse), cuando no se dispone de información de tan solo alguna de las preguntas. Por otra parte, la falta de respuesta se clasifica, (por ejemplo, en Rubin, 1987), en falta de respuesta aleatoria y no aleatoria. En general se acepta que en la mayoría de los casos la falta de respuesta no se produce de forma aleatoria y las unidades que no responden tienen un comportamiento diferente respecto a las características de interés: (las pequeñas empresas tienen peores registros contables,etc.). A su vez, la falta de respuesta no aleatoria puede clasificarse en falta de respuesta no aleatoria ignorable y falta de respuesta no aleatoria no ignorable. La falta de respuesta ignorable se produce cuando el mecanismo de la falta de respuesta es independiente del nivel de la variable de la que no se ha obtenido información pero depende en cambio de otras variables de las que sí se tiene información. Por el contrario, la falta de respuesta es no ignorable cuando el mecanismo de la falta de respuesta depende del nivel de la variable de la que no se ha obtenido respuesta. La falta de respuesta representa un grave problema para una encuesta, ya que no sólo supone una disminución del tamaño muestral, sino que suele sesgar las estimaciones, ya que las unidades que no responden rara vez constituyen una submuestra aleatoria de la muestra total.

Los dos principales procedimientos para el tratamiento de la falta de respuesta son los métodos de reponderación y de imputación. La reponderación consiste en aumentar los pesos de las unidades que responden y se utiliza principalmente para el tratamiento de la falta de respuesta total. La imputación consiste en reemplazar los datos que faltan por valores “plausibles” y se utiliza fundamentalmente para el tratamiento de la falta de respuesta parcial. El método de reponderación puede, teóricamente, utilizarse también para el tratamiento de la falta de respuesta parcial, aplicándose variable a variable. Sin embargo, en la práctica, el procedimiento se hace complicado, ya que exigiría usar diferentes pesos para cada variable, razón por la cual se prefiere normalmente el método de la imputación. De hecho, en encuestas a empresas se utiliza a veces la imputación para la falta de respuesta total, cuando existe información auxiliar de otras fuentes o de encuestas anteriores.

Finalmente, respecto a los outliers o valores atípicos, tienen, igual que los errores y la falta de respuesta, una importante incidencia en la elaboración de datos. Los outliers afectan fundamentalmente en la depuración y en la estimación. El tema de los outliers está presente en prácticamente todas las ramas de la inferencia estadística. Generalmente, en disciplinas estadísticas diferentes al muestreo de poblaciones finitas, se considera que la muestra ha sido generada a partir de un modelo o población que sigue una cierta distribución paramétrica. En este contexto, se considera que los outliers han sido generados a partir de una fuente diferente a dicho modelo o población. Sin embargo, dentro del marco de muestreo de poblaciones finitas basado en diseño, las muestras se seleccionan a partir de poblaciones finitas que son fijas, y los outliers se consideran que son valores legítimos que proceden de la población en estudio. No obstante, son valores extremos, que se encuentran alejados del núcleo central o de la mayoría de los datos. Existe otro tipo de valores que no son extremos, pero que su inclusión o exclusión puede afectar grandemente a las estimaciones. A este tipo de valores se les llama valores influyentes.

Según comparten varios autores (por ejemplo, Gambino 1987, Srinath 1987, y Bruce 1991) la distinción entre valores extremos y valores influyentes es muy útil en las encuestas. La influencia de una observación depende del estimador que se está utilizando. Por ejemplo, una observación puede ser influyente para un estimador de

expansión y no serlo para uno de la razón y viceversa. De acuerdo a esto, una observación influyente se debe definir respecto a un estimador particular

Algunos autores, por ejemplo, Chambers (1986), clasifican los outliers en dos tipos: outliers representativos y outliers no representativos. Los representativos son los que se obtienen sin que exista error en los datos y representan a otras unidades similares en la población. Los outliers no representativos son los que proceden de un error en los datos o son únicos en el sentido de que no existe otra unidad como ellos. Existe una amplia literatura acerca de los outliers en casos paramétricos o en poblaciones infinitas. Por ejemplo, Hawkins (1980) y Barnett y Lewis (1984) describen la detección y tratamiento de outliers para muestras que proceden de poblaciones con distribuciones paramétricas. Belsley et al. (1980) y Cook y Weisberg (1982) revisan métodos para el análisis de regresión. Beckman y Cook (1983) proporcionan una revisión histórica de los outliers incluyendo métodos bayesianos y de regresión robusta. Por su parte, Huber (1981), Hampel et al. (1986), y Rousseeuw y Leroy (1987) presentan distintos aspectos de la teoría de estimación robusta. Por el contrario, se ha escrito mucho menos acerca de los outliers en muestras de poblaciones finitas. En algunas ocasiones, las encuestas que se basan en el muestreo de poblaciones finitas pueden usar métodos que provienen de otras ramas de la estadística. Normalmente puede considerarse que el estudio de los outliers contempla dos aspectos: su detección y su tratamiento. El problema de los outliers en la estimación es esencialmente un problema de estimación robusta. La estimación robusta puede ser abordada mediante el tratamiento de los outliers detectados o por la directa aplicación de técnicas de estimación robustas como el M-estimador.

La detección de outliers tiene interés en la depuración, ya que pueden ser legítimos valores de la población pero pueden ser también errores en los datos que deben ser corregidos antes de la estimación. Tradicionalmente, los outliers son detectados usando sus distancias relativas al centro de los datos. El tratamiento de outliers en la etapa de estimación encuentra un conjunto de procedimientos con una base formal. Existen dos enfoques principales para el tratamiento de los outliers en poblaciones finitas. Un primer conjunto de métodos está basado en la sustitución de los valores outliers o en la reducción de sus pesos. La sustitución de los outliers puede llevarse a cabo mediante los métodos de recorte o “winsorización”. Una alternativa a sustituir los valores que se consideran outliers consiste en reducir sus pesos. Una primera versión de este procedimiento se encuentra en Bershad (1960), tras la cual muchos de estos estimadores han sido propuestos. En esta línea pueden destacarse las aportaciones de Searls (1966), Ernst (1980), Hidiroglou y Srinath (1981), Dalén (1987), Tambay (1988), Ghangurde (1989a, 1989b), Fuller (1991), Hidiroglou (1991), Rivest y Hurtubise (1993), Rivest (1993a, 1993b), y Thorburn (1993). Un segundo conjunto de métodos consiste en usar métodos de estimación robustos como la M-estimación. Huber (1964) introdujo el M-estimador como una alternativa robusta a la media muestral para una distribución que sigue una normal en el centro y una doble exponencial en las colas. Desde la publicación de Huber muchos M-estimadores han sido propuestos y estudiados. En esta línea puede verse, por ejemplo, Andrews et al. (1972). Más recientemente, Chambers (1986), Hampel et al. (1986), Lee (1990, 1991a), Bruce (1991), Gwet y Rivest (1992), y Hulliger (1993) realizan importantes contribuciones en esta materia.

2.2 Panorámica actual

Para afrontar la depuración e imputación existen actualmente multitud de técnicas. Estas técnicas utilizan herramientas teóricas o provienen de la experiencia práctica. Por otra parte, las técnicas pueden haberse diseñado de forma ad hoc para una encuesta determinada o pretender aplicarse de forma general. Inicialmente, los sistemas de depuración consistían en tareas manuales secuenciales, como la construcción de reglas para "detectar y corregir". Los sistemas resultantes eran a menudo grandes y complejos.

2.2.1 Métodos generalizados

A mediados de los 70, Fellegi y Holt (1976) proponen un sistema generalizado de depuración e imputación. Varios sistemas generalizados se han desarrollado a partir de entonces. En los métodos generalizados, las reglas y convenciones están estandarizadas, programadas y contrastadas, y pueden aplicarse a muchas encuestas diferentes. Las ventajas de los métodos generalizados son que utilizan una metodología contrastada y eficiente, que coordinan y sistematizan esfuerzos que se repiten de encuesta en encuesta, que ahorran recursos y tiempo en el estudio y desarrollo de procedimientos específicos y que garantizan la consistencia entre encuestas similares. El paradigma fundamental de los métodos generalizados procede del trabajo de Fellegi y Holt (1976), que propusieron una metodología por la cual un conjunto programas que se ejecutaban en bach llevaban a cabo las siguientes funciones:

1) análisis de los edits
2) detección de errores
3) identificación de las variables a imputar
4) imputación

La labor de los especialistas de la encuesta se limita a proporcionar los diferentes edits. El resto de las labores de depuración e imputación la lleva cabo automáticamente el programa. Los edits se clasifican en explícitos, que son los originalmente especificados por los especialistas de la encuesta y los implícitos, que se deducen lógicamente de los anteriores, y son generados automáticamente por el programa. El corazón de la metodología de Fellegi y Holt es el concepto de conjunto completo de edits, que es el conjunto de todos los edits generados implícitamente a partir del conjunto primitivo de edits explícitos. La generación del conjunto completo de edits, es decir, el conjunto donde se integran los edits explícitos y todos los edits implícitos, es un proceso interactivo. Fellegi y Holt demuestran en su artículo que, si el conjunto de edits explícito cumple unas condiciones bastante generales, el proceso es convergente. El procedimiento actúa siguiendo los siguientes principios: cada registro satisface todos los edits; el sistema minimiza el número de campos a imputar (conocido como principio de cambio mínimo); no es necesario especificar las reglas imputación, sino que se derivan automáticamente de los edits; y las imputaciones mantienen la estructura de frecuencias de los datos sin error.

La metodología de Fellegi y Holt es, en teoría, aplicable tanto a variables cuantitativas como a variables cualitativas. Sin embargo, en su implementación, surgen mayores problemas cuando se aplica a variables cuantitativas. Partiendo del paradigma de Fellegi y Holt se han diseñado otros sistemas generalizados que intentan adaptarse a los

problemas de los datos cuantitativos. Entre ellos pueden destacarse el sistema SPEER (Structured Program for Economic Editing and Referral) desarrollado por el Bureau of the Census, y el sistema GEIS (Generalized Edit and Imputation System), desarrollado por la oficina estadística de Canadá. Estos sistemas están descritos por ejemplo, en Pierzala (1990a, 1990b). Las principales características del sistema SPEER, descritas por Draper et al. (1990) y Winkler y Draper (1996), son la utilización de edits de la razón y su fundamentación en la teoría de grafos y en las técnicas estadísticas. Por su parte, las principales características del sistema GEIS, descritas en Kovar (1993) son la utilización de edits lineales, y su fundamentación en técnicas de investigación operativa y de programación lineal. Entre los diferentes métodos generalizados, pueden citarse los siguientes:

a) sistema DIA (Depuración e Imputación Automática), desarrollado por el INE, García et al (1990)
b) sistema SCIA, desarrollado por el ISTAT, Barcaroli et al. (1995)
c) sistema NIM (New Imputation Methodology), desarrollado también por la oficina estadística canadiense, Bankier et al. (1995 y 1996)
d) sistema CherryPi, desarrollado por la oficina estadística de Holanda, De Waal (1996)
e) sistema DAISY (Design, Analysis and Imputation System), desarrollado por la oficina estadística italiana, Barcaroli y Venturi (1997)
f) sistema MacroView, desarrollado por la oficina estadística de Holanda, Van de Pol et al. (1997)
g) sistema Plain Vanilla, desarrollado por la oficina estadística de Estados Unidos, Grahan (1997)
h) sistema StEPS(Standard Economic Processing System), desarrollado por la oficina estadística de Estados Unidos, Sigman (1997)
i) sistema AGGIES (Agriculture Generalized Imputation and Edit System), desarrollado por el ministerio de agricultura de Estados Unidos, Todaro (1998)
j) sistema Blaise, desarrollado por la oficina de estadística de Holanda, Blaise Reference Manual (1998)
k) sistema GEIS (Generalized Edit and Imputation System), desarrollado por la oficina estadística canadiense, Statistics Canada (1998), que ha derivado más recientemente en el sistema Banf

2.2.2 Métodos de depuración selectiva

La depuración selectiva constituye actualmente uno de los enfoques más prometedores de la depuración, especialmente en encuestas con datos cuantitativos. Su utilización está recomendada por diversos autores e instituciones, por ejemplo, el Grupo de Trabajo de depuración e imputación de las Naciones Unidas (2006), o las *Statistics Canada Quality Guidelines* (2009).

La depuración selectiva surge como un intento de solución de los problemas de la depuración tradicional. Entre estos problemas puede señalarse, en primer lugar, el elevado coste de la misma. A pesar de los esfuerzos realizados para racionalizar los procesos y a pesar de la ayuda que pueden suponer las herramientas informáticas y de comunicaciones, la experiencia de los distintos países muestra que el coste de la depuración no ha disminuido. En realidad, la facilidad que proporcionan los ordenadores ha sido frecuentemente utilizada para programar más edits de los realmente

necesarios o de los que el equipo de depuración es capaz de llevar a cabo. Esos edits adicionales han dado lugar a un incremento en el número de cuestionarios que necesitan una revisión manual. Todo ello ha dado lugar al fenómeno conocido como “sobredepuración”, es decir, los recursos y el tiempo dedicados a la depuración no se justifican por las mejoras resultantes en la calidad de los datos. Cuando se hace referencia al coste de la depuración, no se considera solamente el coste que esta supone para los organismos productores. En realidad, debe valorarse un coste más amplio, integrado además por el coste de los informantes de la encuesta, especialmente en el caso de que exista un número excesivo de recontactos, para discutir y aclarar al depurador la validez de los datos que no pasan los edits. También debe valorarse el coste en términos de tiempo invertido en la depuración, que, o bien recorta tiempo de otras fases del proceso de producción estadístico, o bien alarga el tiempo de difusión de la encuesta, uno de los aspectos de la calidad de una estadística que más interesa habitualmente a los usuarios.

Por otra parte, existe un abundante número de estudios en la literatura de la depuración, que muestran que mucho del impacto que la depuración tiene en los datos finales es atribuible tan sólo a un pequeño porcentaje del total de los edits. Por ejemplo, al evaluar la depuración en los censos económicos de los Estados Unidos, Greenberg y Petkunas (1986) encuentran que el 5% de los errores corregidos durante el proceso de depuración da lugar a más del 90% del cambio total en las estimaciones. Del mismo modo, en un estudio de la encuesta anual de cuentas financieras de Suecia, Wahlström (1990) encuentra que el 2% de los errores corregidos da lugar a más del 90% del cambio en las estimaciones. En la encuesta industrial anual sueca, Hedlin (1993) encuentra que el 8% da lugar al 95%. Finalmente, en la encuesta industrial anual canadiense, Boucher (1991) encuentra que después de corregir el 50% de los errores ya se había alcanzado el 100% del cambio en las estimaciones, de manera que la corrección del otro 50% de los errores no producía ninguna ganancia en la acuracidad de las estimaciones.

Otro problema de la depuración tradicional consiste en que la mayoría de los datos detectados como sospechosos de error no da lugar a correcciones, una vez que ha sido analizada su validez por parte del equipo de depuración (frecuentemente después de recontactar con el informante). El ratio de impacto (el número de datos que se corrigen respecto al total de datos que los edits señalaban como sospechosos de error) resulta ser demasiado bajo en muchas encuestas, evidenciando que la eficiencia del proceso de depuración podía ser mejorada y el número de recontactos disminuido (con la consiguiente reducción del coste y de la carga de respuesta). Por ejemplo, Lindström (1991) realiza un estudio en diversas encuestas y encuentra que el ratio de impacto se mueve entre el 28% y el 47%.

Por otra parte, la depuración también puede esconder problemas serios en la recogida de datos y dar una falsa impresión de la capacidad de respuesta de los informantes. Dos ejemplos de este problema pueden verse en el Australian Bureau of Statistics (1987). Al revisar los depuradores la sección de las compras, en el Censo del Comercio al por menor de Australia, sistemáticamente deducían una pequeña cantidad a partir de la "compra de bienes" y la insertaban en el espacio dejado en blanco por las empresas de "compras de envases y embalajes" para evitar un mensaje de error en la siguiente fase de depuración de controles previamente grabados por el ordenador. El otro ejemplo ocurría en la encuesta de producción de productos industriales. La clasificación de los bienes que figuraban en el cuestionario, era demasiado detallada para los informantes,

que agregaban la producción de algunos epígrafes y la asignaban globalmente en el epígrafe más relevante. El equipo de depuración manual comparaba las respuestas actuales con las respuestas obtenidas en periodos anteriores. A partir de esa información deducían las cantidades de los epígrafes complementados por las empresas, y asignaban los datos de producción a epígrafes que las empresas no habían complementado. Este proceso se conoce en la literatura de depuración como "depuración creativa" porque los depuradores manuales inventan los procedimientos de depuración e imputación mediante unos métodos subjetivos y no homogéneos, dando una falsa impresión de la capacidad de respuesta de los informantes.

Otro problema de la depuración tradicional, que con alta frecuencia se ha venido observando en la mayoría de organismos productores de estadística, es que los responsables de encuesta se marcan como objetivo de depuración utilizar el mayor número de contrastes de depuración posibles y los intervalos de aceptación lo más estrechos posibles. La idea que subyace aplicando estos objetivos, es que la calidad de la encuesta mejorará a medida que aumente el número de contrastes y la amplitud de los intervalos de rechazo (siguiendo el principio de" la seguridad lo primero", es decir cuanto más contrastes y más estrechos mejor). Sin embargo, algunas indicaciones de que esta idea no era cierta se tienen ya desde fechas tan tempranas como 1965. Usando datos de la encuesta industrial anual de Noruega de 1982, Nordbotten (1965) implantó una serie de estudios de simulación de un sistema experimental automático de depuración e imputación. Valores detectados como sospechosos de error por contrastes de la razón fueron imputados automáticamente usando un método hot deck. Nordbotten encontró que los intervalos de aceptación más amplios daban lugar a una calidad más alta, porque los intervalos de aceptación demasiado estrechos identificaban incorrectamente muchos datos válidos como erróneos, y entonces reemplazaban innecesariamente esos datos reales por valores imputados. Incluso esos intervalos de aceptación más amplios fueron considerados demasiado estrechos por Fellegi (1965) en su discusión sobre los estudios de simulación de Nordbotten. A partir de entonces, se han llevado a cabo un elevado número de estudios, que muestran que la calidad no siempre mejora cuando se utiliza el mayor número de contrastes de depuración posibles y cuando se utilizan intervalos de aceptación lo más estrechos posibles. Entre estos estudios pueden destacarse, por ejemplo, los de Corby (1984) y Werking et al. (1988).

Para solucionar estos problemas de la depuración tradicional, se propusieron los métodos conocidos como métodos de depuración selectiva. Estos métodos fueron introducidos por Granquist en la oficina estadística de Suecia a partir de 1984. En su inicio, estos métodos fueron conocidos como métodos de macrodepuración, ya que partían de los datos agregados y no de los microdatos. Sin embargo, actualmente reciben el nombre de métodos de depuración selectiva, dejando el amplio término de macrodepuración para todo procedimiento que realice los contrastes sobre los datos agregados (Glosario de términos de depuración e imputación, Winkler 1999). La depuración selectiva consiste en la detección y corrección selectiva de errores. Pone en relación los microdatos con los macrodatos, para determinar los errores de los microdatos que tienen influencia en los macrodatos. Su objetivo es agilizar las tareas de depuración e imputación en los datos de la encuesta sin detrimento en la calidad del proceso.

Para entender el concepto de depuración selectiva, conviene analizar los errores de una encuesta de la forma en que lo hace Granquist (1984b). En una encuesta, pueden

aparecer errores de negligencia y errores sistemáticos no anticipados. Los errores de negligencia se tienen como resultado de la falta de cuidado del informante o del proceso de la encuesta. Lo esencial es que, en una repetición de la encuesta, el mismo error probablemente no sería encontrado para la misma variable del mismo cuestionario. Por su parte, los errores sistemáticos no anticipados, aparecen debido a la ignorancia o a la mala interpretación de las cuestiones, conceptos o definiciones, o son cometidos deliberadamente. Lo esencial es que en una repetición de la encuesta, el mismo error probablemente afectaría a la misma variable del mismo cuestionario. Granquist encuentra que la depuración tradicional: no es eficiente con los errores de negligencia, está lejos de ser aceptable con los errores sistemáticos no anticipados y genera una ilusoria sensación de confianza

En un intento de resolver estos problemas, Granquist (1984a) propone el método de la depuración selectiva, que se basa en: a) Identificar áreas problemáticas, b) Estimar y documentar el impacto en las áreas problemáticas, c) tomar medidas para ajustar los errores sistemáticos no anticipados en la encuesta actual, y d) tomar medidas para mejorar este tipo de errores en futuras encuestas. El método se basa en un proceso selectivo de detección de errores. De acuerdo con ello, no todos los cuestionarios o variables son objeto de investigación. Son objeto de investigación aquellos cuestionarios que son verdaderamente influyentes en los agregados y se ignoran datos cuya magnitud no es significativa o que se cancelan en el proceso de agregación. La depuración selectiva se trata de una filosofía y no de un conjunto de procedimientos cerrados. Se concibe generalmente como un proceso interactivo. Se aplica preferentemente en datos cuantitativos y en encuestas de empresas caracterizadas por la asimetría de sus poblaciones.

A partir de los trabajos llevados a cabo por Granquist en la oficina estadística sueca, han aparecido un elevado número de métodos de depuración selectiva. Entre ellos mencionaremos los siguientes:

Método de agregación

El método de agregación fue desarrollado por Granquist (1988b) para la oficina estadística de Suecia. Versiones posteriores pueden verse en Lindström (1990a, 1990b). La idea fundamental de este procedimiento es trabajar en dos etapas. En la primera etapa, se pasa los controles de depuración a los datos agregados. En la segunda etapa, se pasan esos mismos controles a los datos individuales, pero únicamente a aquellos pertenecientes a los agregados que no han pasado los controles en la primera etapa.

El procedimiento admite dos versiones: en una primera versión, se depurarían en una primera etapa los datos agregados, a un determinado nivel de agregación (por ejemplo sectores de actividad o regiones geográficas de los cuales se quiere obtener información). En una segunda etapa, se depurarían aquellos microdatos de los datos agregados que no han pasado los controles de la primera etapa. En una segunda versión de esta misma idea, se asignaría a los registros de un agregado detectado como sospechoso de error, una señal específica que mostrara cuál de las variables no pasa los controles. Entonces, la depuración al nivel de microdatos, se aplicaría únicamente en aquellos cuestionarios que pertenecen a los agregados que no pasa los controles para esa variable específica.

Uno de los rasgos esenciales de los métodos de agregación consiste en que los intervalos de aceptación son determinados manualmente mediante la revisión de listas de observaciones clasificadas de acuerdo a las funciones de los contrastes.

Sólo las n más grandes observaciones y las m más pequeñas observaciones de las funciones de chequeo son impresas en las listas. Tanto las funciones de chequeo, que trabajan a nivel agregado, como las funciones de chequeo que trabajan en el nivel desagregado, tienen que ser funciones de los pesos (de acuerdo al diseño muestral).

Los estudios de evaluación llevados a cabo por la oficina estadística sueca, concluyeron que el trabajo de verificación de errores se redujo en un 50% sin descenso en la calidad de los datos. Dentro del método de agregación, Anderson (1989b) propone utilizar diagramas de caja para determinar los intervalos de aceptación.

Método "top-down"

El método "top-down" fue desarrollado por Granquist (1987 y 1991) para la oficina estadística de Suecia. La idea que subyace en este método es clasificar de forma jerárquica los valores de las funciones de chequeo (que son funciones de los valores debidamente ponderados) y comenzar la revisión manual desde arriba o desde abajo y continuar hasta que las correcciones no tengan un efecto significativo en las estimaciones. Por tanto, se seleccionan y ordenan de mayor a menor los n valores extremos de una variable o una función de variables. Por ejemplo, los n mayores cambios positivos, o los n mayores cambios negativos, o las n mayores contribuciones al agregado. De modo interactivo se pueden corregir errores en la pantalla y ver cómo se modifica la lista valores extremos. Se deja de corregir cuando se considera que las correcciones no modifican sustancialmente los totales (Anderson 1989a)

Desagregación en cascada de tablas de series

Es un procedimiento de macrodepuración desarrollado en el ámbito de las encuestas industriales del INE. Permite agregar o desagregar cualquier conjunto de variables y ratios de variables (habitualmente en forma de series temporales) bajo cualquier condición. Este procedimiento permite buscar el o los cuestionarios iniciales responsables del presunto error. No existe una forma previa, obligatoria de usar para buscar ese o esos cuestionarios sospechosos. Cada persona puede decidir cual es la mejor forma de buscar el error, en función de su experiencia sobre el sector o, simplemente según su costumbre. La esencia del método es que la búsqueda de la información la determina el usuario, y no está prefijada de antemano mediante un conjunto fijo de tablas.

Métodos basados en la función score

Un enfoque fundamental en los métodos de depuración selectiva es la construcción de una función de tanteo o puntuación para cada dato o para cada unidad informante. Latouche y Berthelot (1990 y 1992) fueron los pioneros en la utilización de funciones score como instrumento de depuración selectiva. En sus investigaciones establecieron las bases de una función score y propusieron tres funciones. Su objetivo es optimizar la estrategia de investigación y recontacto con los informantes. Su metodología es aplicable a encuestas continuas de empresas, caracterizadas por tres elementos:

investigan fundamentalmente variables cuantitativas, las variables de interés tienen distribuciones altamente asimétricas, donde unas pocas empresas contribuyen a una parte importante de las características poblacionales a investigar, y se dispone de información histórica de encuestas anteriores.

Las funciones tanteo deben tener en cuenta cuatro criterios: el tamaño de la unidad informante, el número de datos sospechosos de error dentro del cuestionario, el tamaño de esos datos sospechosos de error, y la importancia relativa que se asigna a las variables sujetas a posibles errores. Además de estos aspectos teóricos, la función de tanteo debe tener en cuenta aspectos operativos, como que sea fácil de implementar y sea suficientemente flexible para poder ser utilizada en diferentes encuestas. Al mismo tiempo, la fórmula de la función de tanteo debe de tener una interpretación relevante desde el punto de vista del contenido y objeto de la encuesta. Finalmente, otro aspecto operativo a tener en cuenta es que el uso de la función no retrase el proceso de la encuesta. Para ello, las variables utilizadas en la función no deben depender del flujo de respuestas, sino que deben conocerse al comienzo de la encuesta, para que el método pueda aplicarse desde el primer momento, aún cuando la tasa de respuesta sea todavía muy pequeña. Para ello, las variables utilizadas pueden proceder de la misma encuesta en periodos anteriores y, del periodo corriente, solamente sea necesario utilizar los datos proporcionados por el informante que se está depurando.

La función Score es una fórmula matemática que asigna una puntuación relativa a cada unidad informante, utilizando como inputs las características de la unidad informante que están relacionadas con el potencial impacto en las estimaciones, combinando los factores (teóricos y operativos) deseables. Primero se calcula una puntuación para cada variable de un cuestionario, luego se suman estas puntuaciones para obtener una puntuación global de cada unidad informante. Las unidades informantes con más alta puntuación se considera que van a tener mayor influencia en los agregados, y deben ser recontactadas con mayor prioridad

Significance editing

Lawrence y McDavitt (1994), realizan un trabajo pionero aplicándolo a la Encuesta de Ganancias Salariales Semanales de Australia y Lawrence and McKencie, (2000), proponen una metodología general de depuración selectiva. El principio fundamental de *significance editing* es estimar el efecto en las estimaciones de los agregados de resolver un determinado dato sospechoso de error. Un prerequisito es disponer de un modelo adecuado de depuración, formado por el conjunto de *edits.* Por tanto, *significance editing* no mejora un modelo inadecuado de depuración ni detecta errores que no hayan sido detectados por el conjunto de edits. Los elementos que constituyen una estrategia de *significance editing* son la determinación de un “valor arreglado esperado”, el cálculo de *score* locales, la combinación de *score* locales para calcular un *score* global y el establecimiento de umbrales de corte para los *scores.* La determinación del “valor arreglado esperado” se hace a partir del conjunto de edits, partiendo de la idea de que los edits expresan un punto de vista predeterminado de cómo la población se tiene que comportar. Aunque no se contempla un método general para generar el valor arreglado esperado a partir del modelo de depuración, en Lawrence and McKencie, (2000), se ofrecen algunos procedimientos concretos aplicables a conjuntos de edits utilizados habitualmente en encuestas de empresas.

Capitulo 3

Planteamiento teórico y desarrollo empírico de nuevas técnicas de depuración basadas en modelos estadísticos y en técnicas de optimización estocástica

3.1 Depuración e imputación basada en modelos de series temporales

3.1.1 Introducción

Los modelos de series temporales no son de uso común en la depuración e imputación ni en otras fases del proceso de producción de estadísticas, con la excepción del ajuste estacional. Sin embargo, las encuestas continuas dan lugar a un conjunto de observaciones secuenciales recogidas a lo largo del tiempo. Por tanto, el marco teórico no tiene por qué limitarse al de las variables aleatorias, sino que puede ser ampliado al de los procesos estocásticos, y el uso de modelos de series temporales está plenamente justificado. Si existe información sobre periodos anteriores, se debe utilizar al máximo en la depuración e imputación, y ello puede hacerse a través de ese tipo de modelos. Por otra parte, el paradigma de Fellegi-Holt (1976) proporciona la base para la construcción de sistemas generalizados. En los sistemas generalizados las reglas están estandarizadas, programadas y contrastadas y pueden aplicarse a encuestas diferentes. Los sistemas generalizados basados en la filosofía de Fellegi-Holt llevan a cabo el análisis de los edits para chequear su consistencia, la detección de errores, la identificación de las variables a imputar y la imputación. Sin embargo, no resuelven la especificación de los edits, que se considera un input del sistema y deben ser especificados por los usuarios (los responsables de cada encuesta). Habitualmente, los usuarios establecen los edits de acuerdo a su experiencia práctica, sin que exista un marco teórico para los mismos. En este trabajo se propone un procedimiento para tratar la especificación de los edits y se diseñan herramientas concretas de depuración e imputación de datos estadísticos, basándose en la utilización de series temporales y aplicable a encuestas continuas.

Como un criterio esencial de la depuración e imputación consideramos que los datos sean consistentes con la información existente. En las encuestas continuas, la principal información suele ser la histórica de la misma encuesta en periodos anteriores. La información existente se resume en un modelo**.** El modelo debe extraer al máximo la información que contienen los datos. Adicionalmente, debe incorporar la información *a priori.* Para hacer operativo el concepto de que los datos sean consistentes consideramos que un dato es consistente cuando se acerca a la predicción. Las herramientas de depuración e imputación que aquí se proponen se basan fundamentalmente en las predicciones del modelo. Los valores atípicos, y por tanto sospechosos de error, se definen y se detectan en función de cuánto se alejan los datos observados (a depurar) de la predicción. Los valores imputados se definen como la predicción que se deduce del modelo para ese dato.

En este trabajo se hace uso de modelos RegARIMA y se desarrollan a partir de ellos nuevas herramientas de depuración e imputación. Los modelos se utilizan en la microdepuración, en la macrodepuración y en la depuración selectiva. Del mismo modo, se utilizan en la micro y macro imputación. Las propiedades de la depuración e imputación se derivan fundamentalmente a partir de las funciones de predicción de los modelos. Para el trabajo empírico nos centramos en los Índices de producción industrial (IPI). Fórmulas similares pueden derivarse fácilmente si se utilizan para otros indicadores de corto plazo.

Este subcapítulo se estructura en cinco apartados. En el 3.1.1 se realiza la introducción. En el 3.1.2 se describen las herramientas desarrolladas de depuración e imputación y en el 3.1.3 las de macrodepuración y análisis. En el 3.1.4 se presenta un ejemplo con datos reales. Finalmente, en el 3.1.5 se formulan algunas conclusiones.

3.1.2 Construcción de herramientas de depuración e imputación

Los Índices de producción industrial (IPI) se calculan a través de una encuesta mensual. Se utiliza una muestra panel de cerca de 14.000 empresas. Una sola variable, la producción, en una unidad física particular (toneladas, litros, etc), o en valor monetario, se solicita a cada empresa. Como resultado de la encuesta, tenemos un conjunto de microdatos $q_{i,j,t}$, es decir, la cifra de producción para el producto i informado por la empresa *j* en el mes t. Desde el conjunto de microdatos , el índice para el producto i se calcula como:

$$I_{i,t} = I_{i,t-1} \frac{\sum_j q_{i,j,t}}{\sum_j q_{i,j,t-1}}$$

Donde *j* es el conjunto de empresas con valores válidos en tanto t y $t-1$.

Y, a partir de estos índices de productos, los índices de Laspeyres agregados se calculan en los niveles sucesivos de desagregación de la clasificación de actividades económicas (la parte superior de la agregación es el total de la industria). Se utiliza la siguiente fórmula

$$I_t = \sum_i w_i I_{i,t}$$

donde los pesos del año de base se basan en el valor añadido (para las actividades) o el valor de la producción (para los productos).

Utilizamos el mismo tipo de modelos (RegARIMA) para las series macrodatos y microdatos. Dado que el número de serie de tiempo para manejar es muy grande y es difícil y costoso construir modelos para todos ellos, es necesario un procedimiento automático. Utilizamos un método automático desarrollado por Revilla, Rey y Espasa que encaja en la estrategia iterativa de modelización de Box-Jenkins de especificación inicial, estimación y comprobación (Box y Jenkins, 1970). Un uso directo de los modelos ARIMA no es suficiente para capturar variaciones de calendario, porque no

son exactamente periódicas. Se utilizan modelos de regresión para manejar los efectos de calendario y otras variaciones determinísticas. Para especificar las variables de intervención se necesita algún conocimiento a priori sobre el comportamiento de los datos de producción.

Por lo tanto, los modelos globales son la suma de los modelos de regresión y ARIMA (modelos RegARIMA):

$$ln\, q_{i,j,t} = \frac{\theta_{i,j}(B)\,\Theta_{i,j}(B^{12})}{\varphi_{i,j}(B)\,\Phi_{i,j}(B^{12})} a_{i,j,t} + \sum_h \frac{\alpha_{i,j,h}(B)}{\delta_{i,j,h}(B)} A_{i,j,h,t}$$ para modelizar los microdatos.

$$ln\, I_{i,t} = \frac{\theta_i(B)\,\Theta_i(B^{12})}{\varphi_i(B)\,\Phi_i(B^{12})} a_{i,t} + \sum_h \frac{\alpha_{i,h}(B)}{\delta_{i,h}(B)} A_{i,h,t}$$ para modelizar los índices.

donde:

• $ln\, q_{i,j,t}$ es el logaritmo neperiano de la cifra de producción para el producto i, de la empresa j.

• $ln\, I_{i,t}$ es el logaritmo neperiano del índice de producción industrial para el producto (o actividad) i .

• B es el operador retardos, $B^k(I_t) = I_{t-k}$

• $\theta(B), \varphi(B), \Theta(B^{12}), \Phi(B^{12}), \alpha_h(B), \delta_h(B)$ son polinomios en el operador retardos, para el producto (o actividad) i, o para las empresas i, j.

• $a_{i,j,t}$ y $a_{i,t}$ son variables ruido blanco i. i. d. $N(0,\sigma_{i,j})$ y $N(0,\sigma_i)$ respectivamente.

• $A_{i,j,h,t}$ y $A_{i,h,t}$ son variables de intervención.

Para poder utilizar un enfoque de microdepuración se ajusta un modelo RegARIMA para cada serie de datos de empresa.. El modelo se utiliza tanto para la depuración como para la imputación. Un intervalo de confianza (por ejemplo, un intervalo de 95%) puede ser construido a partir del modelo:

$$P\left[\hat{q}_{i,j,t} - 1{,}96\,\sigma_{ij} < q_{i,j,t} < \hat{q}_{i,j,t} + 1{,}96\,\sigma_{i,j}\right] = 0{,}95$$

donde $\hat{q}_{i,j,t}$ es la previsión un periodo por delante para $q_{i,j,t}$, y los valores sospechosos se pueden definir como los microdatos fuera del intervalo. El valor imputado de $q_{i,j,t}$ sería $\hat{q}_{i,j,t}$.

Para el uso de un enfoque macrodepuración, un modelo RegARIMA se ha construido para cada uno de la serie de índices de productos y actividades. El modelo se utiliza

para la depuración e imputación. Un intervalo de confianza (por ejemplo, un intervalo de 95%) puede ser construido a partir del modelo:

$$P\left[\hat{I}_{i,t} - 1{,}96\ \sigma_i < I_{i,t} < \hat{I}_{i,t} + 1{,}96\ \sigma_i\right] = 0{,}95$$

donde $\hat{I}_{i,t}$ es la previsión un periodo por delante para $I_{i,t}$, y los valores atípicos se puede definir como los índices fuera del intervalo. El valor imputado de $I_{i,t}$ sería $\hat{I}_{i,t}$.

En el uso de un enfoque de depuración selectiva tenemos que resolver dos problemas: detectar anomalías en los datos macro (los índices) y detectar los microdatos influyentes. Para enfrentar el primer problema hemos diseñado algunas herramientas, las "sorpresas", que son funciones de la previsión del modelo RegARIMA (en particular, de las previsiones un periodo por delante):

La sorpresa (o sorpresa simple) para el índice es el cambio relativo entre el valor observado y la previsión:

$$S_{i,t} = \frac{I_{i,t} - \hat{I}_{i,t}}{\hat{I}_{i,t}}$$

Si calculamos la previsión un periodo por delante de la serie en logaritmos $\ln \hat{I}_{i,t}$ para $\ln I_{i,t}$ el error un periodo por delante es:

$$e_{i,t} = \ln I_{i,t} - \ln \hat{I}_{i,t}$$

Dado que error de predicción a $e_{i,t}$ se distribuye como un proceso de ruido blanco $N(0,\sigma_i)$ y, por otra parte, $\ln I_{i,t} - \ln \hat{I}_{i,t} \cong \left(I_{i,t} - \hat{I}_{i,t}\right) \big/ \hat{I}_{i,t}$, tenemos que $S_{i,t}$ se distribuye $N(0,\sigma_i)$. Por lo tanto, se puede construir un intervalo de confianza (por ejemplo, un intervalo de 95%) para las sorpresas:

$$P\left[-1.96\sigma_i < S_{i,t} \le 1.96\sigma_i\right] = 0.95$$

y los valores atípicos se pueden definir como los índices cuyos sorpresa está fuera del intervalo.

La sorpresa estandarizada para el índice $I_{i,t}$ es:

$$\frac{S_{i,t}}{\sigma_i} = \frac{I_{i,t} - \hat{I}_{i,t}}{\hat{I}_{i,t}} \frac{1}{\sigma_i}$$

Esta herramienta permite la comparación directa de índices con diferentes variabilidad.

La sorpresa estándar ponderada para el índice $I_{i,t}$ es:

$$\frac{S_{i,t}}{\sigma_i} w_i = \frac{I_{i,t} - \hat{I}_{i,t}}{\hat{I}_{i,t}} \frac{w_i}{\sigma_i}$$

Esta herramienta permite la clasificación de los índices teniendo en cuenta no sólo la magnitud de la sorpresa sino también los diferentes pesos.

Una vez que hemos detectado y clasificado los índices sorprendentes (los índices que no son coherentes con su comportamiento en el pasado y por lo tanto pueden ser considerados como valores atípicos) necesitamos para medir el impacto de cada uno de los microdatos de estos índices sorprendentes. Para ello, usamos las "influencias". La influencia de un dato individual sobre una magnitud agregada se define como la diferencia entre la magnitud agregada observada y el valor de esta misma magnitud cuando el punto de referencia individual no está disponible.

La influencia del microdato $q_{i_0,j_0,t}$ en el índice de productos $I_{i_0,t}$ es:

$$INF_{i_0,j_0}^{I_{i_0,t}} = I_{i_0,t-1} \frac{\sum_j q_{i_0,j,t}}{\sum_j q_{i_0,j,t-1}} - I_{i_0,t-1} \frac{\sum_{j \neq j_0} q_{i_0,j,t} + \hat{q}_{i_0,j_0,t}}{\sum_j q_{i_0,j,t-1}} = I_{i_0,t-1} \frac{q_{i_0,j_0,t} - \hat{q}_{i_0,j_0,t}}{\sum_j q_{i_0,j,t-1}}$$

Donde $\hat{q}_{i_0,j_0,t}$ es un valor imputado para el microdato $q_{i_0,j_0,t}$

La influencia sobre el índice agregado I_t es:

$$INF_{i_0,j_0}^{I_t} = \sum_i w_i I_{i,t} - \left[\sum_{i \neq i_0} w_i I_{i,t} + w_{i_0} I_{i_0,t-1} \frac{\sum_{j \neq j_0} q_{i_0,j,t} + \hat{q}_{i_0,j_0,t}}{\sum_j q_{i_0,j,t-1}} \right] = w_{i_0} I_{i_0,t-1} \frac{q_{i_0,j_0,t} - \hat{q}_{i_0,j_0,t}}{\sum_j q_{i_0,j,t-1}}$$

Esta expresión mide el impacto de los microdatos en el índice por medio de los siguientes factores:

• El peso w_{i_0} del producto (o actividad)

• El índice $I_{i_0,t-1}$ que "actualiza" el peso anterior.

• Una medida de la discrepancia relativa entre el dato observado y el dato imputado (la predicción obtenida del modelo)

$$\frac{q_{i_0,j_0,t} - \hat{q}_{i_0,j_0,t}}{\sum_j q_{i_0,j,t-1}}$$

Este resultado está en línea con la metodología de las funciones de puntuación (funciones score) utilizadas en la literatura, por ejemplo en Latouche y Berthelot (1992) o Hedlin (2003). De acuerdo a ella, las funciones score se basan generalmente en dos componentes multiplicativos, el componente influencia y el componente riesgo. El primer componente mide la influencia de una variable en la estimación de un total y el segundo la probabilidad de un potencial error. En nuestro caso los dos primeros factores representan el componente influencia y el tercero el componente riesgo. En los estudios citados anteriormente el componente riesgo se estima mediante la diferencia entre el valor observado y un valor "anticipado". Este valor anticipado es una estimación para el valor verdadero que se habría obtenido después de la depuración interactiva, utilizándose habitualmente el valor del periodo anterior. En nuestro trabajo, utilizamos como valor anticipado la predicción del modelo, que supone una mejora respecto a los procedimientos habituales. De hecho, únicamente en caso que la serie siguiera un proceso camino aleatorio podrían equipararse los dos procedimientos.

Puede demostrarse que los microdatos que son más influyentes en el índice agregado también son los más influyentes en las sorpresas de ese índice. Estos "influencias" nos permiten dar prioridad a los microdatos de índices sospechosos con el fin de verificar manualmente y recontactar menos empresas.

3.1.3 Construcción de herramientas de macrodepuración y análisis

A partir de los modelos se ha obtenido información muy útil para llevar a cabo la macrodepuración, en línea del análisis exploratorio de datos (Tukey, 1977). Los modelos se utilizan para estimar un conjunto de características de un indicador a corto plazo como el comportamiento tendencial, las variaciones estacionales, las oscilaciones cíclicas, los efectos de calendario y otros efectos determinísticos, los valores atípicos, la volatilidad, (como una huelga), etc. Así, para cada agregado, se construye un vector de valores correspondientes a las características arriba mencionadas. En la macrodepuración y análisis de una encuesta, antes de su difusión, se necesita la mayor cantidad de información posible sobre el fenómeno se trata de medir. Por otra parte, diferentes subconjuntos de datos, procedentes de la misma encuesta, muestran a menudo variabilidades y comportamientos muy diferentes. Por ejemplo, en los índices mensuales de producción industrial español, podemos encontrar los valores muy pequeños (incluso cero) para agosto, debido a las vacaciones de verano se toman generalmente en este mes. Estos datos no deben ser considerados como valores atípicos (es decir, sospechosos de error) si se tiene información sobre este patrón estacional. Sin embargo, este comportamiento estacional en agosto es muy diferente de una rama a otra. Incluso hay ramas en que la producción no disminuye, sino que crece con intensidad en agosto, como en la producción de cerveza. Por esta razón, es de utilidad adquirir información acerca de las diferentes características dinámicas de cada uno de los subconjuntos de datos, tanto para llevar a cabo la macrodepuración y análisis de una encuesta como para mejorar las normas y estrategias de depuración.

Aunque, desde un punto de vista teórico, los modelos multivariantes (que recogen la correlación de todas las variables de la encuesta) proporcionarían mayor información, nos restringimos al entorno univariante, debido a las dificultades de construir modelos multivariantes para un conjunto tan elevado de series. Por otra parte, el uso de modelos ARIMA univariantes para describir las características dinámicas de un fenómeno

económico tiene una base metodológica sólida. En condiciones bastante generales (ver Prothero y Wallis (1976), Wallis (1977), Zellner (1979)) cualquier variable que se determina dentro de un modelo econométrico simultáneo dinámico estructural (SEM) se genera de manera univariante por un modelo ARIMA con Análisis de Intervención. El componente de análisis de intervención recoge la contribución de las variables ficticias del modelo de SEM y/o el efecto de ciertas intervenciones, que afectan a las variables exógenas de ese modelo. En la medida en que el modelo SEM refleja las características del mundo real, el modelo ARIMA con Análisis de Intervención correspondiente a una variable endógena del modelo SEM incorpora consistentemente, aunque sea de forma parcial, las características de esa variable.

Así, para cada uno de los agregados se construye un vector de valores correspondientes a las características anteriormente mencionadas. La metodología propuesta se ha aplicado a los Índices de producción industrial, pudiendo utilizarse de forma análoga con cualquier encuesta continua o indicador de corto plazo, de los que se disponga de una serie temporal suficientemente larga. El IPI se desagrega por ramas de actividad a cinco niveles sucesivos de desglose y por Comunidades Autónomas. Para construir modelos para todos ellos, se ha utilizado un procedimiento automático desarrollado por Revilla, Rey y Espasa (1990), que se enmarca en la estrategia iterativa de modelización de Box-Jenkins, de identificación, estimación y validación. Para especificar las variables de intervención se ha encontrado que los procedimientos automáticos no son adecuados para todas las series y se necesita mejorarlos mediante la modelización manual. A continuación, vamos a considerar los diferentes aspectos que hemos estudiado a partir de los modelos.

a) Comportamiento del nivel

Una descripción adecuada de la naturaleza de la tendencia a largo plazo de la serie se encuentra en los modelos. Esta tendencia se determina por la contribución en la función de previsión final de las raíces unitarias positivas reales del factor de autorregresivo y por la contribución de la posible media distinta de cero de la serie estacionaria. La presencia de *d* raíces del tipo mencionado significa que la tendencia a largo plazo es un polinomio temporal de orden *(d-1)*, cuyos coeficientes se determinan por las condiciones iniciales en que se encuentra el sistema. La presencia de una media distinta de cero aumenta el polinomio anterior con un término de orden d con un coeficiente determinista. Por lo tanto, cuando la serie es estacionaria, el modelo no requiere diferencias. Cuando el modelo especifica una diferencia de la serie mostrará oscilaciones locales de nivel, cuando el modelo especifica dos diferencias de la serie tendrá una tendencia casi lineal, etc.

b) Comportamiento estacional

Los modelos pueden contener también un factor, que recoge un ciclo estacional con un período de 12 unidades de tiempo. Las raíces unitarias complejas y negativas reales del polinomio autorregresivo reflejan este factor. Si ninguna de estas raíces se repite su contribución en la función de predicción final consiste en 12 factores estacionales estables y aditivos, que se determinan por las condiciones iniciales del sistema. Como resultado de ello, la senda a largo plazo de la serie se compone de estos factores estacionales y de la tendencia descrita en a).

c) Efectos de calendario.

Los índices españoles de producción industrial, como otras series de flujos, se ven afectados por efectos de calendario, ya que contienen variaciones debidas a la longitud y a la composición días de la semana de cada mes. Asimismo, se ven afectados por las vacaciones, incluyendo las vacaciones móviles, por ejemplo la Semana Santa. Adicionalmente, los días festivos varían de un año a otro, y de una Comunidad Autónoma a otra. Para modelizar estos efectos de calendario, incorporamos esta información como variables determinísticas. En lugar de las más frecuentemente utilizadas siete variables de *trading day* (Hillmer, Bell y Tiao, 1983) se construye una sola variable, adaptada al comportamiento de la producción industrial española. Esta variable mide el número de días laborables, eliminando los sábados, domingos y festivos. Las fiestas que varían de una Comunidad Autónoma a otra se ponderan por el valor añadido industrial del año base. Al construir el modelo sobra la transformación logarítmica de la serie, el parámetro asociado con esta variable ficticia puede ser interpretado como el aumento proporcional en la producción en comparación con la de un mes similar con un día menos de trabajo. Para completar la descripción de los efectos de calendario, se incluye otra variable determinística que contiene el efecto de Pascua. En marzo y abril toma los valores que indican la proporción de días afectados por estas fiestas y cero en los otros meses (Hillmer, Bell y Tiao, 1983). El parámetro que afecta a esta variable artificial se puede interpretar como la variación proporcional sufrida por la producción como resultado de estas vacaciones. Por lo tanto, los efectos de calendario se resumen utilizando sólo dos variables, lográndose unos modelos más parsimoniosos que con la metodología habitual.

d) Otros efectos determinísticos

Es posible encontrar contribuciones deterministas en la tendencia y/o en los factores estacionales de la serie, que pueden modelizarse mediante el análisis de intervención. Como ejemplo, en febrero de 1997, el sector transporte se declaró en huelga. Para la mayoría de las ramas industriales, la huelga causó una escasez de materias primas. El efecto esperado es una reducción inmediata en el nivel de producción. Después de algunos períodos de observación, se encuentra en algunas series un aumento en los meses de marzo y abril, que compensa la disminución en febrero. En efecto, algunas fabricas han intentado cumplir con los pedidos de los clientes mediante trabajo adicional en meses siguientes a la huelga. Para captar estos comportamientos, hemos construido dos variables diferentes, según se compense o no el efecto de la huelga en los meses siguientes.

1. La H_t variable, donde:

$$H_t = \begin{cases} 1.0, & t = \text{Febrero1997} \\ 0.6, & t = \text{Marzo1997} \\ 0.4, & t = \text{Abril1997} \\ 0.0, & t \neq \text{Feb., Mar. y Abr. 1997} \end{cases}$$

2. La variable de pulso Pt, donde:

$$P_t = \begin{cases} 1, & t = \text{Febrero}1997 \\ 0, & t \neq \text{Febrero}1997 \end{cases}$$

Hemos aceptado que la huelga tuvo un efecto sobre una serie de índices de producción industrial cuando el parámetro variable de intervención es significativamente diferente de cero. Cuando los parámetros son significativos para los dos de ellos, hemos elegido la variable de intervención que produce una desviación estándar residual menor del modelo. Como hemos utilizado la transformación logarítmica de la serie, es posible interpretar el valor de los parámetros como un efecto porcentual sobre el nivel de la serie original.

e) Valores atípicos

Hemos estudiado la existencia de observaciones inusuales o inesperadas. En función de su naturaleza, hemos detectado tres tipos de valores atípicos, outlier aditivo (AO), Cambio de Nivel(LS), y Cambio Temporal (TC), siguiendo el enfoque de Chang et al. (1988), a partir de los residuos estimados del modelo. El estudio de los valores extremos se puede utilizar para detectar los eventos especiales que puedan afectar a la producción en un período de tiempo determinado, para analizar si aparecen fortuitamente o no en las diferentes ramas, el mes en el que más a menudo aparecen, etc.

f) Volatilidad

La última característica que hemos considerado es la volatilidad, a través de una medida de la incertidumbre sobre la evolución futura de la serie, expresada por la desviación estándar de errores de predicción un periodo por delante.

3.1.4 Ejemplo con datos reales

Las herramientas de depuración e imputación pueden utilizarse con distintas estrategias. Un ejemplo del uso de este método en dos listados de trabajo En la tabla 1 se muestran la tasa de variación anual y las sorpresas simples. Por su uso intuitivo, se muestran los intervalos de confianza construidos a partir de los modelos para las tasas de variación anual en vez de para los índices. En este caso la rama 4 se consideraría sospechosa de error por estar fuera del intervalo del 95%. Descendiendo al siguiente nivel de desagregación se observa que la rama 41 está fuera del intervalo. Del mismo modo, lo está la 411 (y no la 412 ni la 413), por lo que las empresas pertenecientes a esta rama pasarían a depuración manual intereactiva. En la tabla 2, los sectores se clasifican en función de la sorpresa estándar ponderada, lo que permite priorizar las empresas que pasan a depuración interactiva. También puede establecerse un criterio de selección, por ejemplo no depurando las empresas de ramas con sorpresa estándar ponderada que en valor absoluto están por debajo de 1,96.

Tabla 1: Tasas de variación y sorpresa simple

Rama de actividad	Tasa Anual observada	Intervalo 95% de confianza Tasa Anual	Predicción Tasa Anual	Sorpresa Simple	Desviación Teórica
4	-2,93	[-2.29; 7.68]	2,57	-5,37	2,48
41	-8,22	[-7.19; 4.10]	-1,71	-6,63	2,93
411	-50,24	[-15.31; 1.52]	-6,89	-70,96	36,95
412	-6,38	[-31.65; 10.22]	-13,21	7,86	12,19
4121	-14,02	[-35.21; 23.47]	-10,56	-3,87	16,45
4123	2,82	[-32.69; 32.72]	-5,48	8,79	17,32
4124	-30,08	[-34.53; -.87]	-19,44	-13,21	10,58
413	1,22	[-39.36; 14.47]	-16,69	21,50	16,21
4131	2,81	[-.95; 11.19]	4,94	-2,03	2,95
4132	-2,58	[-36.56; 28.83]	-9,59	7,75	18,07
4133	16,74	[-15.56; 45.10]	10,69	5,47	13,81

Tabla 2: Sorpresas

Rama de actividad	Tasa observada	Predicción de la Tasa	Sorpresa Simple	Sorpresa Estándar	Sorpresa Estándar Ponderada
4243	70,28	3,32	64,73	3,79	17,10
2511	-27,73	-3,29	-25,25	-3,11	-16,93
4110	-50,24	-6,89	-70,96	-6,84	-16,89
2514	-15,92	4,64	-19,62	-3,00	-16,87
2512	39,39	-11,83	58,12	7,22	16,51
4752	-0,74	2,06	-2,75	-1,09	-15,66
3299	-11,97	4,45	-15,70	-2,02	-15,57
4751	22,82	-7,36	32,55	2,34	14,64
3630	-0,28	3,68	-3,82	-0,81	-14,54
3166	15,97	5,83	9,58	1,89	13,92

En relación con la utilización de los modelos para la macrodepuración y análisis, en la Tabla 3 se presenta un ejemplo de las principales características dinámicas de los índices de producción, para el Total Nacional y las Comunidades Autónomas. Por ejemplo, el Total Nacional muestra un comportamiento tendencial y un carácter estacional estocástico, lo que implica un comportamiento diferente en la actividad de

producción en diferentes meses del año. La producción industrial es sensible a los efectos de calendario. Más específicamente, la existencia de un día laborable menos da lugar a una caída de 1,9% en la producción. Del mismo modo, la semana santa provoca una caída de la producción del 4,2%, distribuido entre marzo y abril de acuerdo a la proporción de días afectados por esta fiesta cada año. La huelga de transporte de febrero provocó una reducción de 3,8% en el nivel de producción de este mes, compensado en marzo y abril. No muestra los valores atípicos por encima de tres desviaciones estándar. El grado de incertidumbre en cuanto a la producción para el próximo mes es del 2,1%. Una información similar puede obtenerse para cada una de los Comunidades Autónomas, observándose diferentes patrones de comportamiento.

Tabla 3: Caracterización IPI

	Comportamiento del nivel	**Estacio-nalidad**	**Efecto día laborable (%)**	**Efecto semana santa (%)**	**Efecto huelga (%)**	**Valores Atípicos**	**Incerti-dumbre**
Total Nacional	Tendencia	Sí	1.9	-4.2	-3,8		2.1
Andalucía	Tendencia	Sí	1.7	-2.0	(*) +7.5	Ene 1996 (-) Feb 1997(+) Feb 1998 (+)	2.7
Aragón	Tendencia	Sí	2.1	-4.1	(*) -6.6	Dic 1994 (+) Feb 1997 (-)	2.2
Asturias	Tendencia	Sí	1.3	-4,3	-4,4		2.3
Baleares	Tendencia	Sí	1.6	-3,5			2.3
Canarias	Oscilaciones locales	No	1.8	-3,7		Mar 1994 (+) Mar 1997 (+)	2.9
Cantabria	Tendencia	Sí	1.9		-4,2		2.8
Castilla-León	Tendencia	Sí	1.6	-5,2	(*) -8,1	Feb 1996 (-) Jul 1995 (-) Dic 1996 (-) Feb 1997 (-) Nov 1997 (-)	2.2
Castilla-La Mancha	Tendencia	Sí	0.8	-4,0			2.7
Cataluña	Tendencia	Sí	2.0	-4,7	-4,6		2.4
Comunidad Valenciana	Tendencia	Sí	2.2	-4,6	-4,4		1.7
Extremadura	Oscilaciones	No	1.3				4.5
Galicia	Tendencia	Sí	1.9	-3,3	(*) -5.6	Dic 1995 (+) Feb 1997 (-)	1.9
Madrid	Tendencia	Sí	1.6	-5,2			3.0
Murcia	Tendencia	Sí	2.1		-3,2		2.7
Navarra	Tendencia	Sí	2.4	-4,7	(*) -7.6	Feb 1997 (-)	2.6
País Vasco	Tendencia	Sí	2.4	-3,8	(*) -7.7	Ago 1993 (-) Feb 1997 (-)	2.6
Rioja	Tendencia	Sí	2.5	-4,9		Ene 1994 (-) Dic 1994 (+)	3.5

3.1.5 Conclusiones

La utilización de los modelos ha permitido hacer un uso intensivo de la información existente de la encuesta en periodos anteriores. Aunque el uso de esta información no es nuevo en la depuración e imputación de datos, las herramientas que hemos construido a partir de los modelos utilizan de manera más eficiente que los métodos tradicionales todo el pasado de la serie y permiten hacer inferencias probabilísticas y construir intervalos de confianza. Con ello hemos conseguido especificar y establecer los valores extremos de los edits, lo que habitualmente se hacía a partir de la experiencia del experto de la encuesta. Se han construido un conjunto de herramientas de depuración selectiva que permiten detectar y corregir los mayores errores. Se ha experimentado con datos reales del IPI. De acuerdo a estos resultados, el uso de modelos de series temporales puede ser muy útil para ahorrar tiempo en la depuración de los indicadores a corto plazo. Del mismo modo, se han obtenido un conjunto de características del comportamiento dinámico de los índices, de utilidad para la macrodepuración y el análisis exploratorio de datos. Se ha experimentado con tres tipos de imputaciones para manejar la falta de respuesta de uno o más datos de la empresa $q_{i,j_o,t}$ que forman parte de los índices $I_{i,t}$:

(i) Un imputación tradicional basada en el dato del periodos anterior y de los datos disponibles de otras empresas

$$\hat{q}_{i,j_0,t} = q_{i,j_0,t-1} \frac{\sum_{j \neq j_o} q_{i,j,t}}{\sum_{j \neq j_o} q_{i,j,t}}$$

(ii) Una microimputación basada en modelos de $q_{i,j_o,t}$ es decir $\hat{q}_{i,j_o,t}$

(iii) Una macroimputación basada en modelos para el índice $I_{i,t}$ es decir $\hat{I}_{i,t}$

La elección entre los métodos anteriores se puede hacer en función de diversas circunstancias. Si la tendencia de las empresas dentro de la rama de actividad es similar, la imputación tradicional por lo general funciona correctamente. Sin embargo, si la tendencia no es similar una microimputación basado en el modelo funciona mejor. Por otra parte, si la tasa de no-respuesta es baja, una imputación micro (tradicional o basada en modelos) por lo general funciona correctamente. Por el contrario, si la tasa de no-respuesta para es alta y no hay muchas empresas en el índice es preferible una macroimputación basada en modelos para todo el índice. Queda abierta como línea de investigación el estudio de un enfoque más sistemático para las imputaciones.

3.2 La depuración selectiva como un problema de optimización estocástica

3.2.1 Introducción

La depuración es una de las fases más costosas del proceso de producción de estadísticas. No sólo consume recursos económicos sino también tiempo. Adicionalmente, si el número de recontactos es grande aumenta de manera significativa la carga de respuesta. Por tanto, disponer de métodos eficientes de depuración es fundamental para los organismos estadísticos. Actualmente, la depuración manual exhaustiva es considerada ineficiente, puesto que la mayor parte del trabajo de depuración no tiene consecuencias en el nivel agregado e incluso puede deteriorar la calidad de los datos (ver Berthelot y Latouche (1993), y Granquist.(1997)).

Los métodos de depuración selectiva son estrategias para seleccionar un subconjunto de los cuestionarios recogidos en una encuesta para someterlos a una depuración minuciosa. Esta depuración puede hacerse de formas diversas, pero en general implica la intervención humana y por tanto, un gasto que conviene reducir. Una razón por la que es conveniente seleccionar algunos cuestionarios es que depurando ciertas unidades es más probable que mejore la calidad que si se depuran otras. Esto puede ser debido a que unas son más sospechosas de contener un error o a que en caso de tenerlo, éste probablemente tenga más impacto en los agregados. Por tanto, es razonable suponer que una buena estrategia de depuración selectiva puede permitir que se compatibilicen dos objetivos: buena calidad de las estimaciones agregadas y reducido trabajo de depuración. El enfoque actualmente predominante es el de calcular algún tipo de función score (FS) que asigne una puntuación a cada unidad, priorizando la depuración de las unidades con mayor puntuación. Cuando para cada unidad se recogen varias variables, se pueden calcular diferentes FS locales y combinarlas en una global. Así, las unidades cuya FS excede un cierto umbral, son depuradas manualmente. En consecuencia, es necesario decidir: (a) las FS locales; (b) cómo combinarlas en una global (suma, máximo, etc.) y (c) el umbral. Hasta ahora, estas cuestiones han sido tratadas de manera empírica debido a la falta de una base teórica. En Latouche y Berthelot (1992), Granquist.(1997) y Lawrence y McKenzie R. (2000) se proponen algunas directrices, pero en esencia se depende del criterio del experto.

Este subcapítulo se estructura en siete apartados. En el 3.2.1 se realiza la introducción En el apartado 3.2.2, se plantea formalmente el problema, en los apartados 3.2.3 y 3.2.4 mostramos como resolver dos variantes del problema, en el 3.2.5 se propone un método para calcular ciertos momentos condicionales que se requieren para obtener la solución, en el 3.2.6 presentamos los resultados de una aplicación práctica y finalizamos en el 3.2.7 con unas conclusiones.

3.2.2 El problema de la depuración selectiva

Introduzcamos un poco de notación,

- x_t^{ij} es el valor *verdadero* de la variable j en el cuestionario i en el periodo t, con $i = 1,\ldots,N$ y $j = 1,\ldots,q$.
- $\tilde{x}_t^{ij} = x_t^{ij} + \varepsilon_t^{ij}$ es el valor *observado*, siendo ε_t^{ij} el error de observación.
- $X_t^k = \sum \omega_{ij}^k x_t^{ij}$ es el k-ésimo estadístico calculado con los valores verdaderos, donde $k = 1,\ldots,p$.

Suponemos que χ_t^{ij} y ε_t^{ij} son variables aleatorias con respecto al espacio probabilístico (Ω, F, P). Puede haber otras variables relevantes, como $\tilde{x}_t^{ij}$, x_s^{ij} para $s < t$ o incluso variables de otras encuestas. Denotaremos por G_t la σ-álgebra generada por toda la información disponible hasta t. Para simplificar la notación, omitimos el subíndice t cuando no hay riesgo de ambigüedad.

Nuestro objetivo es encontrar una adecuada estrategia de selección, que debería indicar para cada i si el cuestionario va a ser depurado o no, empleando la información disponible. En realidad, vamos a permitir que se determine sólo la probabilidad de depuración de una unidad, dejando una cierta indeterminación.

Definición 2. 1. *Una estrategia de selección (ES) con respecto a* G_t *es un vector aleatorio* G_t*-medible,* $r = (r_1,\ldots,r_N)^T$ *tal que* $r_i \in [0,1]$.

Denotamos por $s(G_t)$ el conjunto de todas las ES con respecto a G_t. La interpretación de r es que el cuestionario i es depurado con probabilidad $1 - r_i$. El permitir $0 \leqslant r_i \leqslant 1$ en lugar de obligar a $r_i \in \{0,1\}$ es conveniente tanto desde el punto de vista teórico como del práctico, porque así el conjunto de estrategias es convexo y podemos emplear técnicas de optimización más sencillas que las de la programación entera. Si para un cierto i, $r_i \in (0,1)$, la unidad es depurada si $\chi_t^i < r_i$, donde χ_t^i es una variable aleatoria uniforme en $[0,1]$, e independiente de cualquier otra de las que se consideren. Denotamos por $\tilde{r}_i$ la variable indicatriz del suceso $\chi_t^i < r_i$ y $\tilde{r} = (\tilde{r}_1,\ldots,\tilde{r}_N)^T$. Si una ES cumple que $r_i \in \{0,1\}$ c.s., $\tilde{r} = r$ c.s. y decimos que r es entera. En nuestra aplicación, las soluciones son aproximadamente enteras. También es conveniente tener una definición del concepto de función *score*.

Definición 2.2. *Sea* r *una ES,* $\delta = (\delta_1,\ldots,\delta_N)^T$ *un vector aleatorio y* $\Theta \in R$*, tal que* $r_i = 1$ *si y solo si* $\delta_i \leqslant \Theta$. *Entonces, decimos que* δ *es una función score que genera* r *con el umbral* Θ.

Para plantear formalmente nuestro problema, suponemos que tras la depuración manual, se obtienen los valores verdaderos. Así, solo tenemos que considerar los valores observados y verdaderos. Definimos el estadístico depurado como el calculado con los valores obtenidos tras depurar los que hayan sido seleccionados y lo denotamos por $X^k(r) = \sum \omega_{ij}^k (x_t^{ij} + \tilde{r}_i \varepsilon_t^{ij})$.

La calidad de $X^k(r)$ debe ser medida con arreglo a alguna función de pérdida. En este trabajo, solo consideramos el error cuadrático, $(X^k(r) - X^k)^2$, quedando la adaptación a otras funciones para futuros desarrollos. El valor de la función de pérdida se puede escribir como

$$(X^k(r) - X^k)^2 = \sum_{i,i'} \in_i^k \in_{i'}^k \tilde{r}_i \tilde{r}_{i'}, \qquad (2.1)$$

donde $\in_i^k = \Sigma_j \omega_{ij}^k \varepsilon_t^{ij}$. En forma matricial como $(X^k(r) - X^k)^2 = \tilde{r}' E^k \tilde{r}$, con $E^k = \{E_{i,i'}^k\}_{i,i'}$ y $E_{i,i'}^k = \in_i^k \in_{i'}^k$.

Ahora podemos plantear el problema de la selección como un problema de optimización:

$$[P_Q] \quad \max_r \quad \mathbf{E}[1^T \tilde{r}]$$

$$\text{sujeto a} \quad \begin{array}{l} r \in S(G_t) \\ \mathbf{E}[\tilde{r}^T E^k \tilde{r}] \leqslant e_k^2, k = 1, \ldots, p \end{array}$$

Éste es un problema de optimización lineal con restricciones cuadráticas, que tiene la dificultad de que se busca la solución en un espacio de dimensión infinita. En la sección 3.2.4 mostramos como resolverlo.

Analicemos ahora la expresión (2.1). Podemos descomponerla como

$$(X^k(r) - X^k)^2 = \sum_i (\in_i^k)^2 \tilde{r}_i + \sum_{i \neq i'} \in_i^k \in_{i'}^k \tilde{r}_i \tilde{r}_{i'}. \qquad (2.2)$$

El primer término de la derecha de (2.2) mide el impacto individual de cada error, independientemente de su signo. En el segundo término, los sumandos son negativos cuando los factores tienen signos opuestos. Por tanto, para reducir el error total, sería conveniente dejar sin depurar parejas de unidades con distinto signo. La no-linealidad del segundo término complica los cálculos, así que estudiaremos también la versión despreciando esa parte. Si definimos $D^k = (D_1^k, \ldots, D_N^k)^T$, con $D_i^k = (\varepsilon_i^k)^2$, podemos plantear

$$[P_L] \quad \max_r \quad \mathbf{E}[1^T \tilde{r}]$$

$$\text{s.a.} \qquad r \in S(G_t), \mathbf{E}[D^k \tilde{r}] \leqslant e_k^2, k = 1, \ldots, p.$$

En la sección 3.2.3 veremos que la solución viene dada por una cierta FS. Puesto que no hay justificación teórica para despreciar la parte cuadrática, el problema lineal debe justificarse empíricamente mediante su eficacia en la práctica.

3.2.3 Caso lineal

Vamos a resolver el problema $[P_L]$ mediante un método de dualidad, pero antes debemos expresar la función objetivo como $E[1^T r]$ y la restricción como $E[\Delta^k r] \leq e_k^2$, donde $\Delta^k = E[D^k \mid G_t]$. Definamos ahora la lagrangiana $L(r,\lambda) = \mathbf{E}[1^T r] - \lambda^T \mathbf{E}[\Delta^k r - e_k^2]$. En [1], se demuestra que bajo hipótesis no muy restrictivas, si $\bar{\lambda}$ es una solución del problema dual,

$$[D] \quad \min_{\lambda \geq 0} \quad \varphi(\lambda)$$

con $\varphi(\lambda) = \max_r L(r,\lambda)$, entonces $r = \arg\max L(.,\bar{\lambda})$ es una solución del problema primal $[P_L]$. Como Δ^k es conocido, la maximización de L respecto a r se reduce al problema determinista

$$\max_r \quad 1^T r - \sum_k \lambda_k (\Delta^k r - e_k^2) \qquad (3.1)$$

$$\text{s.a.} \qquad r_i \in [0,1] \qquad (3.2)$$

Aplicando las condiciones de Karush-Kuhn-Tucker (ver [2]) tenemos que,

$$r_i = \begin{cases} 1 & \text{si} \quad \lambda^T \Delta_i < 1 \\ 0 & \text{si} \quad \lambda^T \Delta_i > 1 \end{cases} \qquad (3.3)$$

donde $\Delta_i = (\Delta_i^1, \ldots, \Delta_i^p)^T$. El caso $\lambda^T \Delta_i = 1$ se puede despreciar como un suceso de probabilidad cero, ya que para datos cuantitativos, la distribución de Δ_i es continua. Por tanto, la solución a $[P_L]$ es la ES generada por la FS $\delta_i = \lambda^T \Delta_i$ con el umbral 1.

En el apartado 3.2.5 describimos como usar un modelo para el cálculo de Δ^k. Para estimar $\varphi(\lambda)$ cambiamos la esperanza por la media muestral, donde la muestra puede ser simulada u obtenida a partir de datos reales si se dispone de ellos. Si tenemos una muestra de M valores, usamos la estimación

$$\hat{\phi}(\lambda) = \frac{1}{M} \sum_{l=1}^{M} L(\lambda, r^l(\lambda)).$$

La minimización de $\hat{\varphi}$ se hará por métodos numéricos.

3.2.4 Caso cuadrático

El problema cuadrático plantea dificultades particulares. Si intentamos aplicar el método de dualidad como en el apartado anterior, expresamos la restricción como $E[r^T \Gamma^k r + (\Delta^k)^T r] \leq e_k^2$, donde $\Gamma^k = \{\Gamma_{ij}^k\}_{ij}$ y

$$\Gamma_{ij}^k = \begin{cases} \mathbf{E}[\in_i^k \in_j^k | G_t] & \text{if} \quad i \neq j \\ 0 & \text{if} \quad i = j \end{cases},$$

pero por desgracia, las matrices Γ^k son indefinidas, luego la restricción no es convexa.

Podemos superar esta dificultad reemplazando la restricción por otra convexa y que bajo ciertas hipótesis es equivalente a la original.

Consideremos la restricción $E[r^T M^k r + (v^k)^T r] \leq e_k^2$, donde $M_{ij}^k = m_i^k m_j^k$, $m_i^k = \mathbf{E}[\in_i^k | G_t]$, $v_i^k = V[\in_i^k | G_t]$. Se puede demostrar (ver Arbúes et al. (2008)) que si r es una solución entera al problema con la nueva restricción y $E[\in_i^k \in_j^k | G_t] = m_i^k m_j^k$ para $i \neq j$, entonces r es una solución al problema original. Hemos comprobado que la condición de integridad se cumple aproximadamente en nuestra aplicación. La hipótesis sobre la esperanza condicionada es restrictiva y su relajación puede ser estudiada en futuros desarrollos. En el apartado 3.2.5 describimos un método para obtener Σ^k y v^k. En Arbúes et al. (2008) se describe como resolver con poco coste computacional el problema de programación cuadrática que resulta.

3.2.5 Momentos condicionales basados en un modelo

La aplicación práctica de los resultados de las secciones anteriores requiere un método para calcular los momentos condicionales del error que aparecen. En esta sección omitimos el índice j para simplificar la notación, pero los resultados son aplicables para el caso en el que hay varias variables por cuestionario.
Sea H_t la σ- álgebra generada por toda la información disponible en t a excepción de $\tilde{x}_t^i$. Así, $G_t = \sigma(\tilde{x}_t^i, H_t)$. Sea $\hat{x}_t^i$ un predictor de x_t^i calculado empleando la información de H_t, es decir, H_t-medible. Denotamos el error de predicción por $\xi_t^i = x_t^i - \hat{x}_t^i$. Supongamos que,

(a) ξ_t^i y η_t^i están distribuidas como normales independientes con media cero y varianzas v_i^2 y σ_i^2 respectivamente.

(b) $\varepsilon_t^i = \eta_t^i e_t^i$, donde e_t^i tiene una distribución de Bernoulli tomando los valores 1 y 0 con probabilidades p y 1-p y es independiente de ξ_t^i y η_t^i.

(c) ξ_t^i, η_t^i y e_t^i son conjuntamente independientes de H_t.

Bajo estas hipótesis, los momentos condicionales de ε_t^i con respecto a G_t dependen solo de $u_t^i = \hat{x}_t^i - \tilde{x}_t^i$, es decir, de la diferencia entre el valor predicho y el observado. En las fórmulas siguientes, también omitimos i y t por simplicidad.

$$\mathbf{E}[\varepsilon \mid G] = \frac{\sigma^2}{\sigma^2 + \nu^2} u\zeta \qquad (5.1)$$

$$\mathbf{E}[\varepsilon^2 \mid G] = \left[\frac{\sigma^2 \nu^2}{\sigma^2 + \nu^2} + \left(\frac{\sigma^2}{\sigma^2 + \nu^2} \right)^2 u^2 \right] \zeta \qquad (5.2)$$

donde,

$$\zeta = \frac{1}{1 + \frac{1-p}{p} \left(\frac{\nu^2}{\sigma^2 + \nu^2} \right)^{-1/2} \exp\{- - \frac{u^2(\sigma^2)}{2\nu^2(\sigma^2 + \nu^2)}\}}. \qquad (5.3)$$

3.2.6 Caso práctico

En esta sección presentamos los resultados de la aplicación de los métodos descritos a los datos de la encuesta de Cifras de Negocios y Entradas de Pedidos en la industria que lleva a cabo el Instituto Nacional de Estadística (datos disponibles en http://www.ine.es/inebmenu/mnu_industria.htm). En el momento del estudio, se disponía de datos mensuales desde enero de 2002 hasta septiembre de 2006 ($t = 1\ldots,57$) recogidos para una muestra de alrededor de $N = 13{,}500$ unidades. Solo dos de las variables recogidas fueron consideradas: importe neto de la cifra de negocios (x_t^{i1}) y nuevos pedidos recibidos (x_t^{i2}). Estas dos variables son agregadas separadamente para obtener los dos indicadores, así que $p = q = 2$ y $\omega_{i2}^1 = \omega_{i1}^2 = 0$.

Para aplicar (5.1)-(5.3) es necesario ajustar un modelo a los datos, pero antes empleamos la transformación $y_t^{ij} = \log(x_t^{ij} + m)$, donde m es una constante positiva ajustada por máxima verosimilitud ($m \approx 10^5$€). Para recuperar los momentos condicionales de la variable original, se pueden usar las propiedades de la log-normal o un desarrollo de Taylor de primer orden, que resulta $E[(\tilde{x}_t^{ij} - x_t^{ij})^2 \mid G_t] \approx (\tilde{x}_t^{ij} - m)^2 E[(\tilde{y}_t^{ij} - y_t^{ij})^2 \mid G_t]$. En nuestra aplicación hemos empleado la aproximación. También hemos observado que si $\tilde{x}_t^{ij} - m$ es sustituido por una media de los últimos 12 valores de $\tilde{x}_t^{ij}$ el resultado es más robusto frente a valores pequeños de $\tilde{x}_t^{ij} - m$.

El modelo que hemos planteado para las variables transformadas es muy simple. Suponemos que x_t^{ij} y $x_t^{i'j'}$ son independientes si $(i, j) \neq (i', j')$. Para cada par (i, j) seleccionamos para la serie x_t^{ij} uno de los siguiente modelos

$$(1 - B)\, y_t^{ij} = a_t \qquad (6.1)$$

$$(1 - B^{12})\, y_t^{ij} = a_t \qquad (6.2)$$

$$(1 - B^{12})(1 - B)\, y_t^{ij} = a_t \qquad (6.3)$$

donde B es el operador de retardos $Bu_t = u_{t-1}$ y a_t es un proceso de ruido blanco. El criterio de selección es el de mínimo error residual cuadrático. Una vez seleccionado el

modelo, lo empleamos para calcular la predicción $\hat{y}_t^{ij}$ y su desviación típica v_{ij}. La desviación típica *a priori* del error de observación y de la probabilidad de error se suponen constantes (la transformación logarítmica hace que tenga sentido suponer que los errores de observación tienen la misma d.t.), los denotamos por σ_j y p_j con $j=1,2$ y son estimados usando datos históricos de la encuesta. Para ello, nos aprovechamos de que tanto la primera versión recibida de cada dato como las que posteriormente resulten de posibles correcciones quedan almacenadas en una base de datos. Para este estudio, la primera versión es considerada como dato *observado* y la definitiva como dato *verdadero.*

Una vez que hemos calculado σ_j, p_j, v_{ij} y u_t^{ij}, usamos (5.1)-(5.3) para calcular los momentos condicionales, Δ^k, Σ^k y v^k.

Pretendemos ahora comparar la eficacia de nuestro método con la función score descrita en Hedlin (2003), $\delta_i^0 = \omega_i \,|\, \tilde{x}^i \;\; \hat{x}^i \,|$, donde $\hat{x}_i$ es una predicción de x_i. En Hedlin (2003) se propone usar como predicción el último valor de la misma variable en periodos anteriores. Hemos considerado una modificación de esta función, δ^1, simplemente cambiando la predicción por la que se obtiene mediante los modelos (6.1)-(6.3). Finalmente, δ^2 es la FS calculada usando (5.1)-(5.3). En esta comparación, la combinación de FS locales en una global se ha hecho simplemente sumándolas. Medimos la efectividad de una FS mediante $E^j = \sum_n E^j(n)$, con $E^j(n) = \left[\sum_{i \geqslant n}^{N} \omega_i^j (\tilde{x}^{ij} - x^{ij})\right]^2$, donde consideramos las unidades ordenadas descendentemente según la correspondiente FS. La cantidad $E^j(n)$ puede ser considerada como una estimación del error que queda tras editar n unidades. Los resultados incluidos en la tabla 1 muestran que δ^2 mejora a δ^1, que a su vez es más eficaz que δ^0.

Tabla 1. Comparación de FS.

FS	Cifras de Negocios	Nuevos Pedidos
δ^0	0.44	1.33
δ^1	0.38	0.45
δ^2	0.26	0.37

3.2.7 Conclusiones

Hemos descrito un marco teórico para tratar el problema de la depuración selectiva, definiendo el concepto de estrategia de selección. La búsqueda de una adecuada ES se presenta como un problema de optimización lineal con restricciones cuadráticas. Consideramos también una versión modificada con restricciones lineales. Mostramos un método práctico para resolver ambos problemas. La FS correspondiente a la versión lineal mejora la de referencia.

3.3 Propuesta y desarrollo de un marco teórico de depuración e imputación basado en modelos y optimización

3.3.1 Introducción

La diversidad de errores de medida ha dado lugar a la aparición de diferentes técnicas y algoritmos para su detección y tratamiento, como la depuración interactiva, la depuración automática, la depuración selectiva, o la macrodepuración (véase de Waal et al. (2011) para una revisión completa). Hoy en día se acepta ampliamente que ningún método o enfoque puede hacer frente a todo tipo de errores. Por lo tanto, deben ser combinados en una estrategia global de depuración e imputación.

Los sistemas generalizados han supuesto un gran avance. Sin embargo, no resuelven la especificación de los edits, que deben ser introducidos por los usuarios. Habitualmente, los usuarios establecen los edits de acuerdo a su experiencia práctica, sin que exista un marco teórico. Por su parte, la depuración selectiva se centra en los errores influyentes. En las últimas dos décadas, esta modalidad de depuración ha sido reconocida como un elemento clave en las estrategias de depuración e imputación, pero, hasta la fecha, únicamente se ha tratado de forma heurística.

Dentro de este contexto, el trabajo que aquí se presenta se centra en dos aspectos principales: el uso de modelos estadísticos para resolver los problemas de la especificación de los edits y la utilización de técnicas de optimización para resolver los problemas que presenta la depuración selectiva. Ambos aspectos están estrechamente relacionados, ya que las técnicas de optimización que proponemos necesitan de modelos para substanciar las restricciones y calcular las matrices de las funciones de pérdida. Adicionalmente, los modelos pueden emplearse en la macrodepuración y en la imputación.

En los epígrafes precedentes se han abordado ya estos temas. Así, en el primer bloque de investigaciones (subcapítulo 3.1), se introducen modelos que llevan a cabo todos los objetivos que pretendemos, es decir, la especificación de los edits, y el soporte a la depuración selectiva, la macrodepuración y la imputación. Sin embargo, se hace de una manera parcial, limitándose al uso de modelos univariantes de series temporales y aplicables a datos cuantitativos de encuestas coyunturales. Por su parte, el segundo bloque de investigaciones (subcapítulo 3.2) introduce las técnicas de optimización como solución formal al problema de la depuración selectiva, pero se centra únicamente en la optimización estocástica. En este último bloque de investigaciones, se intenta avanzar en el grado de generalización y formalización, introduciendo un marco general de utilización de modelos y abordando un problema de optimización general, del que se derivan una versión estocástica y otra combinatoria.

Este subcapítulo se estructura en cinco apartados. En el 3.3.1 se realiza la introducción. En el 3.3.2 se analizan distintos modelos estadísticos en función de la información auxiliar disponible. En el 3.3.3 se formula el problema de optimización genérica, que nos conduce a una versión de optimización estocástica o combinatoria dependiendo del momento en que se realice y por tanto del flujo de información disponible. En el apartado 3.3.4 se lleva a cabo la comparación de nuestra propuesta con funciones score

de uso habitual, utilizando datos reales del Índice de Cifras de Negocio y del Índice de Entrada de Pedidos. Por último, se incluyen unas conclusiones en el apartado 3.3.5.

3.3.2 El uso de modelos estadísticos

Como un criterio esencial de la depuración e imputación se considera que los datos a difundir sean consistentes con la información existente hasta la fecha. La información existente se resume en un modelo**.** El modelo debe extraer al máximo la información que contienen los datos e incorporar la información *a priori.*

Para hacer operativo el concepto de que los datos sean consistentes se considera que un dato es consistente cuando se acerque a la predicción que ofrece este modelo para ese dato. Las herramientas de depuración e imputación se basan fundamentalmente en las predicciones del modelo. Los outliers se definen y se detectan en función de cuánto se alejan los datos observados (a depurar) de la predicción. Los valores imputados se definen como la predicción que se deduce del modelo para ese dato.

En este trabajo se desarrollan nuevas herramientas de depuración e imputación basadas en los modelos (fundamentalmente en la función de predicción). Estos modelos se utilizan tanto en la microdepuración como en la macrodepuración y en la depuración selectiva. Las propiedades de la depuración e imputación pueden derivarse a partir de la predicción de los modelos.

Al hacer frente a las tareas de depuración e imputación se recurre a la mejor información auxiliar disponible en ese momento. Con plena generalidad, esta incluirá: (i) los valores observados de las variables de análisis $y_k^{(t)}$ para el presente ($t = T$) y el pasado ($t < T$), (ii) los valores verdaderos de estas variables $y^{(0,t)}$ para aquellas unidades depuradas en el pasado $t < T$ (iii) y los valores de las covariables auxiliares $x_k^{(t)}$ para los períodos de los que se dispone de información. Por tanto, tenemos $y_k = y_k^{(T)}$, $y_k^0 = y_k^{(0,T)}$ y $x_k = y_k^{(t_1)}, y_k^{(0,t_1)}, x_k^{(t_2)}$, con $t_1 < T$ y $t_2 \leqslant T$. Téngase en cuenta que alguno de estos valores puede ser coincidente (por ejemplo, cuando el error de medición es nulo) y que y_k^0 sólo se conoce después de llevar a cabo el trabajo de depuración. También conocemos (por lo menos podemos conocer) una predicción puntual $\hat{y}_k$ para cada variable en función de estas variables auxiliares. Adicionalmente, se puede eventualmente incorporar información a priori. Por ejemplo, podemos hacer uso de un modelo univariante de series temporales $\left\{y_k^{(0,t)}\right\}_{t<T}$ para hacer una predicción puntual $\hat{y}_k^{(T)}$. Surgen diferentes opciones dependiendo de la cantidad y tipo de información auxiliar. Estas predicciones entrarán en el problema de selección como covariables auxiliares, de modo que $x_k = y_k^{(t_1)}, y_k^{(0,t_1)}, \hat{y}_k^{(T)}, x_k^{(t_2)}$, con $t_1 < T$ y $t_2 \leqslant T$.

Llamemos m^* al modelo auxiliar utilizado para hacer las predicciones $\hat{y}_k$ que no debe confundirse con el modelo de error de medición. Este modelo de error de medición se da como es habitual en términos de (i) la distribución condicional de los valores de predicción de los valores verdaderos y^0 y (ii) la distribución de los y^0 condicionada a la información auxiliar disponible X. Para una variable supondremos $y_k = y_k^0 + \varepsilon_k^{obs}$ y $y_k^0 = \hat{y}_k + \varepsilon_k^{pred}$ En otras palabras, estamos usando el valor predicho calculado de acuerdo

con el modelo auxiliar m^* como una variable exógena para el modelo de relación y^0. En este sentido nos referimos a esta propuesta como un modelo de observación-predicción.

Generalizando estas ideas, consideremos:

(i) un modelo de observación $P_{obs|0}\left(y \mid y^0\right)$, es decir, una distribución de probabilidad condicionada para los valores observados, dados los valores verdaderos y^0 ;

(ii) un modelo de predicción $P_{0|pred}\left(y^0 \mid \hat{y}\right)$, es decir, una distribución de probabilidad condicionada para los valores verdaderos, dados los valores predichos y^0 de acuerdo con un modelo auxiliar m^*.

Ahora vamos a denotar $P_{obs|pred}$ por la distribución de probabilidad de los valores observados condicionadas a los previstos $\hat{y}$ y por $P_{0|obs,pred}$ la distribución de probabilidad de los valores verdaderos y^0 condicionada a los valores observados y^{obs} y los valores predichos $\hat{y}$. Entonces por el teorema de Bayes o una generalización del mismo, podemos escribir

$$P_{0|obs,pred} = \frac{P_{obs|0} \times P_{0|pred}}{P_{obs|pred}} \qquad (2)$$

El producto debe ser entendido en una forma generalizada adecuado cuando las distribuciones son completamente generales. Como de costumbre, si las distribuciones de probabilidad son absolutamente continuas con función $f(.)$ de densidad, la ecuación (2) puede ser fácilmente reconocida como

$$f_{0|obs,pred}(\mathbf{y}^0) = \frac{f_{obs|0}(\mathbf{y}^{obs} \mid \mathbf{y}^0, \hat{\mathbf{y}})\, f_0(\mathbf{y}^0 \mid \hat{\mathbf{y}})}{\int_{R^Q} f_{obs|0}\left(\mathbf{y}^{obs} \mid \mathbf{y}^0, \hat{\mathbf{y}}\right) f_0(\mathbf{y}^0 \mid \hat{\mathbf{y}})\, d\mathbf{y}^0}.$$

El caso discreto también se reduce a aplicar el teorema de Bayes. Una vez que tenemos la distribución $P_{0|obs,pred}$, las matrices de pérdida pueden ser calculadas como

$$M^{(q)} = E_{0|obs,pred}\left[\Delta^{(q)} \mid S, Y, \hat{Y}\right] \qquad (3)$$

Para ilustrar esta propuesta, consideremos el siguiente ejemplo genérico con una variable continua. Vamos a definir el modelo de observación $y_k^{obs} = y_k^0 + \varepsilon_k^{obs}$ y de modelos de prediccion
$y_k^0 = \hat{y}_k + \varepsilon_k^{pred}$, con las siguientes especificaciones

1. $\varepsilon_k^{obs} = \eta_k^{obs} e_k$.
2. $e_k \approx Be(p_k)$, donde $p_k \in (0,1)$.

3. $\left(\varepsilon_k^{pred}, \eta_k^{obs}\right) \approx N\left(\mathbf{0}, \begin{pmatrix} \nu_k^2 & \rho_k\sigma_k\nu_k \\ \rho_k\sigma_k\nu_k & \sigma_k^2 \end{pmatrix}\right)$.
4. ε_k^{pred}, η_k^{obs} y e_k son conjuntamente independientes de Z_k^{co}.
5. e_k es independiente de ε_k^{pred} y η_k^{obs}.

Estas especificaciones equivalen a decir que la unidad tiene una probabilidad $1 - p_k$ de declarar un valor sin error de medida ($y_k = y_k^0$) y, si informan de un valor erróneo, el error de medida se distribuye como una variable aleatoria normal con media cero y varianza σ_k^2. Por otra parte, el error de predicción se distribuye como una variable aleatoria normal con media cero y varianza ν_k^2. Ambos errores se distribuyen conjuntamente como una variable aleatoria normal bivariante con correlación ρ_k. Informar de un valor erróneo es independiente de los dos tipos de errores.

Por el momento, vamos a suponer que los parámetros $\theta = (p_k, \sigma_k^2, \nu_k^2, \rho_k)^T$ son conocidos. Centrémonos en la función de pérdida al cuadrado. Arbués et al. (2012a) demuestran que

$$E_m\left[\left(y_k - y_k^0\right) | s_k, y_k, \hat{y}_k\right] = \nu_k \quad \frac{\sigma_k^2 + \rho_k\sigma_k\nu_k}{\sigma_k^2 + \nu_k^2 + 2\rho_k\sigma_k\nu_k}\left(\frac{y_k - \hat{y}_k}{\nu_k}\right) \quad \zeta_k \quad \left(\frac{y_k - \hat{y}_k}{\nu_k}\right), \tag{4}$$

$$E_m\left[\left(y_k - y_k^0\right)^2 | s_k, y_k, \hat{y}_k\right] = \nu_k^2 \cdot \left(\frac{\sigma_k^2 + \rho_k\sigma_k\nu_k}{\sigma_k^2 + \nu_k^2 + 2\rho_k\sigma_k\nu_k}\right)^2 \cdot \left[\frac{\sigma_k^2(1-\rho_k^2)(\sigma_k^2 + \nu_k^2 + 2\rho_k\sigma_k\nu_k)}{\left(\sigma_k^2 + \rho_k\sigma_k\nu_k\right)^2} + \left(\frac{y_k - \hat{y}_k}{\nu_k}\right)^2\right] \zeta_k\left(\frac{y_k - \hat{y}_k}{\nu_k}\right),$$

$$E_m\left[\left(y_k - y_k^0\right)\left(y_l - y_l^0\right) | s_k, y_k, \hat{y}_k\right] = E_m\left[\left(y_k - y_k^0\right) | s_k, y_k, \hat{y}_k\right] E_m\left[\left(y_l - y_l^0\right) | s_k, y_k, \hat{y}_k\right], \quad k \neq l,$$

$$\text{donde } \zeta_k(x) = \frac{1}{1 + \frac{1 - p_k}{p_k}\left(\frac{\nu_k^2}{\sigma_k^2 + \nu_k^2 + 2\rho_k\sigma_k\nu_k}\right)^{-1/2} \exp\left(-\frac{1}{2}\frac{\sigma_k^2 + 2\rho_k\sigma_k\nu_k}{\sigma_k^2 + \nu_k^2 + 2\rho_k\sigma_k\nu_k}x^2\right)}.$$

Si nos centramos en la función de pérdida absoluta, bajo las mismas hipótesis, Arbués et al. (2013) demuestran que

$$E_m\left[\left|y_k - y_k^0\right| | s_k, y_k, \hat{y}_k\right] = \sqrt{\frac{2}{\pi}} \cdot \nu_k \cdot {}_1F_1\left(-\frac{1}{2}; \frac{1}{2}; \frac{(y_k - \hat{y}_k)^2}{2\nu_k^2}\right) \cdot \zeta_k\left(\frac{y_k - \hat{y}_k}{\nu_k}\right) \tag{5}$$

Donde ${}_1F_1(a; b; x)$ denota la función hipergeométrica confluente de la primera clase.

La estimación de los parámetros θ depende del escenario. Para el problema estocástico, como antes, nos vemos obligados a utilizar algunos valores de referencia o medidas heurísticas. Una vez más se recurre a la información auxiliar. La elección depende en gran medida de la cantidad y tipo de la información auxiliar. Desde el histórico de dos conjuntos de datos que comprende períodos de tiempo anteriores (por ejemplo, un panel fijo), podemos calcular

$$\hat{p}_k = \frac{1}{\tau}\sum_{t=1}^{\tau} I_{y_k^{(t)} \neq y_k^{(0,t)}},$$

$$\hat{\sigma}_k^2 = \frac{1}{\tau - 1}\sum_{t=1}^{\tau}\left(\varepsilon_k^{(t)} - \bar{\varepsilon}_k\right)^2,$$

$$\hat{v}_k^2 = \frac{1}{\tau - 1}\sum_{t=1}^{\tau}\left(\varepsilon_k^{(t)} - \bar{\varepsilon}_k\right)^2, \quad \text{donde } \varepsilon_k^{(t)} = \hat{y}_k^{(t)} - y_k^{(0,t)},$$

$$\hat{\rho}_k = \frac{1}{\tau - 1}\sum_{t=1}^{\tau}\left(\varepsilon_k^{(t)} - \bar{\varepsilon}_k\right)\left(\varepsilon_k^{(t)} - \bar{\varepsilon}_k\right).$$

En el caso de paneles rotatorios o de diseños muestrales con muy corta continuidad de las unidades en la muestra, nos vemos obligados a hacer supuestos simplificadores, como la partición de la muestra $s = \cup_{i=1}^{I} s_i$ y postular $\theta_k = \theta_i$ si $k \in s_i$. También podemos adoptar estos supuestos para algunos de los parámetros. El caso extremo sería $\theta_k = \theta = (p, \sigma^2, v^2, \rho)^T$ para todos $k \in s$, que se puede complementar aún más con hipótesis adicionales, tales como $\rho = 0$ Por otro lado, para el problema de combinatorio tenemos (casi) la muestra actual completa, de modo que podamos hacer uso de estos datos, aunque con importantes limitaciones. Es claro que es imposible estimar cada uno θ_k utilizando sólo la muestra del periodo corriente. Estamos obligados a hacer algunos supuestos simplificadores, en línea con los anteriores.

Alternativamente, el modelo de contaminación propuesto por Di Zio y Guarnera (2013) es un ejemplo relevante de una técnica basada en modelos que utiliza exclusivamente datos del período actual para estimar los parámetros del modelo.

3.3.3 El problema de optimización

Antes de identificar las variables, la función objetivo y las restricciones de nuestro problema de optimización, se necesita introducir la siguiente notación. El diseño muestral, según el cual se ha seleccionado una muestra probabilística se denota $p(.)$. El tamaño muestral por *n* y los pesos muestrales por $w_{ks.}$ La dependencia de la muestra de los pesos de muestrales asume implícitamente que no necesitan ser los pesos de diseño. Por ejemplo, en un estimador de razón de la forma $\hat{Y}^{rat} = X \frac{\hat{Y}^{HT}}{\hat{X}^{HT}}$ donde x es una variable auxiliar conocida del marco muestral, $X = \sum_{k \in U} x_k$,es un total poblacional conocido, y $\hat{Y}^{HT} = \sum_{k \in s} \frac{y_k}{\pi_k}$ (análogamente para $\hat{X}^{HT}$) representa el estimador de Horvitz-Thompson del total poblacional $Y = \sum_{k \in U} y_k$, los pesos de muestreo están dados por $\omega_{ks} = \frac{X}{\hat{X}^{HT}}\frac{1}{\pi_k}$, donde π_k es la probabilidad de inclusión de primer orden para la unidad *k*. Situaciones más complejas se encajan en esta notación. Los valores verdaderos, observados y depurados de una variable del cuestionario $y^{(q)}$, $q = 1, \ldots, Q$ (para facilitar la notación se omite el superíndice (q) en adelante, excepto cuando sea estrictamente necesario), para la unidad k se indican, respectivamente, por y_k^0, y_k, e y_k^*.

Se asigna una variable binaria $r_k \in \{0,1\}$ a cada unidad *k* para indicar si se ha seleccionado para posterior depuración ($r_k = 0$) o no ($r_k = 1$). El vector $\mathbf{r} = (r_1, \ldots, r_n)^t$ definido para los elementos muestrales será denominado la *estrategia de selección*. Esta asignación nos permite relacionar los valores anteriores por la ecuación $y_k^*(\mathbf{r}) = (1 - r_k) \cdot y_k^0 + r_k \cdot y_k$ donde hemos hecho explícita la dependencia de los valores depurados sobre la *estrategia de selección*. Nótese que estamos asumiendo implícitamente que el trabajo de depuración nos conduce de los valores observados a los valores verdaderos. Si denotamos el error de medida $\varepsilon_k = y_k - y_k^0$, podemos escribir $y_k^*(r) = y_k^0 + r_k \varepsilon_k$. Los valores depurados son los que se utilizan para calcular los estimadores de la encuesta. Es decir, si hemos de estimar el total para un dominio U_d de la población $Y_{U_d} = \sum_{k \in U_d} y_k^0$ (se omitirá en adelante el subíndice U_d para facilitar la notación), el estimador correspondiente será $\hat{Y}^*(r) = \sum_{k \in s_d} \omega_{ks} y_k^*(r)$. Nos limitaremos a totales poblacionales y estimadores lineales. Todas las covariables auxiliares no incluidas en el cuestionario de la unidad k se indican mediante $\mathbf{x}_k$.

Conviene señalar que el estimador anterior calculado con los valores depurados pudiera no ser el estimador final para la difusión de la encuesta, ya que eventualmente podrian necesitarse de algunos procedimientos posteriores como el ajuste de los pesos o el tratamiento de valores atípicos. La selección de las unidades divide la muestra en un flujo crítico ($r_k = 0$) y en otro no crítico ($r_k = 1$). Aquí estamos suponiendo que las unidades del flujo critico serán sometidas a depuración manual y las otras se dejarán inalteradas. Sin embargo, este procedimiento de selección puede ser utilizado alternativamente con otras estrategias, como por ejemplo que las unidades del flujo crítico sean sometidas a depuración manual y las otras sean sometidas a un procedimiento de depuración automática.

Las variables mencionadas anteriormente se consideran variables aleatorias que siguen un modelo *m* en un espacio de probabilidad $(\Omega, \mathsf{F}, \mathsf{P})$. Esta aleatoriedad no se indica específicamente en la notación, a excepción de la *estrategia de selección*, de manera que $\mathbf{R}$ denota la *estrategia de selección aleatoria*, y $\mathbf{R}(\omega) = \mathbf{r}$ con $\omega \in \Omega$, será una realización numérica particular llamada la *selección*. La predicción de la variable y_k según el modelo *m* se denota por $\hat{y}_k$. Cuando se utilizan variables aleatorias en los estimadores de la encuesta, se escribe indistintamente $\hat{Y}^0 = \sum_{k \in s_d} \omega_{ks} y_k^0$, $\hat{Y} = \sum_{k \in s_d} \omega_{ks} y_k$ e $\hat{Y}^*(R) = \sum_{k \in s_d} \omega_{ks} y_k^*(R)$.

Se denota por $\mathbf{Z}$ el conjunto de variables aleatorias utilizadas para seleccionar las unidades destinadas a depuración en la estrategia de D&I. En particular, vamos a considerar dos opciones, a saber, cualquiera de los dos $Z = Z^{st} \equiv s$ para el problema estocástico o $Z = Z^{co} \equiv \{s, X, Y\}$ para la versión combinatoria. Cuando esta información auxiliar se limita a la unidad k vamos a escribir en consecuencia $Z_k^{st} = s_k$ o $Z_k^{co} = \{s_k, x_k, y_k\}$. El uso de la información auxiliar actúa condicionando las variables aleatorias correspondientes. Las covariables auxiliares se eligen de acuerdo con el modelo estadístico utilizado en el problema de optimización (ver más abajo). Juegan un

papel similar al de las variables auxiliares en el diseño muestral o en el proceso de calibrado, con las que pueden coincidir parcial o totalmente.

El problema de optimización general

En el problema de optimización se quiere minimizar el número de cuestionarios a depurar, a condición de que las funciones de pérdida elegidas para los estimadores de la encuesta dirigidos a los valores poblacionales estén acotadas. Para establecer formalmente el problema de optimización es necesario (i) las variables, (ii) la función de optimizar, y (iii) las restricciones. Aparte de la identificación de estos elementos, es importante para mostrar cómo la información auxiliar entra en la formulación del problema.

Las variables finales son la estrategia de selección $r^T = (r_1, \ldots, r_n)$ para las unidades de la muestra $s = \{1, \ldots, n\}$ donde $r_k = 0$ si se ha seleccionado la unidad k y $r_k = 1$ de otro modo. Sin embargo, ya que el error de medición $\varepsilon_k = y_k - y_k^0$ está concebido para ser de naturaleza aleatoria condicionada a la muestra realizada s y teniendo en cuenta la información auxiliar Z elegida para realizar la selección de unidades, esta selección puede variar dependiendo de los valores realizados de y, de y^0 y de Z. Por lo tanto R denota la estrategia de selección estocástica, de forma que (i) $R(\omega) = r$ es una selección particular realizada y (ii) $E_m[R \mid Z]$ es el vector de las probabilidades de no selección en el marco del modelo específico m teniendo en cuenta la información auxiliar Z. La función objetivo a optimizar, dada la información auxiliar Z se escribe entonces como $E_m[1^T R \mid Z]$ cuya maximización equivale a reducir al mínimo el número de unidades seleccionadas.

Las restricciones se derivan de la aplicación de una función de pérdida a los estimadores de la encuesta. Vamos a concentrarnos en las dos funciones de pérdidas más utilizadas en la práctica, la pérdida absoluta $L = L^{(1)}(a,b) = \mid a - b \mid$ y la pérdida cuadrática $L = L^{(2)}(a,b) = (a-b)^2$.

Para estas funciones de pérdida, cada restricción siempre se puede escribir como una cota en una forma cuadrática, denotada por $E_m[R^T \Delta R \mid Z]$. Las formas particulares adecuadas para los problemas estocásticos y combinatorios serán explicadas mas adelante. La matriz indica las pérdidas potenciales a nivel de unidad. Medidas de sesgo y/o MSE parecen naturales en la práctica. Se derivan de la elección de la función de pérdida absoluta o cuadrática, respectivamente. Estas medidas pueden ser heurísticas por naturaleza, como el pseudo-sesgo para las funciones score tradicionales, o se derivan explícitamente en algunos modelos de medición de error apropiado. En particular, los términos no nulos de fuera de la diagonal de Δ permiten la inclusión de términos cruzadosel en la pérdida "global".

La elección de la matriz Δ está relacionada de forma natural a la elección de la función de pérdida L. Por lo tanto, si Δ es diagonal con entradas $\mid \omega_{ks} \varepsilon_k \mid$, entonces estamos eligiendo la pérdida absoluta de manera que $E_m\left[L\left(\hat{Y}^*(R), \hat{Y}^0\right) \mid Z\right]$ también está acotada por η. Este caso está dirigida al sesgo. Del mismo modo, si $\Delta_{kl} = \omega_{ks}\omega_{ls}\varepsilon_k\varepsilon_l$ entonces

estamos eligiendo la pérdida cuadrática de manera que también $\mathsf{E}_m\left[L\left(\hat{Y}^*(R),\hat{Y}^0\right)|Z\right]$ está igualmente acotada. A su vez, está dirigida al error cuadrático medio. En ambos casos, las técnicas basadas en modelos sobre datos del período actual se pueden aplicar en la versión combinatoria, mientras que en la versión estocástica nos vemos obligados a recurrir a la información auxiliar de otros periodos.

Por ejemplo, la puntuación (local) para una variable dada es generalmente concebida como el producto de un componente de "riesgo" y un componente de "influencia". Una medida genérica puede darse utilizando un enfoque basado en modelos. Sea $p_k = P(y_k^0 \neq y_k \mid y_k)$ la probabilidad a posteriori de que el valor verdadero sea diferente al observado. Sea $\tilde{\mu}_k = \mathsf{E}_m(y_k^0 \mid y_k, y_k^0 \neq y_k)$ la esperanza condicionada del valor verdadero, dado que es diferente del observado. Entonces, tenemos

$$\mathsf{E}_m(y_k^0 \mid y_k) = (1-p_k)y_k + p_k\tilde{\mu}_k \qquad \text{y} \qquad \delta_k = y_k - \mathsf{E}_m(y_k^0 \mid y_k) = p_k(y_k - \tilde{\mu}_k)$$

Resulta que $w_k\delta_k$ puede usarse para constuir el score local de la unidad k con respecto a la variable y, que representa el producto del componente riesgo medido por p_k y el componente influencia medido por $w_k\left(y_k - \tilde{\mu}_k\right)$ donde w_k puede ser, por ejemplo el peso de la muestra.

La principal diferencia entre ambas versiones se plantea al considerar su aplicación real. El problema estocástico, complementado con la suposición de ignorar los términos entre unidades, permite la construcción de las funciones score que pueden aplicarse de forma independiente para cada unidad. El supuesto complementario equivale a considerar los términos cruzados más o menos constantes en el tiempo, no jugando ningún papel significativo en la selección. Por el contrario, el problema combinatorio necesita un número suficiente de observaciones disponibles para llevar a cabo la selección en forma conjunta para todas las unidades.

Teniendo en cuenta la posibilidad de múltiples restricciones, establecemos el siguiente problema de optimización general:

$$\begin{aligned} [P_0] \qquad & \max \ \mathsf{E}_m\left[l^T R \mid \mathbf{Z}\right] \\ & \text{s.t.} \ \ \mathsf{E}_m\left[R^T\boldsymbol{\Delta}^{(q)}R \mid \mathbf{Z}\right] \leqslant \eta_q, \quad q = 1,2,...,Q, \\ & R \in \Omega_0 \end{aligned}$$

Donde Ω_0 denota el espacio de los comportamientos admisible para R y q se refiere a las diferentes restricciones. Las diferentes restricciones pueden surgir del hecho de que existan múltiples variables de interés, o de que las restricciones puedan estar dirigidas a diferentes ámbitos poblacionales. En particular, las matrices de pérdida $\Delta^{(1)},...,\Delta^{(Q)}$ pueden derivarse bajo un único modelo multivariante para el conjunto de datos, incluso cuando las restricciones estén marginalmente especificadas para cada variable de interés.

El problema de optimización estocástica

Como se indicó anteriormente, la suposición principal en esta versión del problema P_0 es dejar de lado los términos entre unidades en cada restricción. Entonces pueden rescribirse como $E_m\left[R^T \Delta R \mid Z\right] = E_m\left[R^T \text{diag}(\Delta) \mid Z\right]$ El problema de optimización estocástica se resuelve en Arbués et al. (2012a) utilizando el principio de dualidad y el de intercambiabilidad. La solución resultante de este problema lineal se da en términos de matrices $M^{(q)} = E_m\left[\Delta^{(q)} \mid Z^{co}\right]$. Pero, puesto que este esquema de selección se aplica unidad por unidad al recibir cada cuestionario y no hay información auxiliar más que con respecto a cada unidad k, los condicionamientos formales sobre Z^{co} se pueden reducir eficazmente al condicionamiento de la información auxiliar $Z_k^{co} = \left\{s_k, x_k, y_k\right\}$ de cada unidad. Así que escribimos $M^{(q)} = E_m\left[\Delta \mid Z^{co}\right] = \sum_{k \in s} E_m\left[\Delta_{kk}^{(q)} \mid Z_k^{co}\right] = \sum_{k \in s} M_{kk}^{(q)}$. Por otro lado, con el fin de obtener los multiplicadores de Lagrange óptimos λ_q^* involucrados en el problema dual, se hace necesario disponer de un fichero histórico con datos observados y depurados. De esta forma, podemos llegar a la solución final, que sólo requiere de las entradas diagonales de las matrices $M^{(q)}$:

$$R_k = \begin{cases} 1 & \text{si} \quad \sum_{q=1}^{Q} \lambda_q^* M_{kk}^{(q)} \leq 1, \\ 0 & \text{si} \quad \sum_{q=1}^{Q} \lambda_q^* M_{kk}^{(q)} > 1. \end{cases}$$

Téngase en cuenta que, dado que el esquema está "entrenado" en los datos históricos, la evaluación de $M_{kk}^{(q)}$ dadas las observaciones de la muestra actual produce necesariamente un pseudo-medida, independientemente de la definición de las matrices de pérdida.

Este hecho proporciona una función score para la selección unidad por unidad. En el caso especial de que Q=1, se selecciona la unidad k cuando $M_{kk} > 1/\lambda^*$ de modo que la expresión a la izquierda de la desigualdad puede considerarse como el valor del score y el de la derecha como el valor del umbral. Igualmente, se puede considerar $\lambda^* M_{kk}$ como una puntuación "estandarizada", en el sentido de que el valor umbral se establece en forma genérica igual a 1. Esto se extiende directamente a la configuración con múltiples restricciones, donde cada $\lambda_q^* M_{kk}^{(q)}$ es un score local estandarizado y $\sum_{q=1}^{Q} \lambda_q^* M_{kk}^{(q)}$ es un score global estandarizado, con el valor del umbral igual a 1.

El score global se deriva de la estructura lineal del problema dual y se permiten pocas variaciones sin una modificación sustancial del problema P_0. Como excepción, si el score global es inicialmente concebido como la suma ponderada de los score locales, se puede incorporar cada peso en la restricción que genera el correspondiente score local estandarizado.

El problema estocástico clarifica, pues, el hecho de que el rendimiento de selección unidad-por-unidad sólo se puede establecer a través de hipotéticas repeticiones del proceso de selección. Al final de cada proceso de selección, tenemos la estrategia de

selección realizada r y la pérdida realizada $\sum_{k=1}^{n} r_k M_{kk}^{(q)}$, que puede ser mayor o menor que el límite especificado η_q para $q = 1,...,Q$. En cualquier repetición hipotética del proceso de selección y_k y y_k^0 variarán, y también lo harán los correspondientes $M_{kk}^{(q)}$ y r_k. Es sobre tales hipotéticas repeticiones que la restricción $E_m\left[R\Delta^{(q)}R \mid Z\right] \leqslant \eta_q$ posiblemente pueda ser satisfecha, pero no para cada realización particular del procedimiento de selección.

El problema de optimización combinatoria

El problema combinatorio trata la selección entre todas (o un grupo de) las unidades. Se permiten ahora los términos entre unidades, y la información auxiliar utilizada es la de la muestra , las covariables auxiliares X, y las variables de interés Y, es decir, por $Z = Z^{co}$. Téngase en cuenta que toda esta información está disponible sólo después de que se han recogido todos los cuestionarios, por lo que sólo se puede aplicar al final de este proceso. Se demuestra que cada restricción se reduce a $E_m\left[R^T\Delta^{(q)}R \mid Z^{co}\right] = r^T M^{(q)} r$, donde $M^{(q)} = E_m\left[\Delta^{(q)} \mid Z^{co}\right]$, que puede ahora ser posiblemente evaluado bajo algún modelo de medición del error. Por consiguiente, se hace posible establecer directamente el rendimiento de la estrategia de selección, considerada como una realización articular. El problema de optimización puede ser reformulado como:

$$
\begin{aligned}
[P_{co}(M, \eta, \Omega_0)] \qquad & \max 1^T r \\
& \text{s.a} \quad r^T M^{(q)} r \leqslant \eta_q, \quad q = 1,2,...,Q, \\
& r \in \Omega_0
\end{aligned}
$$

Obsérvese que una derivación más directa se puede obtener por no haber promovido el vector de estrategia de selección r a un vector R aleatorio, en el momento de modelar los errores de medición.

Este problema combinatorio se resuelve en dos formas diferentes utilizando dos algoritmos voraces, véase Salgado et al (2012). La solución de los dos algoritmos no es exacta a priori, sino subóptima con un buen grado de aproximación. El algoritmo más rápido es notablemente menos preciso que el más lento. Esta falta de precisión implica una pequeña cantidad de sobredepuración en la práctica, es decir, que se seleccionarán más unidades de las que hubiera sido necesario depurar de forma óptima. Estos algoritmos heurísticos investigan localmente el óptimo en cada iteración hasta que la solución satisface todas las restricciones. Para ello se introducen las funciones inviabilidad $h_i(r)$ para cada algoritmo $i = 1,2$ (ver Salgado et al. (2012) para más detalles) que indica si una solución satisface todas las restricciones $h(r) = 0$ o no $h(r) > 0$ Ambos algoritmos comienzan a partir de la solución inicial $r = 1$ y en cada iteración seleccionan de manera localmente óptima la siguiente unidad, hasta que se cumplan todas las restricciones.

Finalmente, se pueden considerar que las dos versiones en relación con dos enfoques diferentes para el problema de optimización bajo incertidumbre, véase, por ejemplo

Wets (2002). La versión combinatoria es consistente con el de "esperar y ver", ya que se posponen todas las decisiones hasta que toda la información está disponible. La versión estocástica es un enfoque "aquí y ahora" ya que la decisión sobre el procedimiento o regla de selección se realiza antes de la recogida de datos.

3.3.4.- Caso práctico

En este apartado, se compara empíricamente la eficiencia de nuestra propuesta con funciones score ampliamente utilizadas. Para ello se emplean datos reales del Índice de Cifras de Negocio (ICN) y del Índice de Entrada de Pedidos (IEP). Como es habitual en los experimentos de depuración selectiva, se trabaja con un fichero de datos doble, que para cada unidad contiene datos observados y datos depurados. Cuando se selecciona una unidad, sus valores observados son sustituidos por sus correspondientes contrapartes depurados, considerados verdaderos. Vamos a denotar por $\hat{Y}^{sel}(n_{ed})$ el estimador que se obtiene cuando n_{ed} cuestionarios han sido seleccionados de acuerdo con una técnica de depuración selectiva y se han depurado correspondientemente. Téngase en cuenta que $\hat{Y}^{sel}(n_{ed}=n)=\hat{Y}^{0}$. Como criterio para valorar la selección de unidades, se utiliza el valor absoluto del pseudo-sesgo de un estimador $\hat{Y}$, dado por $\widetilde{ARB}\left(\hat{Y}^{sel}(n_{ed})\right)=\left|\frac{\hat{Y}^{sel}(n_{ed})-\hat{Y}^{0}}{\hat{Y}^{0}}\right|$.

Se va a comparar $\widetilde{ARB}\left(\hat{Y}^{sel}(n_{ed})\right)$ para una técnica de depuración selectiva sel con $\widetilde{ARB}_0(n_{ed})\equiv\widetilde{ARB}\left(E\left[\hat{Y}^{ran}(n_{ed})\right]\right)$ donde ran significa una selección aleatoria con igual probabilidad, de la que se toma su esperanza. Es inmediato demostrar que $\widetilde{ARB}_0(n_{ed})=\left(1-\frac{n_{ed}}{n}\right)\widetilde{ARB}\left(\hat{Y}^{sel}(0)\right)$. Llamemos $\gamma_0(n_{ed})$ y $\gamma^{sel}(n_{ed})$ las líneas rectas y poligonales con vértices $\gamma_0(n_{ed})\approx\{(0,\widetilde{ARB}_0(0)),(n_{ed},\widetilde{ARB}_0(n_{ed}))\}$ y $\gamma^{sel}(n_{ed})\approx\{(0,\widetilde{ARB}_0\left(Y^{sel}(0)\right)),(1,\widetilde{ARB}_0\left(Y^{sel}(1)\right)),\ldots,(n_{ed},\widetilde{ARB}_0\left(Y^{sel}(n_{ed})\right))\}$, respectivamente. Denotamos por $A_{\gamma}(n_{ed})$ el área de la superficie entre la curva y el eje horizontal a la izquierda de la línea vertical en n_{ed} (ver figura 1). El área se considera positiva si la línea poligonal se encuentra por debajo de la línea recta y negativa en caso contrario. Proponemos la siguiente definición para la eficiencia de la técnica sel:

$$\in^{sel}(n_{ed})\equiv\left(A_{\gamma_0}(n_{ed})-A_{\gamma^{sel}}(n_{ed})\right)/A_{\gamma_0}(n_{ed})=1-\frac{A_{\gamma^{sel}}(n_{ed})}{A_{\gamma_0}(n_{ed})}.$$

Téngase en cuenta que esta medida depende del número de unidades para seleccionar. Esto nos permite reconocer aquellas técnicas que dan prioridad a las unidades más influyentes primero. Una situación típica se representa en la figura 1.

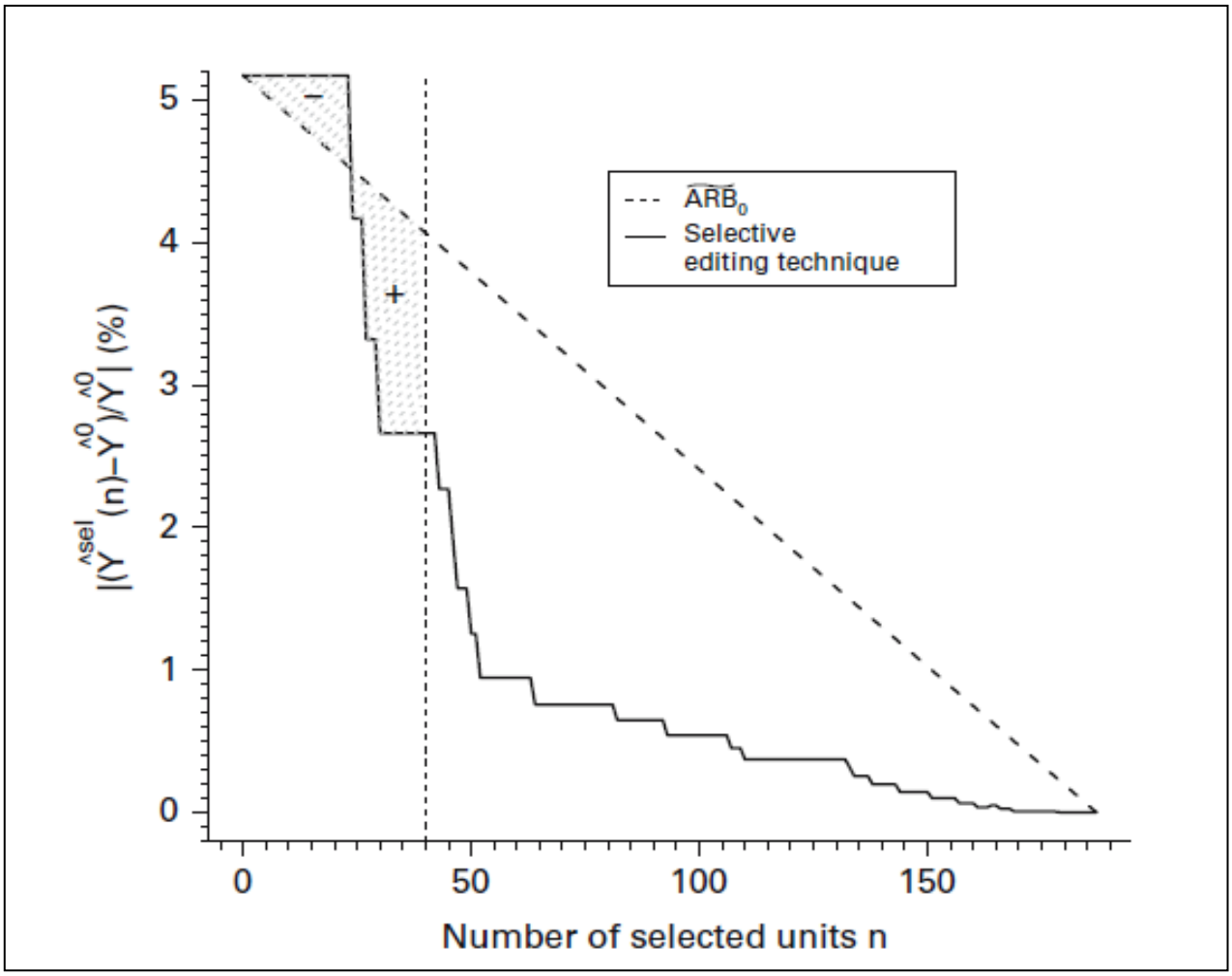

Figura 1. Seudo-sesgo relativo vs. número de unidades seleccionadas

Hemos llevado a cabo una comparación de nuestra propuesta con la obtenida a partir de algunas funciones score en la literatura. Con el fin de evitar posibles interferencias con los datos que faltan y las unidades recién añadidas, hemos utilizado un subconjunto rectangular de los datos de la muestra del ICN y del IEP. Para mayor claridad nos concentraremos en una función score particular, ilustraremos los resultados correspondientes y haremos algunos comentarios sobre el comportamiento similar de todas ellas. Hemos utilizado una versión ligeramente mejorada de la función RATIO de Latouche y Berthelot (1992).

Sea $r_k^{(t)} = \dfrac{y_k^{(t)}}{y_k^{(t-1)}}$ y se define

$$\bar{r}_k^{(t)} = \begin{cases} \left| \dfrac{r_k^{(t)}}{\text{mediana}_k(r_k^{(t)})} - 1 \right| & \text{si } r_k^{(t)} > \text{mediana}_k(r_k^{(t)}), \\ \left| 1 - \dfrac{r_k^{(t)}}{\text{mediana}_k(r_k^{(t)})} \right| & \text{en otro caso.} \end{cases}$$

Se define también $g_k^{(t)} = \omega_{ks} \times \bar{r}_k^{(t)} \times \sqrt{\max(y_k^{(t)}, y_k^{(t-1)})}$ y el score local $s_k^{(t)} = \dfrac{\left| g_k^{(t)} - \text{mediana}_k(g_k^{(t)}) \right|}{\text{RI}_k(g_k^{(t)})}$

Donde RI representa el rango intercuartílico. Para $q = 1,\ldots,Q$ variables, éstas se combinan en la función score global definida como $\text{ratio2}(k,t) = S_k^{(t)} = \sum_{q=1}^{Q} s_k^{(q,t)}$ La mejora surge debido al hecho de la utilización de los datos sólo desde el período de tiempo t-1 y no desde t-2 como en la propuesta original. Por lo tanto esta función Ratio2 sólo se puede utilizar como una forma de depuración de salida después de que se han recogido todos los datos (como en el enfoque combinatorio, con el que haremos la comparación).

En cuanto a la priorización de unidades calculadas bajo el enfoque combinatorio, en primer lugar hay que especificar el modelo auxiliar m^* para encontrar las predicciones $\hat{y}_k$. Para cada unidad hemos estimado tres modelos alternativos de series temporales $\xi_1 : (1-B)z_t = a_t$, $\xi_2 : (1-B^{12})z_t = a_t$ y $\xi_3 : (1-B)(1-B^{12})z_t = a_t$, donde B es el operador retardos, $z_t = \log(m + y_t^0)$ (m siendo un parámetro estimado por máxima verosimilitud) y a_t ruido blanco. Cada predicción $\hat{y}_k$ se calcula de acuerdo con el mejor modelo correspondiente ξ^* (en términos del error cuadrático medio mínimo estimado). Dado que la muestra es un panel fijo elegido por cut-off, las ponderaciones de muestreo ω_{ks} son todas iguales a 1.

A continuación, hemos aplicado el modelo de observación-predicción univariante genérico de las transformaciones logarítmicas de la cifra de negocios y las entrada de pedidos de forma independiente. La probabilidad de error $p_k = p$ y varianza de observación $\sigma_k^2 = \sigma^2$ se han estimado a partir de los últimos tres meses utilizando el fichero doble de datos. La varianza de predicción v_k^2 se ha calculado de acuerdo con el modelo elegido correspondiente ξ^* para cada unidad. Como matrices de pérdida hemos elegido tanto la cuadrática como la absoluta. Por último, para hacer la comparación con una selección aleatoria de unidades, hemos calculado el pseudo-sesgo relativo en valor absoluto para 50 selecciones aleatorias equiprobables. Hemos calculado la media y los cuartiles primero y tercero de la distribución correspondiente. Esto proporciona un intervalo de confianza para cada número de unidades seleccionadas (véase la figura 2). La motivación es proporcionar una visión no sólo respecto de la selección aleatoria promedio, sino también de su distribución.

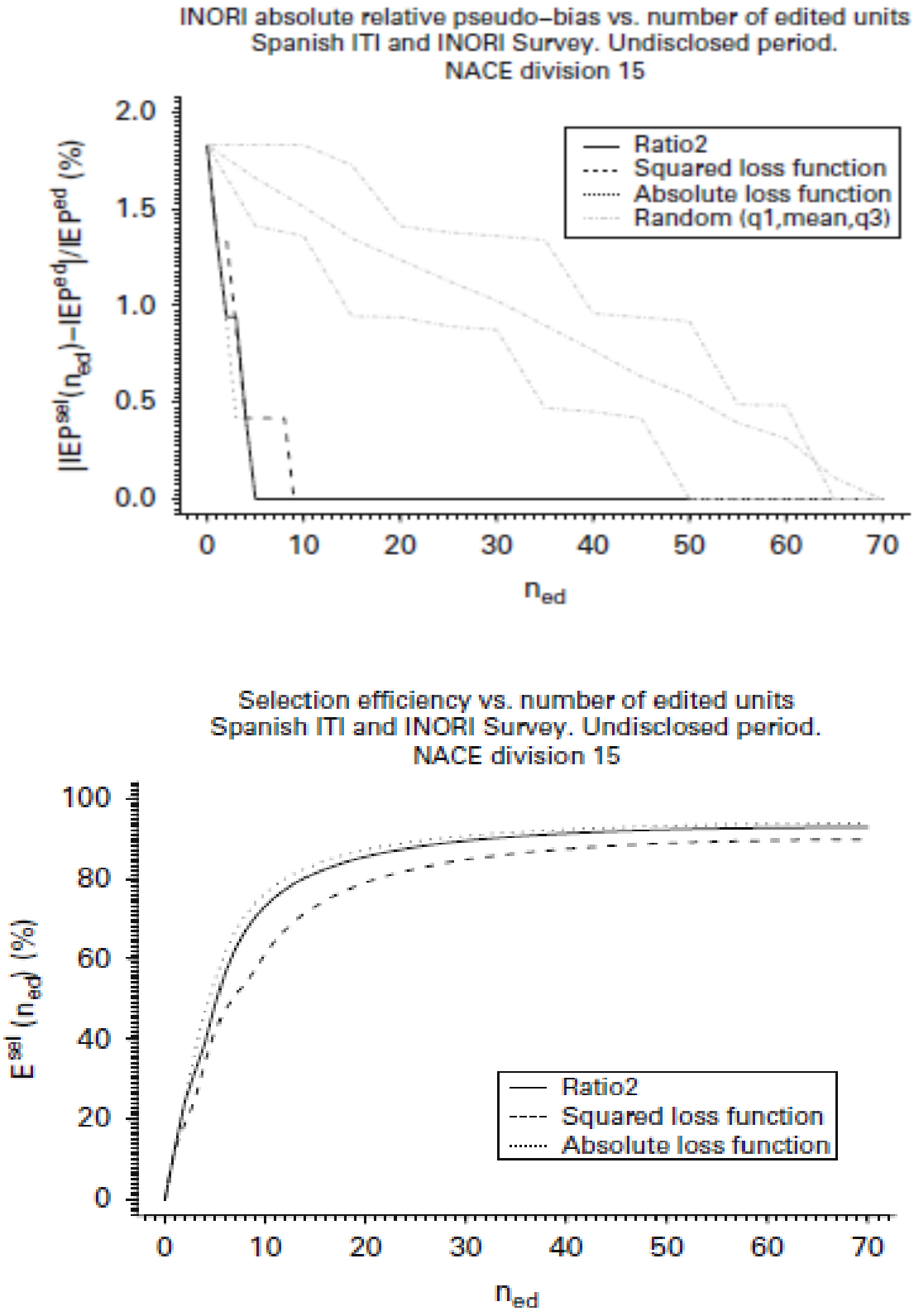

Figura 2. Seudo Sesgo relativo y eficiencia vs. número de unidades seleccionadas ICN (ITI) e IEP (INORI)

Hemos llevado a cabo esta comparación para 23 ramas de actividad, que corresponden a divisiones y subdivisiones de la NACE Rev. 2. En primer lugar, Ratio2 mostró un mejor rendimiento que el resto de las funciones score (RATIO, DIFF, FLAG; ver Latouche y Berthelot 1992) En 15 casos el enfoque de optimización con función de pérdida absoluta daba la priorización más eficiente. En 9 de esos casos Ratio2 resultaba más eficiente que la función de perdida cuadrática (la figura 2 ilustra este comportamiento). En los 5 casos que la función de pérdida cuadrática supera a las otras dos opciones, la pérdida absoluta también hizo mejor que Ratio2. En los restantes 3 casos, Ratio2 superaba ligeramente a la pérdida absoluta, que a su vez se comportó mejor que la pérdida cuadrática.

Por lo tanto, en general, la pérdida absoluta es más eficiente que la pérdida cuadrática en términos del seudo-sesgo, como se esperaba. Esto también sucede con Ratio2, ya que está dirigida también al sesgo. En general, la pérdida absoluta es también más eficiente que las funciones score.

Sin embargo, en condiciones reales de producción, tanto los datos que faltan como los encuestados que se agregan a la muestra deben ser tenidos en cuenta. En estos casos, en el enfoque de optimización, los valores de predicción $\hat{y}_k$ deben ser imputados o fijados bajo algún esquema complementario, ya que los modelos de series de tiempo no obtienen esos valores. Como un test primario, se asignó $\hat{y}_k = y_k$ en estos casos. El resultado general fue un ligero deterioro de la actuación de las funciones score para todos los valores de n_{ed} mientras que en el enfoque de la optimización el comportamiento fue tan bueno como antes para las unidades más influyentes ($n_{ed} = 1,2,\cdots$), pero notablemente más pobre de las últimas unidades ($n_{ed} \geqslant n/2$). No hemos considerado estos temas en la comparación anterior, ya que pertenecen a sofisticaciones del modelo de observación- predicción.

Es importante señalar que los resultados anteriores se han obtenido con modelos de series temporales muy elementales y supuestos muy simplificadores, y no incorporan ningún conocimiento a priori. Así, hay espacio para sofisticar los modelos y mejorar el rendimiento de los métodos de optimización.

3.3.5 Conclusiones

De acuerdo a las líneas de investigación establecidas de utilización de modelos y optimización, se ha intentado avanzar en el grado de formalización de nuestras propuestas. Se han formulado distintos modelos para las variables de interés, como los valores observados, los verdaderos y los depurados así como para los errores de observación. Hemos introducido el modelo que denominamos de observación-predicción, necesario para resolver las matrices de las funciones de pérdida de los procedimientos de optimización. Hemos generalizado el procedimiento de optimización estocástica, que habíamos obtenido anteriormente, llegando a un marco de optimización general, del que derivan las versiones estocástica y combinatoria

Respecto a la limitación que hacemos a estimadores lineales, lo que contrasta con el uso común en la práctica de algunos estimadores no lineales como por ejemplo estimadores de razón o de regresión, puede ser fácilmente superada como sigue. En la práctica la

mayoría de los estimadores no lineales $\hat{Y}_{U_d}^{nl}$ son funciones de estimadores lineales $\hat{Y}_{U_d}^{nl} = f(\hat{Y}_{U_d}^{(1)},\ldots,\hat{Y}_{U_d}^{(M)})$. Entonces, en lugar de considerar la restricción correspondiente para el MSE de $\hat{Y}_{U_d}^{nl}$, consideramos una restricción para cada estimador lineal $\hat{Y}_{U_d}^{(m)}$, $m = 1,\ldots,M$.

Se ha comparado empíricamente nuestra propuesta con las funciones score utilizadas habitualmente, resultando, en general, más eficientes. Se observa que queda camino para refinar las herramientas (usando mejores parámetros, construyendo modelos multivariantes, etc.). Conviene seguir investigando modelos de trabajo que se puedan construir para las variables discretas o semicontinuas, allanando el camino para el uso de técnicas de depuración selectiva también en las encuestas de hogares.

Capítulo 4

Conclusiones

En esta tesis se han llevado a cabo un conjunto de investigaciones relacionadas con la depuración e imputación. Las principales líneas de trabajo han sido el uso de modelos estadísticos y de técnicas de optimización. Las investigaciones desarrolladas se agrupan en tres bloques, que reflejan la forma en que se han ido realizando en el tiempo. Se ha partido de investigaciones que intentaban resolver problemas concretos para posteriormente ir avanzando en grado de generalización y formalización. Así, en el primer bloque, "Depuración e imputación basada en modelos de series temporales", se han abordado preocupaciones que están presentes a lo largo de la tesis, como la utilización de modelos y la optimización de recursos mediante la depuración selectiva, pero restringidas al ámbito de las encuestas continuas y al uso de modelos de series temporales. En el segundo bloque, "La depuración selectiva como un problema de optimización estocástica", se ha intentado dar una solución formal al problema de la depuración selectiva, que hasta ahora había sido tratada de forma heurística en la literatura. Finalmente, en el tercer bloque, "Desarrollo de un marco teórico general de depuración e imputación basado en modelos y optimización", se han propuesto un conjunto de métodos que sirvan para todo tipo de encuestas y de operaciones estadísticas.

En relación con la primera de las investigaciones, "Depuración e imputación basada en modelos de series temporales", se ha explorado la utilización de modelos de series temporales, en particular modelos RegARIMA, tanto para los distintos enfoques de depuración (microdepuración, macrodepuración y depuración selectiva) como para la imputación. Aunque los modelos de series temporales no son de uso común en la depuración ni en otras fases de la cadena de producción de datos estadísticos (con excepción del ajuste estacional), se parte de la idea de que los sucesivos datos de una encuesta continua pueden ser considerados como una realización particular de un proceso estocástico y la utilización de modelos de series temporales tiene plena justificación teórica. En este trabajo se han utilizado modelos RegARIMA, tanto para los microdatos como para los macrodatos. En el enfoque de microdepuración, los modelos se han utilizado para especificar los edits, mediante la construcción de intervalos de confianza para los datos verdaderos, a partir de las funciones de predicción. Los valores observados definidos como atípicos (y por tanto sospechosos de error) son aquellos que se alejan de los valores esperados según los modelos. En el enfoque de macrodepuración, se ha procedido de forma similar, utilizando los modelos construidos para los macrodatos. Para la depuración selectiva se han desarrollado un conjunto de herramientas (que hemos denominado sorpresas, sorpresas estándar, sorpresas estándar ponderadas, e influencias) basadas en la función de predicción un periodo por delante del modelo. En algunos casos estas herramientas se han utilizado siguiendo una estrategia top-down. En otros, constituyen una función score, como las usadas ampliamente en la literatura. Una aportación de esta tesis es que la función score propuesta se obtiene a partir de los modelos y no por razones de tipo heurístico. Adicionalmente, las funciones score utilizan un valor "anticipado" para compararlo con el valor observado, y se suele emplear para ello el dato del periodo anterior. En la función score basada en los modelos hemos utilizado la predicción un periodo por

delante, mejorando el valor anticipado. De hecho, únicamente coincidirían si la serie siguiera un proceso camino aleatorio, lo que es poco habitual en las series de flujos de variables reales.

A partir de los modelos, hemos obtenido las imputaciones a través de la predicción un periodo por delante, lo que resulta razonable intuitivamente, ya que coinciden con la esperanza matemática condicionada a la información disponible. En caso de que no hubiera respuesta para dos o más periodos hemos utilizado la predicción dos periodos por delante, tres periodos por delante, etc. Las propiedades de las imputaciones obtenidas por este método se derivan directamente de las propiedades de la predicción en este tipo de modelos. De este modo obtenemos unas imputaciones que son insesgadas y óptimas en el sentido de que minimizan el error cuadrático medio.

La utilización de los modelos ha permitido hacer un uso intensivo de la información existente de la encuesta en periodos anteriores. El uso de esta información no es nuevo en la depuración e imputación de datos. Los edits de la razón, las tasas de variación y la imputación histórica se utilizan frecuentemente. Sin embargo, estos métodos se basan en un uso parcial de la información de las encuestas anteriores, en tanto que las herramientas de depuración que hemos construido a partir de los modelos utilizan de manera eficiente todo el pasado de la serie, aprovechando toda la estructura de correlaciones. Otra ventaja del uso de modelos es que permite hacer inferencias probabilísticas y construir, por ejemplo, intervalos de confianza. Con ello hemos conseguido especificar y establecer los valores extremos de los edits, lo que habitualmente se hacía a partir de la experiencia del experto de la encuesta o por otras razones de tipo empiricista. El hecho de utilizar distintos modelos para cada uno de los estratos o unidades de observación permite tener en cuenta el diferente comportamiento y variabilidad de los sectores económicos, productos, empresas, etc.

Usando la metodología propuesta basada en modelos de series temporales, se han construido herramientas de depuración que pudieron detectar los valores sospechosos más influyentes. Por lo tanto, los errores más importantes se pueden chequear y corregir de forma rápida, pudiéndose lograr mejoras de puntualidad sin pérdida de acuracidad. De acuerdo a estos resultados, el uso de modelos de series temporales puede ser muy útil para ahorrar tiempo en la depuración de los indicadores a corto plazo.

A partir de los modelos anteriormente mencionados, se ha obtenido información de utilidad para llevar a cabo la macrodepuración, siguiendo la línea del análisis exploratorio de datos (Tukey, 1977). Dado que, en condiciones bastante generales, una variable que se determina dentro de un modelo econométrico simultáneo dinámico estructural (SEM) se genera de una manera univariante por un modelo ARIMA con Análisis de Intervención, a partir de los modelos se ha obtenido información del comportamiento dinámico de los agregados de una encuesta, entre ellos comportamiento tendencial, estacionalidad, efectos de calendario y otros efectos determinísticos, distintos tipos de valores atípicos y volatilidad. Esta metodología se ha aplicado a los agregados del IPI tanto por ramas de actividad como por comunidades autónomas. Para captar el efecto de los días laborables, en vez de las siete variables de *trading day* habitualmente utilizadas (Hillmer, Bell y Tiao, 1983), se ha construido una única variable, denominada días laborables ponderados, consiguiéndose un modelo más parsimonioso y de mejor ajuste.

En el segundo bloque de investigaciones, “La depuración selectiva como un problema de optimización estocástica” hemos descrito un marco teórico que permite dar una solución al problema de la depuración selectiva. El objetivo ha sido encontrar una adecuada estrategia de selección, que indique para cada unidad informante si el cuestionario debe ser depurado o no, utilizando la información disponible. Para ello se ha definido formalmente el concepto de estrategia de selección, como un vector aleatorio medible respecto a la sigma-álgebra generada por toda la información disponible hasta el periodo de observación. La búsqueda de una adecuada estrategia de selección la presentamos como un problema de optimización lineal con restricciones cuadráticas, cuya solución es la selección de unidades a depurar. El objetivo es minimizar la carga de trabajo esperada, con la restricción de que el error esperado del estimador calculado con los datos depurados se sitúe por debajo de una cierta constante. En dicho objetivo, se traduce carga de trabajo como número de cuestionarios a depurar. Las funciones de pérdida se restringen aquí al error cuadrático, quedando el estudio de otras funciones para más adelante en la tesis.

Adicionalmente, hemos considerado también una versión simplificada con restricciones lineales. En este caso, la solución viene dada por una cierta función score. Puesto que no hay justificación teórica para despreciar la parte cuadrática, el uso de la solución lineal debe justificarse por su analogía con las funciones score o, de forma empírica, mediante su eficacia en la práctica.

La principal aportación de nuestro enfoque de optimización estocástica es el desarrollo de un marco teórico para solucionar el problema de la depuración selectiva. Hasta ahora, la función score y los otros procedimientos se basaban en métodos ad hoc. Utilizando datos reales, se ha podido constatar que la función score correspondiente a la versión lineal de optimización mejora la de uso común desarrollada en la literatura. Ambas versiones de optimización producen resultados en los que se satisfacen aproximadamente las restricciones, salvo para valores muy pequeños de las cotas. El método cuadrático parece más conservador y las cotas se satisfacen mejor, pero resultan más unidades a depurar. Por otra parte, la implementación del método lineal es más fácil y computacionalmente menos costoso.

En el tercero de los bloques de esta tesis “Desarrollo de un marco teórico de depuración e imputación basado en modelos y optimización” hemos intentado generalizar las dos líneas de investigación descritas anteriormente. Los sistemas generalizados han supuesto un gran avance en la depuración e imputación. Sin embargo, no resuelven la especificación de los edits, que se establecen habitualmente de acuerdo a la experiencia práctica, sin que exista un marco teórico adecuado. En este trabajo se ha abordado la forma en que pueden obtenerse los edits a partir de modelos construidos con toda la información disponible. Adicionalmente, se han utilizado los modelos para calcular las matrices de pérdida presentes en las restricciones del problema de optimización. Hemos asimilado el problema de depuración selectiva a un problema genérico de optimización. A partir de los principios generales de minimización de los recursos garantizando la calidad de los datos, hemos propuesto una traslación matemática de esos principios en un problema de optimización general. Su solución es la selección de unidades a depurar. Se considera que los recursos son equivalentes al número de cuestionarios a depurar y la calidad a la acuracidad de los estimadores. La propuesta consiste en minimizar el número de cuestionarios a depurar, siempre que las funciones de pérdida de los estimadores estén acotadas por un determinado valor.

Ampliando el enfoque introducido anteriormente, hemos proporcionado dos versiones de optimización, correspondientes a los dos escenarios habituales para la implementación de la depuración selectiva. En el primer caso, la selección se lleva a cabo unidad por unidad, de forma que la selección de una unidad no depende de la selección de las otras unidades. Este procedimiento es adecuado cuando la selección se lleva a cabo en el momento que se recoge cada cuestionario, sin tener que esperar la llegada de otros cuestionarios. Llamamos a este enfoque depuración selectiva mediante optimización estocástica, porque la solución al problema de optimización puede sólo ser establecida con respecto a hipotéticas repeticiones del proceso de selección. En el segundo enfoque, llevamos a cabo la selección conjuntamente para todas (o al menos un grupo de) las unidades. Este procedimiento es adecuado cuando la selección se lleva a cabo en una etapa avanzada del proceso de recogida y se dispone de todos o al menos un número suficientemente grande de cuestionarios. Llamamos a este enfoque depuración selectiva mediante optimización combinatoria, porque el resultado de la solución puede establecerse condicionado a las observaciones de la muestra bajo algún error de medida. Respecto a la optimización estocástica descrita anteriormente en esta tesis, hemos dado un paso más hacia la generalización, al establecer un problema general de optimización, del que se derivan los enfoques de optimización estocástica y optimización combinatoria. Para la resolución de los problemas de optimización se ha construido un modelo estadístico multivariante para los errores de medida, asistido por un modelo auxiliar para hacer predicciones. A este modelo le hemos denominado el modelo de observación-predicción. Se ha evaluado el enfoque de optimización, en comparación con las funciones score, utilizando datos reales de los Índices de Cifras de negocios e Índices de Entrada de Pedidos, obteniéndose, por lo general, mejores resultados.

El enfoque de la función score es hoy en día la técnica más utilizada de depuración selectiva. Por tanto, proporciona un marco adecuado para evaluar las ventajas y desventajas del enfoque aquí propuesto. Anteriormente se ha señalado su carácter empiricista en contraste con la formalización de nuestra propuesta, y se han efectuado comparaciones empíricas entre los dos procedimientos. Otro aspecto a evaluar es el número de decisiones a efectuar. El enfoque de la función score comprende cuatro decisiones (Lawrence y McKenzie 2000): (i) modelo de depuración para la construcción de los valores “anticipados”, (ii) cada una de las funciones score locales, (iii) una función score global, y (iv) un umbral de corte. En el enfoque aquí propuesto, las tres primeras decisiones son sustituidas e integradas conjuntamente en un único paso, la construcción del modelo de observación-predicción, y la última, en la formulación de las restricciones del problema de optimización. La construcción del modelo de observación-predicción es un ejercicio multivariante, por lo que la integración de la selección de las funciones score local y global se hace de manera natural, junto con la construcción del modelo estadístico. Por otra parte, la elección del umbral de corte es ahora sustituido por la elección de los límites en el problema de optimización. En el enfoque de la función score, este umbral debe ser elegido normalmente utilizando datos de las realizaciones anteriores de la encuesta, usando una conexión heurística entre este valor y la función de pérdida de los estimadores de la encuesta. En el nuevo enfoque la elección de los límites hace uso de los valores a priori de las varianzas (o de alguna otra medida similar) como se hace en la etapa de diseño de la encuesta, encajando de forma natural en el proceso de producción de global de la encuesta. Por otra parte, en el enfoque de optimización, ambas versiones se ajustan naturalmente como el comienzo y final de un proceso de depuración en tiempo continuo, en el que las unidades se van

depurando a medida que van llegando en el flujo de recogida. La versión estocástica corresponde a la explotación de la información longitudinal y se sitúa al comienzo del proceso, mientras que la versión combinatoria surge como una técnica de depuración centrada en información de sección transversal y por tanto se sitúa al final del proceso. Por el contrario, el enfoque de la función score y las técnicas tradicionales de macrodepuración apenas pueden verse con los mismos criterios metodológicos. Queda abierto para el trabajo futuro encontrar una formulación más general de este proceso de depuración continuo que incorpore las dos versiones de optimización.

Un cierto paralelismo puede encontrarse en la metodología de Fellegi-Holt, (en particular, con los diferentes enfoques del problema de localización del error), que también hace un amplio uso de técnicas de optimización. La metodología Fellegi-Holt se centra en cada cuestionario individual, con el objetivo de minimizar el número de campos a imputar satisfaciendo todos los edits. La propuesta de optimización se centra en el conjunto de cuestionarios de la muestra, buscando minimizar el número de unidades a depurar, satisfaciendo restricciones sobre funciones de pérdida y utilizando un modelo estadístico en lugar de edits establecidos a priori. Actualmente los puntos comunes se reducen al hecho de que la optimización matemática aparece como una traducción natural de los principios de depuración propuestos. Sin embargo, ante el uso de modelos para especificar los edits (la única fase de la microdepuración que la metodología de Fellegi-Holt no resuelve) se abre la perspectiva de un proceso de depuración coherente, formado por la metodología de Fellegi-Holt auxiliada por el uso de modelos para encarar la microdepuración, la optimización para abordar la depuración selectiva, y de nuevo los modelos para la macrodepuración.

Finalmente, como perspectivas de futuro inmediato, se necesita más investigación metodológica para encontrar modelos multivariantes genéricos que permitan especificar los edits y ajustar el modelo de observación-predicción, y generalizarlos a variables cualitativas y semicontinuas. Desde el punto de vista práctico, todas las aplicaciones se han llevado a cabo con encuestas tradicionales. Sería conveniente investigar los procedimientos con datos administrativos y *Big data*.

REFERENCIAS

[1] Anderson, K. (1989b), *Enhancing Clerical Cost, Effectiveness in the Average Weekly Earnigs*, Unpublished report, Belconnen: Australian Bureau of Statistics.

[2] Andrews, D.F., P.J. Bickel, F.R. Hampel, P.J. Huber, W.H. Rogers, and J.W. Tukey (1972), *Robust Estimates of Location: Survey and Advances*, Princeton, NJ: Princeton University Press.

[3] Arbués, I., Revilla, P., and Salgado, D. (2013). *An optimization approach to selective editing*. Journal of Official Statistics.Vol, 29 No. 4, pp. 1-23.

[4] Arbués I, González M., Revilla P. , *A Class of Stochastic Optimization Problems with Application to Selective Data Editing*, documento de trabajo, Instituto Nacional de Estadística.

[5] Arbués, I., González, M., and Revilla, P. (2012a). *A class of stochastic optimization problems with application to selective data editing.* Optimization, 61, 265–286.

[6] Arbués, I., Revilla, P., and Salgado, D. (2012b). *Optimization as a theoretical framework to selective editing*. UNECE Work Session on Statistical Data Editing, WP1, 1–10.

[7] Banim, J. (2000), *An assessment of macro-editing methods*. ECE, Work Session on Statistical Data Editin, Cardiff, working Paper No.7.

[8] Bankier, M., Luc, M., Nadeau, C., and Newcombe, P. (1995). *Additional Details on Imputing Numeric and Qualitative Variables Simultaneously*. Proceeding of the Section on survey Research Methods, American Statistical Association.

[9] Bankier, M., Luc, M., Nadeau, C., and Newcombe, P. (1996). *Imputing Numeric and Qualitative Variables Simultaneously*. Statistics Canada Technical Report, 120 pages

[10] Barcaroli, G. y Venturi, M. (1997): *DAISY (Design Analysis and Imputation System):Structure, Methodology, and First Applications.* Statistical Data Editing, Volume II, United Nations Statistical Commission Statistical Standards and Studies.No 48

[11] Barnett, V., and T. Lewis (1984), *Outliers in Statistical Data,* 2nd ed., New York: Wiley.

[12] Bazaraa M. S., Sherali H.D.y Shetty C.M. (1993), *Nonlinear Programming: Theory and Algorithms*, Nueva York: Wiley.

[13] Beckman, R. J., and R. D. Cook (1983), *Outliers Technometrics,* 25, pp. 119-149.

[14] Bell, W.R. (1999). *An overview of REGARIMA modelling.* Forthcoming Research Report. Statistical Research Division. U.S. Census Bureau.

[15] Belsley, D.A., E. Kuh, and R. E. Welsch. (1980), *Regression Diagnostics*, New York: Wiley.

[16] Bershad, M. A. (1960), *Some Observations on Outliers*, unpublished report, Washington, DC: U.S. Bureau of the Census.

[17] Berthelot J.y Latouche M. (1993), *Improving the efficiency of Data Collection: A Generic Respondent Follow-Up Strategy for Economic Surveys*, Journal of Business and Economic Statistics, 11, 417–424.

[18] Bethlehem, J (2009). *Applied Survey Methods: A Statistical Perspective*. Journal of Official Statistics 20, 233–264., Copyright © 2009 John Wiley & Sons

[19] Biemer, P.P y Fecso, R.S.(1995). *Evaluating and Controlling Measurement Errors in Business Surveys*. Business Survey Methods, John Wiley & Sons.

[20] Boucher, L. (1991), *Micro-Editing for the Annual Survey of Manufactures. What is the Value-Added?*, Proceedings of the Annual Research Conference, Washington, DC: U.S. Bureau of the Census, pp. 765-781.

[21] Box, G.E.P. and Tiao, G.C. (1975), *Intervention Analysis with Applications to Economic and Environmental Problems*, Journal of the American Statistical Association, 349.

[22] Box, G.E.P. and Jenkins, G.M. (1970). *Time Series Analysis, Forecasting and Control*, ed. Holden-Day, San Francisco.

[23] Bruce, A. G. (1991), *Robust Estimation and Diagnostics for Repeated Sample Surveys*, Mathematical Statistics Working Paper 1991/1, Wellington, Statistics New Zealand.

[24] Carling, J (1999): *Quality and Statistics. Quality Work and Quality Assurance within Statistics*. Office for Official Publications of the European Communities.

[25] Chambers, R. L. (1986), *Outlier Robust Finite Population Estimation*, Journal of the American Statistical Association, 81, pp 1063-1069.

[26] Chambers R. (2000) *Evaluation Criteria for Statistical Editing and Imputation*, T001.05, EUREDIT report.

[27] Chang, I., Tiao, G. C. and Chen, C. (1988), *Estimation of Time Series Parameters in the Presence of Outliers*, Technometrics, 30, pp.193-204.

[28] Chinnppa, N., R. Collins, J. F. Gosselin, T. S. Murray, and S. Simard (1990), *Macro-Editing al Statistics Canada. A Status Report*, unpublished report, Ottawa: Statistics Canada

[29] Cochran, W.G. (1977). *Sampling Techniques*, 3rd edition. New York: Wiley & Sons.

[30] Cook, R. D. and S. Weisberg (1982), *Residuals and Influence in Regression*, London: Chapman and Hall.

[31] Dalén, J. (1987), *Practical Estimators of a Population Total Which Reduce the Impact of Large Observations*, R & D Report, Stockholm: Statistics Sweden.

[32] De Jong, W. (1996): *Designing a Complete Edit Strategy*, Combining Techniques, Statistics Netherlands, Research paper No. 9639.

[33] De Vries, W. and Van Braquel, R. (1999): *Quality systems and statistical auditing: A pragmatic approach to statistical quality management*. Work and Quality Assurance within Statistics. Office for Official Publications of the European Communities.

[34] De Waal, T. (1996): CherryPi: *A computer program for automatic edits and imputation*. Paper presented at the UN Work Session on Statistical Data Editing, Voorburg.

[35] De Waal, T., Pannekoek, J., and Scholtus, S. (2011). *Handbook of statistical data editing and imputation*. Wiley. New York

[36] Deming (1982): *Quality, Productivity and Competitive Position*, MIT.

[37] Deming (1986): *Out of the Crisis*, MIT.

[38] Di Zio, M., and Guarnera, U. (2013). *A contamination model for selective editing*. Journal of Official Statistics, xx, xxx–xxx.

[39] Draper, L., Petkunas, T. And Greenberg, B. (1990): *On-line Capabilities in SPEER*, Proceedings of Symposium 90, Measurement and Improvement of Data Quality, Ottawa: Statistics Canada, pp. 235-243.

[40] EDIMBUS (2007). *Recommended practices for editing and imputation in cross-sectional business surveys*. ISTAT, CBS, SFSO, EUROSTAT. Available from http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPM_EDIMBUS.pdf. Accessed January, 10, 2013.

[41] Engström, P. (1996). *Monitoring the editing process*, Sweden

[42] Engström (1997a)

[43] Ernst, L. R. (1980), *Comparison of Estimators of the Mean Which Adjust for Large Observations*, Sankhya, Series C, 42, pp. 1-16

[44] Fellegi, I,P. (1965), *Invited Discussion paper of the ISI.*. Session on: Automatic Proceedings of the International Statistical Institute Meetings, pp. 468-470.

[45] Fellegi, I. P. And Holt, D. (1976), A *Systematic Approach to Automatic Editing and Imputation*, Journal of the American Statistical Association, March. Vol. 71,

nº 353, pp. 17-35.

[46] Fellegi, I.P. (1998) D*ata Statistics, Information: Some issues of the Canadian Social Statistics Sscene*. Journal of official Statistics

[47] Fuller, W. A. (1991), *Simple Estimators of the Mean of Skewed Populations*, *Statistica Sinica*, 1, pp. 137-158.

[48] Gambino, J (1987), D*ealing with Outliers: A Look at Some Methods Used at Statistics Canada*, paper prepared for the Fifth Meeting of the Advisory Committee on Statistical Methods, Ottawa: Statistics Canada.

[49] Garcia-Rubio, E. and Villan, I. (1990). *DIA System: Software for the automatic imputation of qualitative data. Proceedings of the U.S*. Bureau of the Census 1990 Annual Research Conference, pp.525-537.

[50] Garvin (1983): Quality on the line, *Harvard Business Review,* September-October.

[51] Garvin (1987): *Competing on the eight dimensions of quality, Harvard Business Review*, November-December.

[52] Ghangurde, P. D. (1989a), *Outlier Robust Estimation in Finite Population Sampling, unpublished report*, Ottawa: Statistics Canada.

[53] Ghangurde, P. D. (1989b), *Outliers in Sample Surveys, Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 736-739.

[54] González, M. and Revilla, P. (1997). *Total Quality Management and the INE*. Eurostat.

[55] Graham, R (1997). *Functional Requirements of Plain Vanilla Modules and System Capabilities Technical report from the U.S*. Bureau of the Census.

[56] Granquist, L. (1984a), *On the Role of Editing*, Statistisk Tidskrift, 2. pp. 105-118.

[57] Granquist, L. (1984b), D*ata Editing and Its Impact on the Further Processing of Statistical Data Proceedings of the Workshop on the SCP*, Invite Papers, UNDP/ECE, Statistical Computing Project: ECE/UNDP/SCP/H.3, pp. 25-45.

[58] Granquist, L. (1987). *A report of the main features of a macro-editing procedure which is used in Statistics Sweden for detecting errors in individual observations*. Presented at the Data Editing Joint Group Meeting in Madrid, April 22-24,1987.

[59] Granquist, L. (1988). *A Report of the Main Features of a Macro-editing Idea Applied on the Monthly Survey on Employment and Wages in Mining, Quarrying and Manufacturing*. Report presented at the Data Editing Joint Group Meeting in Budapest, April 18-22, 1988.Statistics Sweden.

[60] Granquist, L. (1991), *Macro-Editing-A Review of Some Methods for Rationalizing the Editing of Survey Data*, Statistical Journal, 8, pp. 137-154.

[61] Granquist, L., (1995). *Improving the traditional editing process*. In: Cox, Binder, Chinnappa, Christianson, Colledge and Knott, eds., Business Survey Methods (John Wiley, New York), pp.385-401.

[62] Granquist L.(1997), *The new view on editing, International Statistics Review*, 65, 381–387.

[63] Granquist, L. (1999): I*mproving Quality by Modern Editing*. UN/ECE Work Session on Statistical Data Editing, Working Paper No, 23. Roma.

[64] Greenberg, B. (1986). *The use of implied edits and set covering in automated data editing. Statistical Research Division Report Series,* U.S. Bureau of the Census. CENSUS/SRD/RR-86-02.

[65] Gwet, J. P., and L.-P. Rivest (1992), *Outlier Resistant Alternatives to the Ratio Estimator,* Journal of the American Statistical Association, 87, pp. 1174-1182.

[66] Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W.A. Stahel (1986), *Robust Statistics: The approach Based on Influence Functions,* New York: Wiley.

[67] Hansen, M.H., Madow, W.G., and Tepping, B.J. (1983). *An evaluation of model-dependent and probability sampling inferences in sample surveys*. Journal of the American Statistical Association, 78, 776–793

[68] Hawkins, D. M. (1980), *Identification of Outliers,* London: Chapman and Hall.

[69] Hedlin, D. (1993). *Raw data compared with edited data*. Presented at the Conference of European Statisticians' Work Session on Statistical Data Editing, Stockholm.

[70] Hedlin D. (2003), *Score Functions to Reduce Business Survey Editing at the U.K.* Office for National Statistics, Journal of Official Statistics, 19, 177–199.

[71] Hedayat, A.S. and Sinha, B.K. (1991). D*esign and Inference in Finite Population Sampling*. Wiley. New York

[72] Hidiroglou, M. A., and K. P. Srinath (1981), *Some Estimators of a Population Total Containing Large Units.* Journal of the American Statistical Association, 78, pp. 690-695.

[73] Hidiroglou, M. A. (1991), *Structure of the Generalized Estimation System (GES)*. Statistics Canada.

[74] Hidiroglou, M. A., and Srinath, K. P.(1993), *Problems Associates with Designing Sub-Annual Business Surveys*. Journal of Economic Statistics, 11, 397-405.

[75] Hidiroglou, M. A. (1994), *Sampling and Estimation for Establishment Surveys: Stumbling Blocks and Progress. Proceedings of the Section on Survey Research Methods*, American Statistical Association, 153-162.

[76] Hillmer, S.C., Bell W.R and Tiao, G.*C.* (1983), *Modeling. Considerations in the Seasonal Adjustment of Economic Time Series*, Applied Time Series Analysis of Economic Data, U.S. Department of Commerce, Bureau of the Census.

[77] Holt, T.and Jones, T. (1999): *Quality work and conflicting quality objectives. Work and Quality Assurance within Statistics.* Office for Official Publications of the European Communities.

[78] Huber, P. J. (1964), *Robust Estimation of a Location Parameter*, Annals of Mathematical Statistics, 35, pp. 73-101.

[79] Huber, P. J. (1981), *Robust Statistics*, New York: Wiley.

[80] Hulliger, B. (1993), *Robustified Horvitz-Thompson Estimators, in-house report*, Swiss Federal Statistical Office.

[81] INE Spain (2010). *Industrial Turnover Indices. Industrial New Orders Received Indices. Base 2005. CNAE-09.* Methodological Manual.

[82] Juran, J.M. (1986): *The quality trilogy: A universal approach to managing quality*, Quality Progress, August.

[83] Juran, J.M. (1988): *Juran of Planning for Quality*, Free Press.

[84] Juran, J.M. (1989): *Management of Quality*, Juran Institute.

[85] Kovar, J. (1997). *What to do when an edit fails. Statistical data editing.* Volume No.2. Conference of European Statisticians. Statistical standards and studies-No.48.

[86] Kovar, J.G. (1993): *Use of Generalized Systems in Editing of Economic Survey Data*, Bulletin of the International Statistical Institute: Proceedings of the 49th Session, Contributed Papers, Book 2, pp. 77-78.

[87] Latouche M.y Berthelot J. (1992), *Use of a score function to prioritize and limit recontacts in editing business surveys*, Journal of Official Statistics, 8, 389–400.

[88] Lawrence, D., and McDavitt, C. (1994). *Significance Editing in the Australian Survey of Average Weekly Earnings*, Journal of Official Statistics, 10, pp.437-447

[89] Lawrence D. y McKenzie R. (2000), *The general application of Significance Editing*, Journal of Official Statistics, 16, 243–253.

[90] Lepp, H. and Linacre, S. (1993). *Improving the efficiency and effectiveness of editing in a statistical agency*. Bulletin of the 49th session of the International Statistical Institute, Florence, Italy.

[91] Lee, H. (1990), *Outlier-Resistant Regression Estimators*, presented at the annual Meeting of the Statistical Society of Canada, St. John's Newfoundland, Canada.

[92] Lee, H. (1991a), *Models-Based Estimators That Are Robust to Outliers*, Proceedings of the Annual Research Conference, Washington, DC: U.S. Bureau of the Census, pp. 178-202.

[93] Lee, H., E. Rancourt, and C. –E. Särndal (1991), *Experiments with Variance Estimation from Survey Data with Imputed Values*, Proceedings if the Survey Research Methods Section, American Statistical Association, pp. 690-695.

[94] Linacre, S.J. (1991): *Approaches to Quality Assurance in the Australian Bureau of Statistics Business Surveys*, Bulletin of the International Statistical Institute: Proceedings of the 48th Session, Cairo, Egypt, Book 2, pp. 237-321.

[95] Lindstrom, K. (1990). *Functions of Macro-Editing:* The Aggregate Method.

[96] Lindstrom, K. (1991). *A macro-editing - A review of some methods for rationalizing the editing of survey data*. Statistical Journal of the United Nations Economic Commission for Europe, Vol. 8, No. 2, Special Issue on Data Editing, pp.155-166.

[97] Lyberg, L., Biemer, P., and Japec, L.(1998): *Quality Improvement in Surveys – A Process Perspective*, Proceedings of American statistical Association, Dallas.

[98] Madsen, B. and Larsen, B. S. (2000). *The uses of neural networks in data editing.* Invited paper, International Conference on Establishment Surveys (ICES II), June, Buffalo, N.Y.

[99] Madsen B. and Solheim L. (2000). How to measure the effect of data editing. UN/ECE Work Session on Statistical Data Editing , Cardiff, October 18-20.

[100] Marler, R.T. and Arora, J.S. (2004). *Survey of multi-objective optimization methods for engineering*. Structural and Multidisciplinary Optimization, 26, 369–395.

[101] Morganstein, D. y Marker, D.A. (1997). *Continuous Quality Improvement in Statistical Agencies.* Survey Measurement and Process Quality, New York, Wiley, pp.475-500.

[102] M. S. Bazaraa, H. D. Sherali y C. M. Shetty (1993), *Nonlinear Programming: Theory and Algorithms*, Nueva York: Wiley.

[103] Nordbotten, S. (1965), *The Efficiency of Automatic Detection and Correction of Errors in Individual Observations as Compared with Other Means for Improving the Quality of Statistics*, Proceedings of the International Statistical Institute Meetings, pp. 442-465.

[104] Norbotten, S. (1998): *Estimating Population proportions from imputed data. Computational Statistics & Data Analysis*, Vol.27,pp.291-309.

[105] Nordbotten, S. (1998): *Improving Editing Strategies*, Proceedings of the third International Conference on Methodological Issues in Official Statistics in Stockholm.

[106] Nordbotten, S. (2000): *Evaluating Efficiency of Statistical Data Editing: General Framework*, UN/ECE Conference of European Statisticians - Methodological Material, Geneva.

[107] Pannekoek, Scholtus, and van der Loo] Pannekoek, J., Scholtus, S., and Van der Loo, M. (2013). *Automated and manual data editing: a view on process design and methodology*. Journal of Official Statistics, xx, xxx–xxx.

[108] Pierzchala, M. (1990a): *A Review of the State of the Art in Automated Data Editing and Imputation*, Journal of official Statistics, 6, pp.355-377.

[109] Pierzchala, M. (1990b): *A Review of Three Editing and Imputation Systems, Proceedings of the Survey Research Methods Section,* American Statistical Associations, pp.111-120

[110] Pritzker, L., Ogus, J. and Hansen M.H. (1965). *Computer Editing Methods – Some Applications and Results.* Bulletin of the International Statistical Institute, Proceedings of 35th Session, Belgrade, 41 (September 1965), pp.442-65.

[111] Prothero, D.L. and Wallis, K.F. (1976),*Modeling Macroeconomic Time Series*, Journal of the Royal Statistical Society, Series A, 139, Part 4**,** pp.468-85.

[112] Relander, T. (1999): *Total Quality Management and Statistics*. Work and Quality Assurance within Statistics. Office for Official Publications of the European Communities.

[113] Revilla, P., Rey, P. and Espasa., A. (1990), *Automatic Univariate Modeling of Time Series*: Application to the Industrial Production Indices, unpublished.

[114] Revilla, P. and Rey, P. (1999). *Selective editing methods based on time series modeling*. UN/ECE Work Session on Statistical Data Editing. Rome.

[115] Rey, P. and Revilla. P, (2000). *Analysis and quality control from ARIMA modelling*. UN/ECE Work Session on Statistical Data Editing. Cardiff.

[116] Rivest, L. -P. (1993b), W*insorization of Survey* Data, presented at the 49th Session of the International Statistical Institute, Firenze, Italy.

[117] Rivest, L. -P.(1993a), S*tatistical Properties of Winsorized Means for Skewed Distributing*, unpublished technical paper, Quebec City, Canada: Department of Statistics, Laval University.

[118] Rivest, L. –P., and D. Hurtubise (1993), *Some Sampling Properties of Winsorized and Truncated Means, unpublished technical paper*, Quebec City, Canada: Department of Statistics, Laval University.

[119] Rivière, P. (2002), *General data editing tools are often unsuitable to use in complex business surveys why*, ECE, .Work Session on Statistical Data Editing, Helsinki, working Paper No.30.

[120] Rousseeuw, P. J., and A. M. Leroy (1987), *Robust Regression and Outlier Detection*, New York: Wiley.

[121] Rubin, D.B. (1987). *Multiple Imputation for No response in Surveys*. J. Wiley & Sons, New York.

[122] Salgado, D., Arbués, I., and Esteban, M.E. (2012). *Two greedy algorithms for a binary quadratically constrained linear program in survey data editing*. INE Spain Working Paper 02/12. January, 10, 2013.

[123] Särndal, C.-E., Swensson, B., and Wretman, J.H. (1992). *Model assisted survey sampling*. Springer.

[124] Searls, D. T. (1966), *The Estimators for a Population Mean Which Reduces the Effect of Large True Observations*, Journal of the American Statistical Association, 61, pp. 1200-1205.

[125] Sigman R. and Wagner, D.(1997).*Algorithms for Adjusting Survey Data that Fail Balance Edits Proceedings of the Section on Survey Research Method*, American Statistical Association.

[126] Smith y Weir (2000)

[127] Statistics Sweden (2011). *User's Guide to SELEKT 1.1– A Generic Toolbox for Selective Data Editing*. Statistics Sweden document, February 17, 2011.

[128] Tambay, J.-L. (1988), *An Integrated Approach for the Treatments of Outliers in Sub-Annual Surveys*, Proceedings on the Survey Research Methods Section, American Statistical Association, pp.229-234.

[129] Tate, P., Underwood, C., Thomas, P. and Small, C., (2001) *Challenges in Developing and Implementing New Data Editing Methods for Business Surveys*. Proceedings of Statistics Canada Symposium. Achieving Data Quality in a Statistical Agency: a Methodological Perspective.

[130] Thorburn, D. (1993), *The Treatment of Outliers in Economic Statistics*, paper presented at the International Conference on Establishment Surveys, Buffalo, NY.

[131] Todaro, T.A. (1998). *Evaluation of the AGGIES automated Edit and Imputation System*. Technical report from the National Agricultural Statistics Service, U.S. Department of Agriculture.

[132] Tukey, J. W. (1977) *Exploratory data analysis*. Addison-Wesley Publishing Company, Reading, Massachusetts.

[133] Valliant, R., Dorfman, A and Royal, R. (2000): *Finite Population Sampling and Inferences*. John Wiley & Sons, INC.

[134] Wallis, K.F (1977), *Multiple Time Series Analysis and the Final Form of Econometric Models*, Econometrical, v.45, n.6, September, pp.1481-98.

[135] Van de Pol, F., Buijs, A., van der Horst, G., and de Waal, T. (1997). *Towards Integrated Business Survey Processing*. New Directions in Surveys And Censuses, Proceedings of the 1997 International Symposium, Statistics Canada.

[136] Wahlström, C. (1990), *The Effects of Editing- A Study on the Annual Survey of Financial Accounts in Sweeden*, unpublished report, F-Method No. 27, 1990-02-26, Stockholm: Statistics Sweden (in Swedish).

[137] Wein E. (2000) *The planning of data editing processes*. UN/ECE

[138] Weir, P. (1997): *Data Editing Performance Measures*, ECE, Work Session on Statistical Data Editing, Prague, working Paper No.38.

[139] Wets, R.-B. (2002). *Stochastic programming models: wait-and-see versus here-and-now*. Institute for Mathematics and Its Applications, 128.

[140] Winkler, W and Draper, L. (1996). *The New SPEER Edit System, in Statistical policy*. Working Paper 25: Data Editing Workshop And Exposition, Washington, DC: Office of Management and Budget, pp. 50-58.

[141] Zellner, A. (1979), *Statistical Analysis of Econometric Models*, Journal of the American Statistical Association v.74, n.367, September, pp.628-651.

Glosario

A.O	Aditive outlier
AGGIES	Agriculture Generalized Imputation and Edit System
ARIMA	Modelos autoregresivos integrados de medias móviles
CAPI	Computer Assisted Personal Interview
CATI	Computer Assisted Telephone Interview
DAISY	Design, Analysis and Imputation System
DIA	Depuración e Imputación Automática
D & I	Depuración e imputación
ES	Estrategía de selección
FS	Función Score
GCT	Gestión de Calidad Total
GEIS	Generalized Edit and Imputation System
ICN	Índice de Cifras de Negocios
IEP	Índice de Entradas de Pedidos
INE	Instituto Nacional de Estadística
LS	Level Shift
MSE	Error cuadrático medio
NACE	Sistema de Clasificación de Actividades Económicas
NIM	New Imputation Methodology
OCR	Optical Character Recognition
OMR	Optical Mark Recognition
RegARIMA	Modelo de regresión con residuos ARMA
SEM	Modelo econométrico simultáneo dinámico estructural
SPEER	Structured Program for Economic Editing and Referal
StEPS	Standard Economic Processing System
TC	Temporary change

Difusión de esta tesis doctoral

Algunas de las investigaciones de esta tesis han sido difundidas en varios artículos publicados en revistas especializadas o presentados en congresos científicos. Entre los de revistas especializadas destacan los publicados recientemente en el Journal of Official Statistics (JOS) y en Optimization. Entre los de congresos científicos los presentados en el seminario organizado por Naciones Unidas Work Session on Satistical Data Editing.

A continuación se relacionan estos artículos, diferenciando los relacionados directamente con el objeto de esta tesis, los relacionados con otras fases del proceso de producción de estadísticas, y, por otra parte, los presentados a congresos. En el caso de que la autoría del artículo sea compartida se menciona entre paréntesis el resto de los autores.

En el Anexo se incluyen cuatro artículos que contienen investigaciones especialmente relevantes para esta tesis.

Publicaciones

Relacionadas con la depuración e imputación

- **"**An optimization approach to selective editing". Journal of Official Statistics. Vol. 29, No. 4, 2013, pp. 1–23 (con Arbués I. y Salgado D.).

- **"**Selective Editing as a Combinatorial Optimization Problem: a General Overview". UN/ECE Work Session on Statistical Data Editing. Oslo, Norway, 24-26 September 2012. (con Arbués I. y Salgado D.).
  http://www.unece.org/stats/documents/2012.09.sde.htm

- **"**Selective Editing as a Stochastic Optimization Problem". UN/ECE Work Session on Statistical Data Editing. Ljubljana, Slovenia, 16-18 May 2011. (con Arbués I. y Saldaña S.)
  http://www.unece.org/stats/documents/2011.05.sde.htm

- "A class of stochastic optimization problems with application to selective data editing" .2010. Optimization 61, 265-268. (con Arbues, I. y Gonzalez, M.).
  http://www.tandfonline.com/doi/abs/10.1080/02331934.2010.511670

- **"**Editing Multimode Data Collections: the Spanish experience". UN/ECE Work Session on Statistical Data Editing. Neuchâtel, Switzerland, 5-7 October 2009. (con Arbués, I. y Saldaña, S.)
  http://www.unece.org/stats/documents/2009.10.sde.htm

- La depuración selectiva como un problema de optimización estocástica. Boletín de Estadística e Investigación Operativa. Vol. 25, No. 1, Febrero 2009, pp. 32-41. (con Arbués I.)
  http://www.seio.es/BEIO/Selective-editing-as-a-stochastic-optimization-problem.html

- "Editing web questionnaires: challenges and opportunities". UN/ECE Work Session on Statistical Data Editing. Bonn. Sep 2006. (con Arbués, I. y González, M.)
  http://www.unece.org/stats/documents/2006.9.sde.htm.

- "EDR Impacts on Editing" United Nation Statistical Commission. Statistical Data Editing. Volume 3. United Nation Publication. 2006.

- "EDR Impacts on Editing" UN/ECE Work Session on Statistical Data Editing. Ottawa. Mayo 2005. (con Arbués, I., González M., González Margarita y Quesada, J.)
  http://www.unece.org/stats/documents/2005.10.sde.htm.

- "Data editing by reporting enterprises". Statistical Journal of the United Nations Economic Commission for Europe. Volume 21, number 1. 2004. (con Arbués, I., González M., González Margarita y Quesada, J.)
  http://www.deepdyve.com/browse/journals/statistical-journal-of-the-united-nations-economic-commission-for-europe/2004/v21/i1

- "Encouraging respondents in Spain". The Statistics Newsletter. OCDE. Diciembre 2002. Issue nº 12. (con González, M.)

- "An E&I method based on time series modelling designed to improve timeliness". UN/ECE Work Session on Statistical Data Editing. Helsinki. Mayo 2002.
  http://www.unece.org/stats/documents/2002.05.sde.htm

- “Spanish methods to improve timeliness in the industrial production indices”. Eurostat Seminar on Short-Term Statistics-improving timeliness and cooperation. Marzo 2001. Oficina de publicaciones oficiales de la Comunidad Europea.

- "Analysis and quality control from ARIMA modelling". UN/ECE Work Session on Statistical Data Editing. Cardiff. Octubre 2000. (con Rey, P.)
  http://www.unece.org/stats/documents/2000.10.sde.htm.

- "Selective editing methods based on time series modelling". UN/ECE Work Session on Statistical Data Editing. Rome. Junio 1999. (con Rey, P.)
  http://www.unece.org/stats/documents/1999.06.sde.htm

- "Characterisation of production in different branches of Spanish industrial activity". Working Paper 91-28. Universidad Carlos III de Madrid. 1991.(con Espasa, A. y Rey, P.).

- "Caracterización de la producción en las distintas ramas de actividad industrial, mediante el análisis de series temporales". Investigaciones Económicas. 1989.

Relacionadas con otras fases de la producción estadística

- **“**Towards a Corporate-Wide Electronic Data Collection System at the National Statistical Institute of Spain”. UN/ECE Work Session on Statistical Data Editing. Ljubljana, Slovenia, 16-18 May 2011. (con Bercebal, J. M. y Maldonado, J.L.)
  http://www.unece.org/stats/documents/2011.05.sde.htm

- “El Censo Agrario 2009 retos y oportunidades”. Revista Índice Nº 37. Noviembre 2009. (con Cortina, F.)
  http://www.revistaindice.com/numero37/

- “Trying to Improve Editing Tasks Through EDR Methods”. Work Session on Statistical Data Editing Vienna, Austria, 21-23 April 2008. (con Arbués, I., González, M. y Yun, I.)
  http://www.unece.org/stats/documents/2008.04.sde.htm

- “Las grandes operaciones estadísticas estructurales del sector agrario”. Revista Índice Septiembre 2005.
  http://www.revistaindice.com/numero12/

- “Los Indicadores Industriales Coyunturales”. Información económica y técnicas de análisis en el siglo XXI. Homenaje al Profesor Dr. Jesús B. Pena Trapero. Edición coordinada por José Miguel Casas Sánchez y Antonio Pulido San Román. Publicación del INE. Madrid 2003.

- “Data editing by reporting enterprises”. UN/ECE Work Session on Statistical Data Editing. Madrid. Octubre 2003. (con González, M., González, Margarita y Quesada, J.).
  http://www.unece.org/stats/documents/2003.10.sde.htm

- “Using Total Quality Management to improve Spanish industrial statistics”. Statistics Sweden-Eurostat. The International Conference on Quality in Official Statistics. Stockholm. Mayo 2001.
  http://www.scb.se/Pages/Standard___299820.aspx

- “Total quality management and the INE”. Web de Eurostat. Diciembre 1997. (con González, M.).
  http://www.forum.europa.eu.int

- “El IPI como principal indicador económico de oferta”. Fuentes Estadísticas. Nº 30. 1997.
- "La modernización de las estadísticas industriales". Fuentes Estadísticas. Nº 17. 1996.
- "La Producción Industrial en 1994". Boletín Coyuntura Industrial. Nº6. 1995.
- "El INE y EUROSTAT actualizan las encuestas tecnológicas". Fuentes Estadísticas. Nº 4. 1995.
- "La modernización de las estadísticas industriales. Hacia un sistema integrado de encuestas". Economía Industrial. Nº 299. 1994.
- "Principales características de los nuevos índices de producción y precios industriales". Situación Española. BBV. 1993.
- "Las nuevas encuestas industriales". Situación Española. BBV. 1993.
- "La producción y el empleo en el sector industrial". La Gaceta de los negocios. 1991.
- "La nueva encuesta de salarios". Economía y Sociología del Trabajo. Nº 8/9. 1990.
- "Algunas reflexiones en torno al Índice de Producción Industrial". Boletín Trimestral de Coyuntura, Nº 21. INE. 1986.

Trabajos Presentados a Congresos

- **“**Selective Editing as a Combinatorial Optimization Problem: a General Overview”. UN/ECE Work Session on Statistical Data Editing. Oslo, Norway, 24-26 September 2012. (con Arbués, I. y Salgado, D.).
  http://www.unece.org/stats/documents/2012.09.sde.htm
- “Implementing a corporate-wide metadata driven production process at INE Spain”. The International Conference on Quality in Athens, Greece. June 2012. (con Bercebal, J.M., Hernández, P. y Maldonado, J.L.).
  http://www.q2012.gr/default.asp?p=14
- “Implementing a Quality Assurance Framework based on the Code of Practice at the National Statistical Institute of Spain”. The International Conference on Quality in Athens, Greece. June 2012. (con Piñan, A.).
  http://www.q2012.gr/default.asp?p=14
- “Trying to improve the efficiency of statistical production processes through selective data editing methods”. The 58rd Session of the International Statistical Institute. Dublin August 2011. (con Arbués, I.).
  http://www.isi2011.ie/
- **“**Selective Editing As A Stochastic Optimization Problem”. UN/ECE Work Session on Statistical Data Editing. Ljubljana, Slovenia, 16-18 May 2011. (con Arbués, I. y Saldaña, S.).
  http://www.unece.org/stats/documents/2011.05.sde.htm
- **“**Towards a Corporate-Wide Electronic Data Collection System at the National Statistical Institute of Spain”. UN/ECE Work Session on Statistical Data Editing. Ljubljana, Slovenia, 16-18 May 2011. (con Bercebal, J.M. y Maldonado, J.L.).
  http://www.unece.org/stats/documents/2011.05.sde.htm

- **"**Editing Multimode Data Collections the Spanish". UN/ECE Work Session on Statistical Data Editing. Neuchâtel, Switzerland, 5-7 October 2009. (con Arbués, I. y Saldaña, S.). http://www.unece.org/stats/documents/2009.10.sde.htm

- "A TQM approach for re-engineering business surveys" The 57rd Session of the International Statistical Institute. Durban. August 2009. http://www.statssa.gov.za/isi2009/

- **"**Offering reporting enterprises free of charge data tailored to their needs**".** The International Conference on Quality in Roma. July 2008. (con Arbués, I. y González, M.). http://q2008.istat.it/papers.html

- "Trying to Improve Editing Tasks Through Edr Methods". Work Session on Statistical Data Editing Vienna, Austria, 21-23 April 2008. (con Arbués, I., González, M. y Yun, I.). http://www.unece.org/stats/documents/2008.04.sde.htm

- "Data editing methods and strategies in modern business surveys". The 56rd Session of the International Statistical Institute. Lisboa. August 2007. (con Arbués, I.) http://isi.cbs.nl/iamamember/CD7-Lisboa2007/default_002.html

- "Editing web questionnaires: challenges and opportunities". UN/ECE Work Session on Statistical Data Editing. Bonn. Sep 2006. (con Arbués, I. y González, M.) http://www.unece.org/stats/documents/2006.9.sde.htm.

- "Forecasting Spanish short-term indicators using factor models". 26th International Symposium on Forecasting. Santander. June 2006. (con Arbués, I.)

- "Using a TQM approach to get high quality incoming data in the Spanish industrial surveys". The International Conference on Quality in Cardiff. April 2006. (con Arbués, I., González, M., González, Margarita y Quesada, J.). http://www.ons.gov.uk/ons/about-ons/user-engagement/events/past-events/q2006---european-conference-on-quality-in-survey-statistics-24-26-april-2006/index.html

- "EDR Impacts on Editing" UN/ECE Work Session on Statistical Data Editing. Ottawa. Mayo 2005. (con Arbués, I., González, M., González, Margarita y Quesada, J.). http://www.unece.org/stats/documents/2005.10.sde.htm.

- "The Measurement of Sustainable Development: The Spanish Experience". Conference of European Statisticians. Geneva. June 2005. (con Angulo, A., Berrade, C., Egido, M.L. y Saralegui, J.) http://www.unece.org/stats/douments/2005.06.ces.htm

- "Auditing by reporting enterprises. The 55rd Session of the International Statistical Institute. Sydney. April 2005. http://www.tourhosts.com.au/archive/isi2005/invit.asp

- "Implementation of Quality Programs at the National Statistical Institute of Spain: some progress in improving relationship with reporting industrial enterprises". European Conference on Quality and Methodology in official Statistics. Mainz. May 2004. (con Arribas, C.). http://q2004.destatis.de/

- "Data editing by reporting enterprises". UN/ECE Work Session on Statistical Data Editing. Madrid. Octubre 2003. (González, M., González, Margarita y Quesada, J.). http://www.unece.org/stats/documents/2003.10.sde.htm

- "Using a TQM approach to reduce enterprise burden in the Spanish industrial surveys". The 54rd Session of the International Statistical Institute. Berlin. August 2003. http://isi.cbs.nl/NLet/NLet023-03.htm

- "Approaches for improving timeliness in production index: a selective editing method". Business Statistics Director Meeting. Luxemburgo. June 2002.

- "An E&I method based on time series modelling designed to improve timeliness". UN/ECE Work Session on Statistical Data Editing. Helsinki. Mayo 2002. http://www.unece.org/stats/documents/2002.05.sde.htm

- "Time Series modelling as a tool to produce short-term indicators". The 53rd Session of the International Statistical Institute. Seul. August 2001. http://www.stat.auckland.ac.nz/~iase/publications.php?show=4

- "Using Total Quality Management to improve Spanish industrial statistics". The International Conference on Quality in Official Statistics. Stockholm. Mayo 2001. http://www.scb.se/Pages/Standard____299820.aspx

- "Spanish methods to improve timeliness in the industrial production indices". Eurostat Seminar on Short-Term Statistics-improving timeliness and cooperation. Luxemburgo. March 2001.

- "Analysis and quality control from ARIMA modelling". UN/ECE Work Session on Statistical Data Editing. Cardiff. Octubre 2000. (con Rey, P.). http://www.unece.org/stats/documents/2000.10.sde.htm.

- "Calendar Effects in the Spanish Index of Industrial Production ". The 52rd Session of the International Statistical Institute. Helsinki. August1999. http://www.stat.fi/isi99/

- "Selective editing methods based on time series modelling". UN/ECE Work Session on Statistical Data Editing. Rome. Junio 1999. (con Rey, P.). http://www.unece.org/stats/documents/1999.06.sde.htm

- "Working Day and Seasonal Adjustment in the Spanish Index of Industrial Production". International Seminar on Seasonal Adjustment Methods SAM'98. Bucarest. October 1998. http://cordis.europa.eu/fetch?CALLER=MSS_RO_NEWS_EN&ACTION=D&DOC=127&CAT=NEWS&QUERY=012d23a14a88:00a4:6ddbe530&RCN=10096

- "A Macroediting Procedure For Short Term Indicators". The 51rd Session of the International Statistical Institute. Estambul . August 1997. http://www.isi-web.org/publications/proceedings

- "Desing and Management for the Agricultural Censuses and Structural Surveys: The Spanish Experience" International Seminar on Taking the first Agricultural Census in China. Beijing. November 1995.

- "Reducing Burden on Businesses in the Spanish Industrial Surveys". The 50rd Session of the International Statistical Institute. Beijing. August 1995.

- "Automatic Procedures for the construction of Arima models". INSEE-Eurostat Workshop on Short Term Indicators. Paris. January 1994.

- "Automatic procedure for the construction of univariate Arima seasonal models". International Workshop on Seasonal Adjustment method and diagnostics. US Bureau of the Census. Washington. 1992.

- "Automatic procedure for the construction and use of Arima models with intervention Analysis". International Symposium on Forecasting. New York. June 1991.

- "Characterisation of production in different branches of Spanish industrial activity". European Meeting of the Econometric Society. Munich. Septiembre de 1989. NSF/NBER Seminar on Time Series. Madrid. September 1989.

- "Univariate models as a way to characterise an aggregate economic variable". International Symposium on Forecasting. Vancouver. 1989.

STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working paper No. 13
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UNECE Work Session on Statistical Data Editing
(27 – 29 May 2002, Helsinki, Finland)

Topic (ii): Measuring and evaluating data editing quality

AN E&I METHOD BASED ON TIME SERIES MODELLING DESIGNED TO IMPROVE TIMELINESS

Invited paper

Submitted by the National Statistical Institute, Spain[1]

Abstract

An edit and imputation method, based on time series modelling, is presented in this paper. The method has been designed to improve timeliness, nowadays an essential quality requirement of public statistics (in particular of short-term indicators), maintaining the current levels of accuracy. REGARIMA modelling is used as a tool to carry out both editing and imputation. The method is being implemented to produce Spanish monthly short-term indicators. An example of using this method in the Industrial Production Indices is also presented.

KEYWORDS: short- term indicator, TQM, selective data editing, REGARIMA modelling

I. INTRODUCTION

1. This paper follows the approach already presented in previous meetings (i.e. using time series modelling for data editing). In Rome we presented a selective editing procedure based on a kind of tools that we named “surprises”, which are functions of ARIMA forecast. In Cardiff, we used time series modelling in a different way: we obtain a set of characteristics from an estimated ARIMA with Intervention Analysis model in order to use them to improve the edits. For this meeting, the focus is on using time series modelling to improve timeliness.

2. Nowadays, public statistical offices are under continuous pressure from society, which demands more and more data, to be produced at a lower cost, with a lower respondent burden, and especially with a shorter delay. Achieving timeliness though with losses of accuracy, by trying to solve the problem in the short term, may be tempting. But it would not be a very good strategy, because public confidence in statistics would probably suffer. Therefore, the only suitable way is trying to reduce dissemination time, without any losses in accuracy.

3. Timeliness and accuracy are usually considered as a major trade-off in the production of statistics. Nevertheless, new IT tools and statistical methodologies offer the opportunity for re-engineering statistical production processes in order to improve timeliness and accuracy simultaneously.

4. Traditionally, data editing is linked with accuracy, but increasingly is also linked with other quality aspects (Kovar, 1997). In this context, some crucial questions arise. For example, what

[1] Prepared by Pedro Revilla (previlla@ine.es).

role should data editing play in achieving timeliness? Or, how can we redesign data editing processes in order to improve timeliness? Or, which new techniques can we introduce to help us to get this target?

5. This paper discusses some of these questions, presenting our experiences in improving timeliness. An edit and imputation method based on time series modeling, in particular, on REGARIMA modeling (Bell, 1999), is a major methodological tool for improving timeliness in short-term indicators. It fits into a more general programme carried out for all the Spanish industrial statistics. The target of the program is to reduce the dissemination dates maintaining the level of quality. The implementation of the programme is a direct consequence of the TQM approach used to produce industrial statistics.

6. In the following section, a general description of the TQM approach and the timeliness programme used is presented. In section III, the use of time series modeling for editing short-term indicators is discussed. An application for the Spanish Industrial Production Indices is presented in section IV. The paper ends with some conclusions.

II. TQM APPROACH AND TIMELINESS PROGRAM USED FOR INDUSTRIAL STATISTICS

7. Trying to follow TQM principles (in particular, customer satisfaction, cornerstone of TQM), the approach used in the industrial statistics (Gonzalez and Revilla, 1997) is quite simple. The starting point is asking customers about our main failures. Then, specific programs and actions are launched to correct those failures and to improve customer satisfaction.

8. Based on our customers' opinions, we have learned that the main failure of our statistics is the delay in the data production. Eighty percent of customers consider the delay our main failure. Therefore, it should constitute an essential priority to improve our timeliness.

9. To face the problem of timeliness we have launched a general programme for the industrial statistics in order to improve timeliness, maintaining the level of accuracy. Adapting questionnaires to the accounting practices of the enterprises and improving our relationship with enterprises are two key factors in achieving timeliness. Introducing selective editing methods is probably the main technical factor.

10. We have taken several measures in order to adapt questionnaires to the accounting practices of the reporting enterprises (adapting variables and valuation rules, using different models of questionnaires, using personalized questionnaires, etc.). The underlying principle in this approach is that enterprises provide data in the same way they produce them for their own use, and the statistical agency re-elaborates them for analytical purposes, if necessary. Adapting questionnaires to the accounting practices makes answering them easier and quicker. It also results in fewer errors made by the enterprises. All of this leads to an improvement both in accuracy and time of dissemination.

11. Another key success factor in achieving timeliness is improving our relationship with reporting enterprises (our suppliers). We offer tailored data on market share, in exchange for the questionnaires. Hence we change our suppliers into customers. We are trying to work using a new model of relationship with enterprises, that we name the joint venture model. The underlying principle is that our relationship with reporting enterprises should be based on a relationship of mutual use and collaboration, rather than on the legal duty of the enterprises to fill in compulsory questionnaires. Because of the increasing interest of enterprises, they are filling in

questionnaires sooner and with more care. An important point of using the joint venture model is that both timeliness and accuracy can be improved, reducing the perceived enterprise burden at the same time.

12. Timeliness should be considered a major goal in our production rather than just a small problem. Therefore, we need to save time in all the phases of the statistical process. Data editing is one of the most time-consuming statistical phases. Hence, re-engineering the data editing procedures is a need for improving timeliness. The use of selective editing methods is a key factor to improve timeliness. Using time series modelling joined to the selective editing philosophy is being quite useful to save time in editing short-term indicators.

III. THE USE OF TIME SERIES MODELLING FOR EDITING SHORT-TERM INDICATORS

13. Time series modelling is not commonly used in statistical agencies for producing short-term indicators, with the exception of seasonal adjustment.

14. Nevertheless, continuous surveys lead to a set of sequential observations collected over time. Therefore, in these surveys, the appropriate theoretical framework for their study should not be limited to that of the random variables but should rather be enlarged on random variables varying with time (i.e. the stochastic processes). Therefore, the use of models that have stochastic processes as a theoretical framework (such as time series analysis models) may be very useful. Indeed, if useful information on previous surveys is available, it should be used to the maximum in different phases of the statistical production process.

15. Certainly, the use of information of previous surveys is not new in data editing. One of the most frequent ratio edits is based on the data of the previous survey, and monthly, quarterly and annual rates are of general application. However, these methods are based on a partial use of the information of the previous surveys. It would be convenient to use, in an efficient way, the whole set of available information, that is, the entire past of the series. This means taking advantage of the whole structure of correlation (cross and auto-correlation). Another advantage of using time series modelling is that it enables us to use probabilistic data editing. This is very useful because it allows taking into account the different variability of the economic sectors, products, etc.

16. Time series modelling can be used adopting very different editing strategies. We can either build models for the microdata (or a subset of them, for example the biggest enterprises) for editing microdata and hence use a microediting approach, or we can build models for the macrodata and use a macroediting or a selective editing approach. For a specific survey the best approach would probably be combining all of them in order to improve timeliness.

IV. AN APPLICATION FOR THE INDUSTRIAL PRODUCTION INDICES

17. An edit and imputation method based on REGARIMA modelling is being used in the Spanish National Statistical Institute to elaborate industrial short-term indicators. The models are used for microediting, macroediting and selective editing. Micro and macro imputation are also carried out based on the models. In this paper we focus on the Industrial Production Indices. Similar formulas are used for other short-term indicators.

18. A monthly survey is carried out by mail in order to calculate the Industrial Production Indices. A panel sample of about 14000 enterprises is used. One single variable, the production, in a particular physical unit (tons, litres, etc.) or in monetary value, is requested from each

enterprise. As a result of the survey, we have a microdata set $q_{i,j,t}$, that is, the production figure for the product i, reported by the enterprise *j* at month t. From the microdata set, the index for product i is calculated as:

$$I_{i,t} = I_{i,t-1} \frac{\sum_j q_{i,j,t}}{\sum_j q_{i,j,t-1}}$$

Where *j* is the set of enterprises with valid values at both t and $t-1$.

And, from these product indices, Laspeyres aggregated indices are calculated at successive levels of breakdown of the economic activities classification (at the top of the aggregation is the total industry). The following formula is used:

$$I_t = \sum_i w_i I_{i,t}$$

where the base year weights w_i are based on the value added (for activities aggregation) or the value of the production (for products aggregation).

19. We use the same kind of models (REGARIMA) for the micro and the macrodata series. Since the number of time series to handle is very large and it is difficult and time consuming to build models for all of them we need an automatic procedure. We use an automatic method developed by Revilla, Rey and Espasa that fits into the Box-Jenkins iterative modelling strategy of identify, estimate and diagnostic checking (Box and Jenkins, 1970). A straightforward use of ARIMA models is not sufficient to capture calendar variations, because they are not exactly periodic. Regression models are used to handle calendar effects and other deterministic variations (for example, a strike). To specify the intervention variables we have found that some subject matter knowledge about the behaviour of the production data is needed.

20. Therefore, the overall models are a sum of ARIMA and regression models (REGARIMA models):

$$ln\, q_{i,j,t} = \frac{\theta_{i,j}(B)\,\Theta_{i,j}(B^{12})}{\varphi_{i,j}(B)\,\Phi_{i,j}(B^{12})} a_{i,j,t} + \sum_h \frac{\alpha_{i,j,h}(B)}{\delta_{i,j,h}(B)} A_{i,j,h,t}$$ for modelling the microdata.

$$ln\, I_{i,t} = \frac{\theta_i(B)\,\Theta_i(B^{12})}{\varphi_i(B)\,\Phi_i(B^{12})} a_{i,t} + \sum_h \frac{\alpha_{i,h}(B)}{\delta_{i,h}(B)} A_{i,h,t}$$ for modelling the indices.

where:

- $ln\, q_{i,j,t}$ is the neperian logarithm of the production figure for the product i, reported by the enterprise *j*.
- $ln\, I_{i,t}$ is the neperian logarithm of the industrial production index for product (or activity) i.
- B is the backshift operator, $B^k(I_t) = I_{t-k}$.
- $\theta(B), \varphi(B), \Theta(B^{12}), \Phi(B^{12}), \alpha_h(B), \delta_h(B)$ are polynomials in the backshift operator, for the product (or activity) i, or for the enterprises i,j.

- $a_{i,j,t}$ and $a_{i,t}$ are white noise variables i. i. d. $N(0,\sigma_{i,j})$ and $N(0,\sigma_i)$ respectively.
- $A_{i,j,h,t}$ and $A_{i,h,t}$ are intervention variables.

21. For using a **microediting approach** we would have to fit a REGARIMA model for each series of enterprise data. In our application, we do not fit a model for all of the enterprises but only for the most influential ones (usually the biggest enterprises). The model is used for both editing and imputation. A confidence interval (for example a 95% interval) can be constructed from the model:

$$P\left[\hat{q}_{i,j,t} - 1{,}96\ \sigma_{ij} < q_{i,j,t} < \hat{q}_{i,j,t} + 1{,}96\ \sigma_{i,j}\right] = 0{,}95$$

Where $\hat{q}_{i,j,t}$ is the one-step forecast for $q_{i,j,t}$, and the outliers can be defined as the microdata outside the interval. The imputed value for $q_{i,j,t}$ would be $\hat{q}_{i,j,t}$

22. For using a **macroediting approach**, an REGARIMA model has been constructed for each of the index series of products and activities. The model is used for both editing and imputation. A confidence interval (for example a 95% interval) can be constructed from the model:

$$P\left[\hat{I}_{i,t} - 1{,}96\ \sigma_i < I_{i,t} < \hat{I}_{i,t} + 1{,}96\ \sigma_i\right] = 0{,}95$$

Where $\hat{I}_{i,t}$ is the one-step forecast for $I_{i,t}$, and the outliers can be defined as the indices outside the interval. The imputed value for $I_{i,t}$ would be $\hat{I}_{i,t}$

23. In using a **selective editing approach** we have to solve two problems: to detect outliers in the macrodata (the indices) and to detect the influential microdata. In order to face the first problem we have designed some tools, the "surprises", that are functions of the REGARIMA model forecast (in particular, from the one-step ahead forecasted values):

The ***Surprise (or simple surprise)*** $S_{i,t}$ for the index $I_{i,t}$ is the relative change between the observed and the forecasted data:

$$S_{i,t} = \frac{I_{i,t} - \hat{I}_{i,t}}{\hat{I}_{i,t}}$$

If we calculate the one-step ahead forecast $\ln \hat{I}_{i,t}$ for $\ln I_{i,t}$, the one-step ahead forecast error is:

$$e_{i,t} = \ln I_{i,t} - \ln \hat{I}_{i,t}$$

Since the one-step ahead forecast error $e_{i,t}$ is a $N(0,\sigma_i)$ white noise process and $\ln I_{i,t} - \ln \hat{I}_{i,t} \cong \left(I_{i,t} - \hat{I}_{i,t}\right) \big/ \hat{I}_{i,t}$, we have that $S_{i,t}$ is approximately $N(0,\sigma_i)$. Hence, a confidence interval (for example, a 95% interval) for the surprises can be constructed:
$P\left[-1.96\sigma_i < S_{i,t} \le 1.96\sigma_i\right] = 0.95$

and the outliers can be defined as the indices whose surprise is outside the interval.

The ***Standard surprise*** for the index $I_{i,t}$ is:

$$\frac{S_{i,t}}{\sigma_i} = \frac{I_{i,t} - \hat{I}_{i,t}}{\hat{I}_{i,t}} \frac{1}{\sigma_i}$$

It allows the direct comparison of indices with different variability.

The ***Weighted standard surprise*** for the index $I_{i,t}$ is:

$$\frac{S_{i,t}}{\sigma_i} w_i = \frac{I_{i,t} - \hat{I}_{i,t}}{\hat{I}_{i,t}} \frac{w_i}{\sigma_i}$$

It allows the ranking of the indices taking into account not only the surprise magnitude but also the different weights.

24. Once we have detected and ranked the surprising indices (i.e., indices that are not coherent with their past behaviour and therefore can be considered as outliers) we need to measure the impact of each of the microdata on these surprising indices. For this purpose, we use the "influences".

The ***Influence of an individual datum over an aggregated magnitude*** is defined as the difference between the observed aggregated magnitude and the value for this same magnitude when the individual datum is not available.

The ***Influence of the individual datum*** $q_{i_0,j_0,t}$ ***over the product index*** $I_{i_0,t}$ is:

$$INF_{i_0,j_0}^{I_{i_0,t}} = I_{i_0,t-1} \frac{\sum_j q_{i_0,j,t}}{\sum_j q_{i_0,j,t-1}} - I_{i_0,t-1} \frac{\sum_{j \neq j_0} q_{i_0,j,t} + \hat{q}_{i_0,j_0,t}}{\sum_j q_{i_0,j,t-1}} = I_{i_0,t-1} \frac{q_{i_0,j_0,t} - \hat{q}_{i_0,j_0,t}}{\sum_j q_{i_0,j,t-1}}$$

where $\hat{q}_{i_0,j_0,t}$ is an imputed value for the individual datum $q_{i_0,j_0,t}$.

and the ***Influence over the aggregated index*** I_t is:

$$INF_{i_0,j_0}^{I_t} = \sum_i w_i I_{i,t} - \left[\sum_{i \neq i_0} w_i I_{i,t} + w_{i_0} I_{i_0,t-1} \frac{\sum_{j \neq j_0} q_{i_0,j,t} + \hat{q}_{i_0,j_0,t}}{\sum_j q_{i_0,j,t-1}} \right] = w_{i_0} I_{i_0,t-1} \frac{q_{i_0,j_0,t} - \hat{q}_{i_0,j_0,t}}{\sum_j q_{i_0,j,t-1}}$$

This expression measures the impact of the microdata on the index by means of the following factors:

- The product (or activity) weight w_{i_0}.
- The index $I_{i_0,t-1}$ which "updates" the above weight.

- A measure of the relative discrepancy between the real and the imputed individual datum $\frac{q_{i_0,j_0,t} - \hat{q}_{i_0,j_0,t}}{\sum_j q_{i_0,j,t-1}}$.

It may be proved that the microdata which are more influential on the aggregated index are also the more influential on the surprises of that index.

25. These "influences" allow us to prioritize the suspicious values in the microdata in order to verify and recontact fewer enterprises. Hence improvements in timeliness are achieved.

26. As an example of using this methodology two computer printouts are shown. In table 1 the sectors are ranked according the weighted standard surprise. Sector 4243 is the one with the highest. Looking in table 2 for the enterprises that most influence the index of this sector, we find an enterprise whose influence is not only the highest, but also much higher in comparison with the others. Hence this enterprise should be checked.

27. Using this methodology, the most influential suspicious values could immediately be detected. Hence the most important errors could be checked and corrected before the index is disseminated for first time.

28. Three kinds of **imputations** would be possible to handle the non-response for one or more enterprise data $q_{i,j_o,t}$ to disseminate the indices $I_{i,t}$:

(i) A traditional imputation based on the available enterprise data

$$\hat{q}_{i,j_0,t} = q_{i,j_0,t-1} \frac{\sum_{j \neq j_o} q_{i,j,t}}{\sum_{j \neq j_o} q_{i,j,t}}$$

(ii) A model-based imputation for $q_{i,j_o,t}$ based on its past series, that is $\hat{q}_{i,j_o,t}$

(iii) A model-based imputation for the whole index $I_{i,t}$ based on its past series, that is $\hat{I}_{i,t}$

29. A choice among the former methods can be made depending on various circumstances. For example, if the trend of the enterprises inside the product (or activity) is similar, the traditional imputation usually works properly. However, if the trend is not very similar a model-based imputation works better. In the same way, if the non-response rate is low for the index, a micro imputation (traditional or model-based) usually works properly. On the contrary, if the non-response rate for the index is high and there are not many enterprises in the index a model-based imputation for the whole index is preferable. Therefore, model-based imputation may improve the estimates of the indices.

V. CONCLUSIONS

30. Improving timeliness without any losses in accuracy is a major challenge for public statisticians today. Data editing should not only be linked to accuracy but also to other quality

aspects, for example timeliness. Data editing is one of the most time-consuming statistical phases. Hence, re-engineering the data editing procedures is a need for improving timeliness.

31. According to our experience, the use of time series modeling is being quite useful to save time in editing short-term indicators.

REFERENCES

[1] Bell, W.R. (1999). "*An overview of REGARIMA modeling*". Forthcoming Research Report. Statistical Research Division. U.S. Census Bureau.
[2] Box, G.E.P. and Jenkins, G.M. (1970*). "Time Series Analysis, Forecasting and Control*", ed. Holden-Day, San Francisco.
[3] González, M. and Revilla, P. (1997). **"***Total Quality Management and the INE"*. Eurostat Web, www.forum.europa.eu.int/.
[4] Kovar, J. (1997). "*What to do when an edit fails*". Statistical data editing. Volume No.2. Conference of European Statisticians. Statistical standards and studies-No.48.
[5] Revilla, P. and Rey, P. (1999*). **"**Selective editing methods based on time series modelling"*. UN/ECE Work Session on Statistical Data Editing. Rome.
[6] Rey, P. and Revilla. P, (2000). **"***Analysis and quality control from ARIMA modelling"*. UN/ECE Work Session on Statistical Data Editing. Cardiff.

Table 1 **SURPRISES**

Sector	Actual Rate	Forecasted Rate	Simple Surprise	Standard Surprise	Weighted Standard Surprise
4243	70,28	3,32	64,73	3,79	17,10
2511	-27,73	-3,29	-25,25	-3,11	-16,93
4110	-50,24	-6,89	-70,96	-6,84	-16,89
2514	-15,92	4,64	-19,62	-3,00	-16,87
2512	39,39	-11,83	58,12	7,22	16,51
4752	-0,74	2,06	-2,75	-1,09	-15,66
3299	-11,97	4,45	-15,70	-2,02	-15,57
4751	22,82	-7,36	32,55	2,34	14,64
3630	-0,28	3,68	-3,82	-0,81	-14,54
3166	15,97	5,83	9,58	1,89	13,92

Table 2 **INFLUENCES**

Enterprise	Influence
1	143,41
2	37,60
3	-22,80
4	14,38
5	8,90
6	-7,52
7	7,35
8	6,81

STATISTICAL COMMISSION and Working Paper No. 19
ECONOMIC COMMISSION FOR EUROPE English only

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Data Editing
(Cardiff, United Kingdom, 18-20 October 2000)

Topic III: New techniques and tools for editing imputation

ANALYSIS AND QUALITY CONTROL FROM ARIMA MODELLING

Submitted by the National Statistical Institute and Sociological Research Centre, Spain[1]

Invited paper

Abstract: In this paper, we use ARIMA modelling to estimate a set of characteristics of a short-term indicator (for example, the index of industrial production), as trends, seasonal variations, cyclical oscillations, unpredictability, deterministic effects (as a strike), etc. Thus for each sector and product (more than 1000), we construct a vector of values corresponding to the above-mentioned characteristics, that can be used for data editing.

KEYWORDS*: continuous surveys, data editing, ARIMA models, Intervention Analysis.*

I. INTRODUCTION

1. This paper follows the approach already presented in the last meeting in Rome (i.e. the use of time series analysis for data editing). The central idea of this approach is that, if useful information from previous surveys is available, it should be used as much as possible in the editing process.

2. In this context, the appropriate theoretical framework for editing continuous surveys should not be limited to that of the static random variables, but should rather be enlarged on that of sequences of random variables varying with time. Therefore, stochastic processes and, in particular, time series models may be used.

3. At the previous meeting in Rome, we presented a selective editing procedure based on a kind of tools that we named "surprises", which are functions of ARIMA forecasts. For this meeting, we use time series modelling in a different way. We propose to estimate a univariate ARIMA with Intervention Analysis model (Box-Tiao, 1975) to describe the characteristics of a short time indicator.

4. The idea comes from the fact that in editing a survey, we need as much information as possible about the phenomenon that we try to measure. On the other hand, different subsets of data often show a very different variability and behaviour. For example, when editing the Spanish monthly indices of industrial production, we can find extremely small (even zero) production data values for August, because summer holidays are usually taken in this month in Spain. Of course, these data must not be considered as outliers (i.e. suspicious data items) if we have information about this seasonal pattern. But an additional problem is that, this seasonal behaviour in August is very different from one branch to another (even there are branches which production does not decrease but strongly grows in August, as in beer production). For this reason, it is important to acquire information about the different dynamic characteristics of each of the branches to improve editing rules and strategies.

[1] Prepared by Pedro Revilla and Pilar Rey.

5. In this paper, we use ARIMA with Intervention Analysis modelling to estimate a set of characteristics of a short term indicator, as trends, cyclical oscillations, seasonal variations, calendar effects, special event effects (as a strike), unpredictability, etc. Thus for each of the branches and products (more than 1000), we construct a vector of values corresponding to the characteristics above mentioned, that can be used as a useful tool by the data editing team.

6. Even when, from a theoretical point of view, multivariate models (that picked up the correlation of all the variables of the survey) would be appropriate, we restrict ourselves to the univariate environment, because of the difficulties in using multivariate time series modelling in practice. On the other hand, the use of univariate ARIMA models to describe the data characteristics of an economic phenomenon has a sound methodological foundation.

7. Under fairly general conditions (see Prothero and Wallis (1976), Wallis (1977), Zellner (1979), etc.) any variable, which is determined within a structural simultaneous and dynamic econometric model (SEM) is generated in a univariate way by an ARIMA with Intervention Analysis model. In the latter model, the intervention analysis component picks up the contribution of the dummy variables of the SEM model and/or the effect of certain intervention analysis, which affect the exogenous variables of that model.

8. To the extent that the SEM model reflects the characteristics of the real world, the ARIMA with Intervention Analysis model corresponding to an endogenous variable of the SEM model incorporates, inefficiently but certainly consistently, the basic characteristics of that variable.

II. APPLICATION TO THE INDUSTRIAL PRODUCTION INDICES

9. The approach presented here is being used in the National Statistical Institute of Spain for editing the Industrial Production Indices. It could also be implemented for other short-term indicators. A monthly survey is carried out by mail in order to calculate the Industrial Production Indices. A panel sample of about 9000 enterprises is used. The response rate is about 95%.

10. One single variable, the production volume, measured in physical units (tons, litres, etc.) or in monetary value, is requested from each enterprise. The indices for products are calculated as chain indices. From these product indices, Laspeyres aggregated indices are calculated at successive levels of breakdown of the economic branches classification (at the top of the aggregation is the total industry). The following formula is used:

$$I_t = \sum_i w_i \, I_{i,t}$$

where the base year weights w_i are calculated using the value added (for branches aggregation) or the value of the production (for products aggregation). The branches have been studied at five successive levels of breakdown, in accordance with the National Classification of Economic Activities existing in Spain.

11. In all of the series the sample contains 96 observations, from January 1992 to December 1999. To characterise the evolution of production in the different branches of industrial activity ARIMA models with Intervention Analysis (ARIMA-IA) are used.

12. Since the number of time series to handle is very large and it is difficult and time consuming to build models for all of them, we need an automatic procedure. We use an automatic method developed by Revilla, Rey and Espasa (1990), that fits into the Box-Jenkins iterative modelling strategy of identify,

estimate and diagnostic checking. Using this method, an ARIMA model has been constructed for each of the series of indices.

13. A straightforward use of ARIMA models is not sufficient to capture calendar effects and other deterministic variations (for example, a strike). Regression models are used to handle them. Therefore, the overall models are a sum of ARIMA and regression models:

$$\ln I_{i,t} = \frac{q_i(B)\Theta_i(B^{12})}{j_i(B)\Phi_i(B^{12})}a_{i,t} + \sum_h \frac{a_{i,h}(B)}{d_{i,h}(B)}A_{i,h,t}$$

where:

- $\ln I_{i,t}$ is the neperian logarithm of the industrial production index for product (or activity) i.
- B is the backshift operator, $B^k(I_t) = I_{t-k}$.
- $q_i(B), j_i(B), \Theta_i(B^{12}), \Phi_i(B^{12}), a_{i,h}(B), d_{i,h}(B)$ are polynomials in the backshift operator.
- $a_{i,t}$ are white noise variables i. i. d. $N(0, s)$.
- $A_{i,h,t}$ are intervention variables.

14. To specify the intervention variables we have found that automatic procedures are not suitable for all of the series and some subject matter knowledge about the behaviour of the indices is needed.

15. In the following, we will consider one by one the different aspects we have studied in the models, describing how they have been implemented.

A. Level behaviour

16. An adequate description of the nature of the long-term trend of the series is contained in the models. This trend is determined by the contribution in the final forecast function of the real positive unit roots of the autoregresive factor and by the contribution of the possible non-zero mean of the stationary series. The presence of d roots of the type already mentioned means that the long-term trend is a time polynomial of order (d-1), the coefficients of which are determined by the initial conditions in which the system is located. The presence of a non-zero means increases the previous polynomial with a term of order d with a deterministic coefficient.

17. Thus, whenever the series is stationary, it would not require differences. When the model specifies one difference the series will show local oscillations in level; when the model specifies two differences the series will have a quasi-linear trend, etc.

B. Seasonal behaviour

18. The models also can contain a factor, which picks up a seasonal cycle with a period of 12 time units. The complex unit roots and real negatives of the autoregressive polynomial form this factor. If none of these roots is repeated its contribution in the final forecast function consists of 12 stable, additive, seasonal factors which at all times are determined by the initial conditions of the system. As a result, the long-term path of the series is made up of these seasonal factors and the trend described in A.

Whenever a seasonal cycle has been found the seasonal factors of the final forecast function have been calculated.

C. Calendar effects

19. Spanish Indices of Industrial Production series are strongly affected by calendar effects. Because these data are reported monthly and represent a total of the production for each month, they contain variations due to the length and day-of-the-week composition of the month. Monthly production data are also affected by holidays, in which the levels of production are lower than on ordinary working days. There are holidays that occur each year at the same time, and movable holidays, for example Easter, that occur at various dates during March and April. To make the situation more complicated, public holidays in Spain vary from one year to another, and from one "Autonomous Community" to another. In order to handle these calendar effects, we incorporate this information as deterministic input variables.

20. Instead of the most commonly used (see Hillmer, Bell and Tiao, 1983) seven trading-day variables (number of Mondays, Tuesdays, etc.), we construct one single variable, adapted to the behaviour of the Spanish industrial production. This variable tries to measure the number of working days, removing Saturdays, Sundays and holidays. To achieve a suitable computation, those holidays that vary from one "Autonomous Community" to another have to be weighted by the industrial value added of the base year.

21. By constructing the model on the logarithmic transformation of the series, the parameter associated with this dummy variable can be interpreted as the proportional increase in production in comparison with that of a similar month with one less working day.

22. Another deterministic variable containing the Easter effect (which takes in March and April the values that indicate the proportion of days affected by these holidays and in the other months value zero) is included in the model (see Hillmer, Bell and Tiao, 1983). The parameter, which affects this artificial variable, can be interpreted as the proportional variation suffered by production as a result of this effect. Hence, all calendar effects are summarised using only two variables.

D. Other deterministic effects

23. It is possible to find deterministic contributions in the trend and/or seasonal factors of the series from the intervention analysis. In February 1997 the transporting sector went on strike. For most of the industrial branches, this strike caused a shortage of raw materials. The expected effect would have been an immediate reduction in the level of production. But, in fact, after some periods of observation, we found in some series an increase in the months of March and April that compensated the decrease in February. After consulting some respondent factories, we learned that they have tried to fulfil the orders of the clients by working extra in the next two months.

24. So, we have constructed two different variables for peaking up the effects of the strike:

1) The H_t variable, where:

$$H_t = \begin{cases} 1.0, & t = February\ 1997 \\ -0.6, & t = March\ 1997 \\ -0.4, & t = April\ 1997 \\ 0.0, & t \neq Feb., Mar.\ or Apr.\ 1997 \end{cases}$$

2) The P_t pulse variable, where:

$$P_t = \begin{cases} 1, & t = \textit{February } 1997 \\ 0, & t \neq \textit{February } 1997 \end{cases}$$

25. We have accepted that the strike had an effect on an industrial production index series when the intervention variable parameter is significantly different from zero. When the parameters are significant for the two of them, we have chosen the intervention variable that produces a smaller residual standard deviation of the ARIMA-IA model. As we have used the logarithmic transformation of the series, it is possible to interpret the value of the parameters as a percentage effect on the level of the original series.

E. Outliers

26. We have made in the series a detection of unusual or unexpected observations. Depending on their nature, we have detected three types of outliers: Additive Outlier (A. O), Level Shift (L. S.), and Temporary Change (T. C.).

27. For defining and detecting the location and type of an outlier, we follow the approach of Chang et al. (1988) using the estimated residuals of the Arima model.

> *The study of outliers can be used to detect special events which may affect production in a particular period of time, to analyse if they appear fortuitously or not in the different branches, the month in which they often appear, etc.*

F. Uncertainty

28. The last characteristic we have considered is a measure of the uncertainty about the future evolution of the series, expressed by the standard deviation of the one period ahead forecasting errors. Each series at a given time can be broken down into two components: its prediction based on previous observations and a prediction error. A good measure of uncertainty is given by the standard deviation of these prediction errors.

III. RESULTS OF THE STUDY

29. As an example of the information that may be obtained from the models, the main dynamic characteristics of the indices for the total industry and for the branches at two digits level are presented in Table 1.

30. The following are shown in the table: 1) Identification of the branch by means of the code of the National Classification of Economic Activities of Spain, 2) Behaviour of the level, 3) Whether the sector has seasonal behaviour or not, 4**)** A measure of the effect of working days, when it affects the series, 5) A measure of the effect of Easter, when it affects the series, 6) A measure of the effect of the 1997 strike, when it affected the series, 7) The outliers, the date when they are produced and their type, 8) The degree of uncertainty about future production levels.

31. To illustrate the information that can be obtained from Table 1, the characteristics of total industry are described. Industrial production from 1992 onwards shows a trend in its level and a stochastic seasonal nature, which implies different behaviour in production activity in different months of the year (in particular, it shows a decrease in the holiday months, especially in August).

32. The industrial output is sensitive to day-of-the-week composition and periods of holidays. More specifically, the existence of one less working day causes a 1,9% drop in production. Likewise, Easter causes a fall in production of 4,0% distributed in March and April according to the proportion of days affected by this holiday in each year.

33. The transport sector strike in February 1997 caused a 3,6% reduction on the level of production of this month, compensated in March and April. It does not show outliers above three standard deviations. The degree of uncertainty regarding production for the next month is 2,1%.

34. Likewise, useful information can be gained of the different industrial branches. It is easy to see that they show very different patterns of behaviour.

REFERENCES

Box, G.E.P. and Tiao, G.C. (1975),"Intervention Analysis with Applications to Economic and Environmental Problems", Journal of the American Statistical Association, 349.

Chang, I., Tiao, G. C. and Chen, C. (1988), “Estimation of Time Series Parameters in the Presence of Outliers”, Technometrics, 30, pp.193-204.

Hillmer, S.C., Bell W.R and Tiao, G.*C.* (1983), "Modelling. Considerations in the Seasonal Adjustment of Economic Time Series", Applied Time Series Analysis of Economic Data, U.S. Department of Commerce, Bureau of the Census.

Prothero, D.L. and Wallis, K.F. (1976),"Modelling Macroeconomic Time Series", Journal of the Royal Statistical Society, Series A, 139, Part 4, pp.468-85.

Revilla, P., Rey, P. and Espasa., A. (1990), "Automatic Univariate Modelling of Time Series: Application to the Industrial Production Indices", unpublished.

Wallis, K.F (1977), "Multiple Time Series Analysis and the Final Form of Econometric Models", Econometrica, v.45, n.6, September, pp.1481-98

Zellner, A. (1979), "Statistical Analysis of Econometric Models", Journal of the American Statistical Association v.74, n.367, September, pp.628-651.

Table 1

Series	Level Behaviour	Seasonal Behaviour	Working-Days Effect (%)	Easter Effect (%)	Feb 1997 Strike Effect (%)	Outliers Date	Outliers Type	Uncertainty (%)
0	Trend	Yes	1.9	-4.0	-3.6			2.1
11	Local Oscillations	No	2.7					8.3
12	Trend	Yes				3/94 2/95 10/95 10/97 1/98	A. O. A. O. L. S. L. S. L. S.	13.0
13	Local Oscillations	Yes						6.4
14	Local Oscillations	Yes				9/93 11/93 5/96 8/93 3/98 10/98 11/98 12/99	L. S. A. O. A. O. A. O. A. O. A. O. A. O. A. O.	17.7
15	Trend	Yes		-5.4	-7.8			3.6
21	Local Oscillations	No				6/94 2/96 4/96 11/96 3/97 8/98 5/98	L. S. T. C. T. C. L. S. T. C. L. S. L. S.	11.4
22	Trend	Yes		-7.2	-6.7	1/94 12/95 8/96 9/99	A. O. A. O. A. O. A. O.	3.6
23	Trend	Yes	3.1					5.7
24	Trend	Yes	1.2	-2.2		2/93 8/93 1/94 12/94 9/96 3/97	A. O. A. O. A. O. A. O. T. C. L. S.	3.7
25	Trend	Yes	1.8	-4.1		12/93 10/95	L. S. A. O.	3.4
31	Trend	Yes	3.2			8/93	T. C.	4.5
32	Trend	Yes	2.9		-7.5	8/94	A. O.	6.2
33	Local Oscillations	No				5/93 7/93 1/94	A. O. A. O. A. O.	20.6
34	Trend	Yes	3.3			8/93	A. O.	5.1
35	Trend	Yes						13.0
36	Trend	Yes	3.5	-4.8	-12.2	8/93 2/94 4/94 8/94 8/95 8/96 4/98	A. O. L. S. A. O. A. O. A. O. A. O. A. O.	3.0
37	Local Oscillations	No				5/94 9/94 12/94 2/96 9/96 5/97 11/98 3/99	L. S. L. S. L. S. L. S. T. C. L. S. L. S. T. C.	6.5

Series	Level Behaviour	Seasonal Behaviour	Working-Days Effect (%)	Easter Effect (%)	Feb 1997 Strike Effect (%)	Outliers Date	Outliers Type	Uncertainty (%)
38	Trend	Yes	2.5	-6.1		5/94 8/94 3/97	T. C. T. C. L. S.	7.0
39	Trend	Yes	3.0			8/94 6/99	A. O. L. S.	5.9
41	Trend	Yes	2.4		5.3	1/94 7/95 1/96	A. O. A. O. T. C.	2.7
43	Trend	Yes	3.2			9/93 8/99	L. S. A. O.	5.2
44	Trend	Yes	4.4			4/93 8/93 7/97 6/99	A. O. L. S. L. S. L. S.	5.9
45	Trend	Yes	3.1			4/93 2/94 9/94 12/95 4/96 9/97 5/99 6/99	T. C. L. S. L. S. T. C. A. O. A. O. T. C. A. O.	2.3
46	Trend	Yes	1.6	-5.3	-3.7	8/93 8/96	A. O. A. O.	3.9
47	Trend	Yes	2.1		-3.3			3.0
48	Trend	Yes	2.7	-3.2	-10.0	8/93 8/95 8/98 8/99	A. O. A. O. A. O. A. O.	3.7
49	Trend	Yes	4.2	-14.5		4/93 8/93 5/94 10/94 12/94 3/95 7/96 10/96 12/96 6/97 11/98	L. S. A. O. A. O. A.O. A. O. A. O. L. S. L. S. A. O. L. S. A. O.	5.4

Optimization
Vol. 00, No. 00, Month 200x, 1–19

RESEARCH ARTICLE

A Class of stochastic optimization problems with application to selective data editing

Ignacio Arbués[†‡], Margarita González[†] and Pedro Revilla[†]

[†] *S. G. de Estadísticas Industriales y Agrarias, Instituto Nacional de Estadística, Castellana 183, 28071, Madrid, Spain*

(*Received 00 Month 200x; in final form 00 Month 200x*)

We present a class of stochastic optimization problems with constraints expressed in terms of expectation and with partial knowledge of the outcome in advance to the decision. The constraints imply that the problem cannot be reduced to a deterministic one. Since the knowledge of the outcome is relevant to the decision, it is necessary to seek the solution in a space of random variables. We prove that under convexity conditions, a duality method can be used to solve the problem. An application to statistical data editing is also presented. The search of a good selective editing strategy is stated as an optimization problem in which the objective is to minimize the expected workload with the constraint that the expected error of the aggregates computed with the edited data is below a certain constant. We present the results of real data experimentation and the comparison with a well known method.

Keywords: Stochastic Programming; Optimization in Banach Spaces; Selective Editing; Score Function

AMS Subject Classification: 90C15; 90C46; 90C90

1. Introduction

Let us consider the following elements: a decision variable x under our control; the outcome ω of a probability space; a function f that depends on x and ω that we want maximize; and a function g also depending on x and ω that we want in some sense not to exceed zero. Various optimization problems can be posed under these assumptions.

First of all, the problem is quite different depending on whether we know ω before we take the the decision about x or not. In the first case (wait and see), ω is fixed and thus, we have deterministic functions $f(\cdot, \omega)$ and $g(\cdot, \omega)$. Therefore, we can solve a nonstochastic optimization problem for each outcome ω. Consequently, the solution depends on ω and it is a random variable (this case is studied, for example, in [2]).

In the case ω is unknown when x is decided (here and now), the usual approach is to pose a stochastic optimization problem by setting as the new objective function $\mathbf{E}[f(x, \omega)]$ (see [4] for a discussion on this matter).

In this paper, we study problems in which we have a partial knowledge of ω and thus, in that respect, they are in somewhere in the middle between here-and-now and wait-and-see. In subsequent sections, we will explain in full detail the meaning of "partial knowledge".

ISSN: 0233-1934 print/ISSN 1029-4945 online

DOI: 10.1080/0233193YYxxxxxxxx
http://www.informaworld.com

The use of some knowledge about ω to decide about x implies that x will depend on ω and thus it will be a random variable. On the other hand, in our problem we have also constraints expressed in terms of expectation, that is, $\mathbf{E}[g(x,\omega)] \leq 0$.

We also analyse an application of these problems to statistical data editing. Efficient editing methods are critical for the statistical offices. In the past, it was customary to edit manually every questionnaire collected in a survey before computing aggregates. Nowadays, exhaustive manual editing is considered inefficient, since most of the editing work has no consequences at the aggregate level and can in fact even damage the quality of the data (see [1] and [6]).

Selective editing methods are strategies to select a subset of the questionnaires collected in a survey to be subject to extensive editing. A reason why this is convenient is that it is more likely to improve quality by editing some units than by editing some others, either because the first ones are more suspect to have an error or because the error if it exists has probably more impact in the aggregated data. Thus, it is reasonable to expect that a good selective editing strategy can be found that balances two aims: (i) good quality at the aggregate level and (ii) less manual editing work.

This task is often done by defining a score function (SF), which is used to prioritise some units. When several variables are collected for the same unit, different *local* score functions may be computed and then, combined into a *global* score function. Finally, those units with score over a certain threshold are manually edited.

Thus, when designing a selective editing process it is necessary to decide,

- Whether to use SF or not.
- The local score functions.
- How to combine them into a global score function (sum, maximum, ...).
- The threshold.

At this time, the points above are being dealt with in an empirical way because of to the lack of any theoretical support. In [8], [9] and [6] some guidelines are proposed to build score functions, but they rely in the criterion of the practitioner. In this paper, we describe a theoretical framework which, under some assumptions, answers the questions above. For this purpose, we will formally define the concept of *selection strategy*. This allows to state the problem of selective editing as an optimization problem in which the objective is to minimize the expected workload with the constraint that the expected remaining error after editing the selected units is below a certain bound.

We present the general problem in section 2. We also describe how to use a duality method to solve the problem. In section 3, we show the results of a simulation experiment. In sections 4 through 7 we describe how to apply the method to selective editing. In section 8, results of the application of the method to real data are presented. Finally, some conclusions are discussed.

2. The general problem

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $N, m, p \in \mathbb{N}$. We consider the functions $f(x,\omega) : \mathbb{R}^N \times \Omega \mapsto \mathbb{R}$, $g_1(x) : \mathbb{R}^N \mapsto \mathbb{R}^m$ and $g_2(x,\omega) : \mathbb{R}^N \times \Omega \mapsto \mathbb{R}^p$. We need some assumptions.

Assumption 2.1 f and g_2 are $\mathcal{F}-$measurable and integrable in ω.

Assumption 2.2 $-f$ and g_2 are convex in x; g_1 is convex.

The convexity implies that f and g_2 are continuous with respect to x (see theorem

2.35 in [12]) and thus, they are Carathéodory functions (that is, continuous in x and measurable in ω). For the same reason, g_1 is continuous. With these assumptions, for any random vector $\omega \in \Omega \mapsto X(\omega) \in \mathbb{R}^N$, $f(X(\omega),\omega)$ and $g_2(X(\omega),\omega)$ are $\mathcal{F}-$measurable.

Let $\mathcal{G} \subset \mathcal{F}$ be a $\sigma-$field. We denote by $\mathcal{O}(\mathcal{G})$ the set of the $\mathcal{G}-$measurable functions such that $\mathbf{E}[f(X(\omega),\omega)]$ and $\mathbf{E}[g_2(X(\omega),\omega)]$ exist in $\mathbb{R}$.

We can now pose the following problem.

$$[P] \qquad \max_X \qquad \mathbf{E}[f(X(\omega),\omega)] \tag{1}$$

$$X \in \mathcal{O}(\mathcal{G}), \text{s.t. } g_1(X(\omega)) \leq 0 \ a.s., \mathbf{E}[g_2(X(\omega),\omega)] \leq 0. \tag{2}$$

The fact that we seek X among the $\mathcal{G}-$measurable random variables means that all the information we have about ω is whether $\omega \in A$, for any $A \in \mathcal{G}$, that is, if the event A happened.

In order to prove the existence of solutions to $[P]$, we make some further assumptions.

Assumption 2.3 For $\theta = -f, g_2$, there exists $\alpha \in (1,+\infty)$ such that $\mathbf{E}[\theta(X(\omega),\omega)] \in \mathbb{R}$ when $\mathbf{E}|X|^\alpha < +\infty$ and the function $X \in L^\alpha(\Omega;\mathcal{G}) \mapsto \mathbf{E}[\theta(X(\omega),\omega)]$ is lower semicontinuous with respect to the strong topology of $L^\alpha(\Omega;\mathcal{G})$.

The semicontinuity condition holds, for example, when the function θ satisfies a Lipschitz condition with respect to the first argument, say $|\theta(x,\omega) - \theta(y,\omega)| \leq C(\omega)\|x-y\|$, such that $C(\omega)$ has a finite moment of order greater than one, in which case we have not only semicontinuity, but continuity.

Assumption 2.4 The set $\{x \in \mathbb{R}^N : g_1(x) \leq 0\}$ is bounded.

Assumption 2.5 $\mathcal{G}$ is countably generated.

Assumption 2.5 is technical and do not seem to imply an important loss of generality for most of practical applications. With these assumptions, we can prove an existence result for $[P]$.

Proposition 2.6: *If assumptions 2.1–2.5 hold then [P] has at least one solution.*

Proof: From assumption 2.4, it follows that the solutions must be bounded with probability one, but since the measure of Ω is finite, they must have finite L^α-norm. Consequently, we can seek a solution in $L^\alpha(\Omega;\mathcal{G})$, which is a Banach space (Theorem 3.11 in [13]). Under assumption 2.5, the closed unit ball B in $L^\alpha(\Omega;\mathcal{G})$ is compact with respect to the $*$-weak topology (see [3], p.246). On the other, hand $L^\alpha(\Omega;\mathcal{G})$ is reflexive, so we do not have to distinguish between the $*$-weak topology and the weak topology.

From assumption 2.3 the set $M = \{X \in L^\alpha(\Omega;\mathcal{G}) : g_1(X) \leq 0 \ a.s., \mathbf{E}[g_2(X(\omega),\omega)] \leq 0\}$ is strongly closed and convex and then, weakly closed. From assumption 2.4, it is also bounded. Then, there exists some $\varepsilon > 0$ such that $\varepsilon M \subset B$. Since εM is a weakly closed subset of a weakly compact set, it is weakly compact itself. M is also weakly compact because is homothetic to εM. On the other hand, $X \mapsto -\mathbf{E}[f(X(\omega),\omega)]$ is convex and strongly lower semicontinuous. Thus, it is weakly lower semicontinuous and it attains a minimum in M, that is, $[P]$ has a solution. □

In the remaining of section 2, we will show a practical way to obtain the solutions. We will analyse in 2.1 the case that $\mathcal{G} = \mathcal{F}$ (full information) and then, in 2.2 we

will describe how to reduce the general case (partial information) to the former one.

2.1. *Duality*

Now, we assume that $\mathcal{G} = \mathcal{F}$. Under this condition, we will establish two results. First, we will show (proposition 2.8) that solutions can be obtained to problem $[P]$ by solving a dual problem $[D]$ formulated using a Lagrangian function. The second result (proposition 2.9) states that the dual problem can be reduced to a family of deterministic optimization problems in $\mathbb{R}^N$ depending on ω. For this, we need an additional assumption.

Assumption 2.7 There is a random vector X_0 such that $g_1(X_0) \leq 0$ *a.s.* and $\mathbf{E}[g_2(X_0(\omega), \omega)] < 0$.

Assumption 2.7 is a classical regularity condition necessary for the duality methods and it is usually known as Slater's condition.

Let us define the Lagrange function,

$$\mathcal{L}(X, \lambda) = \mathbf{E}[f(X(\omega), \omega)] - \lambda^T \mathbf{E}[g_2(X(\omega), \omega)]. \tag{3}$$

We can now define the problem

$$\begin{aligned} [P(\lambda)] \qquad & \max_X \qquad \mathcal{L}(X, \lambda) \\ & X \in \mathcal{O}(\mathcal{G}), \text{s.t. } g_1(X(\omega)) \leq 0 \ a.s. \end{aligned} \tag{4}$$

The dual function is defined as $\varphi(\lambda) = \sup\{\mathcal{L}(X, \lambda) : X \in \mathcal{O}(\mathcal{G}), g_1(X(\omega)) \leq 0 \ \ a.s.\}$. If the supremum is a maximum, we denote by X_λ the point where it is attained. The dual problem is,

$$\begin{aligned} [D] \ & \min_\lambda \varphi(\lambda) \\ & \text{s.t.} \ \ \lambda \in \mathbb{R}^p, \lambda \geq 0. \end{aligned}$$

This problem is of great interest for us because of the following proposition.

Proposition 2.8: *If the assumptions of proposition 2.6 and assumption 2.7 hold then,*

i) There exist solutions to $[D]$.
ii) If X *is a solution to* $[P]$ *and* $\bar{\lambda}$ *is a solution to* $[D]$ *then,* X *is a solution to* $[P(\bar{\lambda})]$.

Proof: Since $[P]$ has a solution, then $\mathbf{E}[f(X(\omega), \omega)]$ is bounded from above in M and thus, we can apply theorem 1, p. 224 in [10]. The first part of the theorem states that $[D]$ has a solution and the second part implies *(ii)*. □

Therefore, if we know how to solve $[D]$ and $[P(\lambda)]$ for any λ, we can obtain solutions to $[P]$ by solving $[P(\bar{\lambda})]$, where $\bar{\lambda}$ is a solution to $[D]$. If we know how to compute $\varphi(\lambda)$, then $[D]$ can be solved by numerical methods since it is a finite–dimensional optimization problem, but if X_λ is a solution to $[P(\lambda)]$, then $\varphi(\lambda) = \mathcal{L}(X_\lambda, \lambda)$. Thus, all we need is a method to solve $[P(\lambda)]$.

The advantage of $P[\lambda]$ with respect to $[P]$ is that the stochastic terms are all in the objective function. Consequently, the solution can be easily obtained by solving

a deterministic optimization problem for any outcome $\omega \in \Omega$.

$$[P_D(\lambda,\omega)] \max_x L(x,\lambda,\omega)$$
$$\text{s.t.} \quad x \in \mathbb{R}^N, g_1(x) \leq 0$$

where $L(x,\lambda,\omega) = f(x,\omega) - \lambda^T g_2(x,\omega)$. We can relate this problem to $[P(\lambda)]$ by virtue of the following direct consequence of theorem 14.60 in [12].

Proposition 2.9: *In the assumptions of proposition 2.6, there exists a solution X_λ to $[P(\lambda)]$ such that for any $\omega \in \Omega$,*

i) $X_\lambda(\omega)$ is a solution to $P_D(\lambda,\omega)$.
ii) $\varphi(\lambda) = \mathbf{E}[L(X_\lambda(\omega),\lambda,\omega)]$.

The optimal $\bar{\lambda}$ will be obtained maximizing φ. Since φ is described in terms of expectation, it is necessary either to know the real distribution of the terms in L and compute explicitly its expectation or to estimate it. In practical applications, the explicit computation will usually not be feasible.

2.2. *Partial information*

In the proof of proposition 2.9, we use the $\mathcal{F}-$measurability of f and g_2 to obtain a measurable solution X_λ. The validity of this argument depends on the fact that $\mathcal{G} = \mathcal{F}$, so the proposition breaks down if we seek $\mathcal{G}-$measurable solutions X while f and g_2 are only measurable with respect to a strictly larger $\sigma-$field $\mathcal{F}$. Fortunately, we can still apply the ideas of the previous section with some changes.

Before introducing the main results of this section, we will show an example. Let us assume that f is linear with respect to x, that is, we can write $f(x,\omega) = Z(\omega)^T x$, where $Z(\omega)$ is a $N \times 1$ random vector. Then, for a $\mathcal{G}-$measurable X, it holds

$$\mathbf{E}[f(X(\omega),\omega)] = \mathbf{E}[\mathbf{E}[f(X(\omega),\omega)|\mathcal{G}]] = \mathbf{E}[\mathbf{E}[Z^T X|\mathcal{G}]] = \mathbf{E}[\mathbf{E}[Z^T|\mathcal{G}]X],$$

where the first identity is a consequence of the Law of Total Expectations and the third to the $\mathcal{G}-$measurability of X. Consequently, if we define $f^*(x,\omega) = U(\omega)^T x$, where $U = \mathbf{E}[Z|\mathcal{G}]$, then we get that $\mathbf{E}[f(X(\omega),\omega)] = \mathbf{E}[f^*(X(\omega),\omega)]$, but now f^* is $\mathcal{G}-$measurable with respect to ω. We can define g_2^* in the same way and thus we can express $[P]$ in such a way that the results of section 2.1 can be directly applied.

We will prove that appropriate functions f^* and g_2^* can be defined without the linearity assumption. We will show that if $\theta^*(x,\omega)$ is defined as the conditional expectation of $\theta(x,\omega)$ (later, f and g_2) with respect to $\mathcal{G}$ with x fixed, then for any $\mathcal{G}-$measurable X it holds that $\mathbf{E}[\theta(X(\omega),\omega)|\mathcal{G}] = \theta^*(X(\omega),\omega)$.

In other words, when we calculate the conditional expectation of $\theta(X(\omega),\omega)$ with respect to $\mathcal{G}$ and X is $\mathcal{G}-$measurable, we can take conditional expectation in the second argument as if the first were deterministic. In order to prove this, we need first the following generalization of the monotone convergence theorem.

Lemma 2.10: *Let $\{\xi_n\}_n$ be a sequence of integrable random variables and $\mathcal{G}$ a $\sigma-$field. If $\xi_n(\omega) \nearrow \xi(\omega)$ and ξ is integrable, then $\mathbf{E}[\xi_n|\mathcal{G}] \nearrow \mathbf{E}[\xi|\mathcal{G}]$ with probability one.*

Proof: The sequence of random variables $\eta_n := \mathbf{E}[\xi_n|\mathcal{G}]$ is nondecreasing with probability one. Consequently, $\eta_n(\omega) \nearrow \eta(\omega)$ with probability one. The random variable η is $\mathcal{G}-$measurable since it is limit of $\mathcal{G}-$measurable r.v.'s (corollary to

theorem 1.14, in [13]). On the other hand, for any $A \in \mathcal{G}$ it holds that

$$\int_A \eta P(\mathrm{d}\omega) = \lim_n \int_A \mathbf{E}[\xi_n|\mathcal{G}]P(\mathrm{d}\omega) = \lim_n \int_A \xi_n P(\mathrm{d}\omega) = \int_A \xi P(\mathrm{d}\omega),$$

where the first and third identities hold by the monotone convergence theorem and the second by the definition of the conditional expectation (page 445 in [3]). Then, η is a version of $\mathbf{E}[\xi|\mathcal{G}]$. □

With this lemma, we can prove the following proposition.

Proposition 2.11: *Let $(x,\omega) \in \mathbb{R}^N \times \Omega \mapsto \theta(x,\omega) \in \mathbb{R}$ be lower semicontinuous in x and $\mathcal{F}$-measurable in ω. If $\theta^*(x,\omega) = \mathbf{E}[\theta(x,\omega)|\mathcal{G}]$, then for any $\mathcal{G}-$measurable X*

$$\mathbf{E}[\theta(X(\omega),\omega)|\mathcal{G}] = \theta^*(X(\omega),\omega). \tag{5}$$

Proof: In order to simplify the notation, we will prove it for $N = 1$ and the arguments can be easily generalized for $N > 1$. Let us first assume that for a certain $L \in \mathbb{N}$, $\theta(x,\omega) = 0$ when $|x| > L$ and consider for $k = -L2^n, \ldots, L2^n - 1$, the intervals $I_{n,k} = [k2^{-n}, (k+1)2^{-n})$. Then, we can define the functions $\theta_{n,k}(\omega) = \inf_{z \in I_{nk}} \theta(z,\omega)$ and

$$\Psi_{n,k}(z) = \begin{cases} 1 \text{ if } z \in I_{n,k} \\ 0 \text{ if } z \notin I_{n,k} \end{cases}$$

We can write

$$\theta_n(z,\omega) = \sum_{k=-n2^n-1}^{n2^n} \theta_{n,k}(\omega)\Psi_{n,k}(z).$$

Let us see that $\theta_n(z,\omega) \nearrow \theta(z,\omega)$. For any z and n, there exist k, l such that $z \in I_{n+1,k} \subset I_{n,l}$. Then, $\theta_n(z,\omega) \leq \theta_{n+1}(z,\omega) \leq \theta(z,\omega)$. Now, for any $\epsilon > 0$, there exists n_0 such that for any $n \geq n_0$, $|z - z'| \leq 2^{-n}$ implies $\theta(z',\omega) \geq \theta(z,\omega) - \epsilon$. Consequently, for any $n \geq n_0$, $\theta(z,\omega) \geq \theta_n(z,\omega) \geq \theta(z,\omega) - \epsilon$. Therefore, $\theta_n(z,\omega) \nearrow \theta(z,\omega)$.

Now, by lemma 2.10,

$$\theta^*(z,\omega) = \mathbf{E}[\theta(z,\cdot)|\mathcal{G}] = \lim_n \sum_{k=-n2^n-1}^{n2^n} \Psi_{nk}(z)\mathbf{E}[\theta_{n,k}|\mathcal{G}].$$

If $X(\omega)$ is $\mathcal{G}-$measurable, then

$$\theta^*(X(\omega),\omega) = \lim_n \sum_{k=-n2^n-1}^{n2^n} \Psi_{nk}(X(\omega))\mathbf{E}[\theta_{n,k}|\mathcal{G}] = \lim_n \mathbf{E}[\theta_n(X(\omega),\omega)|\mathcal{G}]$$

and we conclude by using again lemma 2.10. The case that the support of θ is not bounded can be dealt with by using $\theta(x,\omega)\kappa_L(x)$, where $\kappa_L(x)$ equals one in $[-L, L]$ and has bounded support. □

By definition of conditional expectation, θ^* is $\mathcal{G}-$measurable with respect to ω. If θ is convex in x, so is θ^* and the identity (5) ensures that if the strong semicontinuity

of assumption 2.3 holds for θ, then it holds for θ^* as well. Consequently, if we define $f^*(x,\omega) = \mathbf{E}[f(x,\omega)|\mathcal{G}]$ and $g_2^*(x,\omega) = \mathbf{E}[g_2(x,\omega)|\mathcal{G}]$, they inherit from f and g_2 the properties required for the assumptions of proposition 2.8. Let us consider the problem

$$[P^*] \qquad \max \qquad \mathbf{E}[f^*(X(\omega),\omega)] \tag{6}$$

$$X \in \mathcal{O}(\mathcal{G}), \text{s.t. } g_1(X(\omega)) \leq 0 \ a.s., \mathbf{E}[g_2^*(X(\omega),\omega)] \leq 0. \tag{7}$$

Proposition 2.11 ensures that $[P^*]$ is equivalent to $[P]$ in the sense that the domain of the problem, the objective function and the constraints are the same, if only expressed in a different way. Consequently, every solution to $[P]$ is solution to $[P^*]$ and conversely. However, since f^* and g_2^* are measurable with respect to the same $\sigma-$field as X, we can use the results of subsection 2.1.

3. Simulation.

Proposition 2.8 provides the optimal solution to $[P]$, but the dual problem has to be solved by numerical methods in order to obtain the Lagrange multipliers. Thus, we have designed an example of problem $[P]$ such that λ can be computed analytically and we have obtained estimate values from simulation in order to compare with the true ones.

Let us consider the following case,

$$f(x,\omega) = \mathbf{1}^T x; \qquad g_1(x) = (x^T - \mathbf{1}^T, -x^T)^T; \qquad g_2(x,\omega) = \sum_{i=1}^{N} \delta_i(\omega) x_i - d,$$

where $\{\delta_i\}_{i=1,\ldots,N}$ are independent and uniformly distributed in $[0,1]$ and d is a positive constant. It is easy to see that the solution to $[P(\lambda)]$ is

$$X_i = \begin{cases} 1 \text{ if } \lambda\delta_i < 1 \\ 0 \text{ if } \lambda\delta_i > 1 \end{cases}. \tag{8}$$

Then, we can see that for $\lambda \geq 1$, $\mathbf{E}[X_i] = \lambda^{-1}$ and $\mathbf{E}[X_i\delta_i] = (2\lambda^2)^{-1}$, whereas for $\lambda < 1$, $\mathbf{E}[X_i] = 1$ and $\mathbf{E}[X_i\delta_i] = 1/2$. Hence,

$$\varphi(\lambda) = \begin{cases} N(1-\frac{\lambda}{2}) + \lambda d \text{ if } \lambda < 1 \\ \frac{N}{2\lambda} + \lambda d \qquad\quad \text{if } \lambda \geq 1 \end{cases},$$

and consequently, for $d < N/2$ the minimum is attained at $\bar{\lambda} = (\frac{N}{2d})^{1/2}$.

We will estimate $\bar{\lambda}$ by applying the sample–path optimization or sample average approximation method (SAA, see [5], [11], [15]). The SAA method can also be applied to constrained problems (see [14]), but we do not apply this method to the original problem $[P]$ because there the decision variable is a $\mathcal{G}-$measurable random variable.

We simulate a sample of size M of $\delta = (\delta^1, \ldots, \delta^M)$ and then we minimize the function $\hat{\varphi}(\lambda) = M^{-1}\sum_{j=1}^{M}\varphi_j(\lambda)$, where $\varphi_j(\lambda) = L(x^j,\lambda)$, $L(x,\lambda) = \mathbf{1}^T x - \lambda(\sum \delta_i x_i - d)$, $x^j = (x_1^j, \ldots, x_N^j)$ and x_i^j is defined as in (8) for the j-th simulated value of δ. The convergence of the solution of the approximate problems is guaranteed, for example, by theorem 3.9 in [5]. The hypotheses of the theorem can

be easily checked for our problem. This method is also applied in our application of section 8, where we use a sample of real data instead of a simulated one.

The minimization of $\hat{\varphi}$ has been performed using the function *fmincon* of the mathematical pack MATLAB.

The results for a range of values of N and M are presented in table 1, suggesting that when N is large, even moderate values of M allow to achieve considerable accuracy. We have chosen $d = N/4$ and thus, the theoretical $\bar{\lambda}$ is $\sqrt{2}$.

N	M=1	M=5	M=10	M=25	M=50	M=100	M=250
10	0.2052	0.0966	0.0686	0.0413	0.0330	0.0236	0.0132
50	0.0919	0.0375	0.0313	0.0190	0.0134	0.0104	0.0057
100	0.0694	0.0278	0.0224	0.0117	0.0097	0.0062	0.0044
500	0.0286	0.0125	0.0097	0.0060	0.0044	0.0030	0.0018
1,000	0.0177	0.0090	0.0063	0.0047	0.0030	0.0018	0.0014
5,000	0.0089	0.0041	0.0032	0.0019	0.0014	0.0010	0.0006
10,000	0.0070	0.0028	0.0024	0.0012	0.0011	0.0007	0.0004

Table 1. RMS error in the estimation of $\bar{\lambda}$.

4. The selective editing problem

Let us introduce some notation,

- x_t^{ij} is the *true* value of variable j in questionnaire i at period t, with $i = 1, \ldots, N$ and $j = 1, \ldots, q$.
- $\tilde{x}_t^{ij} = x_t^{ij} + \varepsilon_t^{ij}$ is the *observed* value of variable j in questionnaire i at period t, ε_t^{ij} being the observation error.
- $X_t^k = \sum \omega_{ij}^k x_t^{ij}$ is the k-th statistic computed with the true values ($\tilde{X}_t^k$ is computed with the observed ones), according to certain weightings ω_{ij}^k with k ranging from 1 to p.

The linearity assumption implies a loss of generality, which is nevertheless not very important in the usual practice of statistical offices. Many statistics are in fact linear aggregates of the data. In some other cases such as indices, they are ratios whose denominator depends on past values that can be considered as constant when editing current values. When the statistic is nonlinear, the applicability of the method depends on the accuracy of a first-order Taylor expansion in $\{x_t^{ij}\}$.

Let $(\Omega, \mathcal{F}, P)$ be a probability space. We assume that x_t^{ij} and ε_t^{ij} are random variables with respect to that space. There can be other random variables relevant to the selection process. Among them, some are known at the moment of the selection, such as $\tilde{x}_t^{ij}$, x_s^{ij} with $s < t$ or even variables from other surveys. The assumption that x_s^{ij} is known is equivalent to assume that when editing period t, the data from previous periods have been edited enough and does not contain errors. Deterministic variables such as working days may also be useful to detect anomalous data. We will denote by $\mathcal{G}_t$ the $\sigma-$field generated by all the information available up to time t. In order to avoid heavy notation, we omit the subscript t when no ambiguity arises.

Our aim is to find an adequate selection strategy. A selective editing strategy should indicate for any i whether questionnaire i will be edited or not and this has to be decided using the information available. In fact, we will allow the strategy not to determine precisely whether the unit is edited but only with a certain probability.

Definition 4.1: A selection strategy (SS) with respect to $\mathcal{G}_t$ is a $\mathcal{G}_t$-measurable random vector $R = (R_1, \ldots, R_N)^T$ such that $R_i \in [0, 1]$.

We denote by $S(\mathcal{G}_t)$ the set of all the SS with respect to $\mathcal{G}_t$. The interpretation of r is that questionnaire i is edited with probability $1 - R_i$. To allow $0 \leq R_i \leq 1$ instead of the more restrictive $R_i \in \{0, 1\}$ is theoretically and practically convenient because then, the set of strategies is convex and techniques of convex optimization from section 2 can be used. Moreover, it could happen that the optimal value over this generalized space were better than the restricted case (just as in hypothesis testing a randomized test can have a greater power than any nonrandomized one). If for a certain unit, $R_i \in (0, 1)$, then the unit is effectively edited depending on whether $\chi_t^i < R_i$, where χ_t^i is a random variable distributed uniformly in the interval $[0, 1]$, and independent from every other variable in our framework (in order to accommodate these further random variables, we consider an augmented probability space, $(\Omega^*, \mathcal{F}^*, P^*)$, which is the product space of the original one times the suitable choice of $(\Omega_1, \mathcal{F}_1, P_1)$; only occasionally we have to refer to the augmented one). We denote by $\tilde{R}_i$ the indicator variable of the event $\chi_t^i < R_i$ and $\tilde{R} = (\tilde{R}_1, \ldots, \tilde{R}_N)$. If a SS satisfies $R_i \in \{0, 1\}$ a.s., then $\tilde{R} = R$ a.s. and we say that R is integer. The set of integer SS is denoted by $S_I(\mathcal{G}_t)$. In our case study, the solutions obtained are integer or approximately integer.

It is also convenient to have a formal definition of a Score Function.

Definition 4.2: Let R be a SS, $\delta = (\delta_1, \ldots, \delta_N)^T$ a random vector and $\Theta \in \mathbb{R}$, such that $R_i = 1$ if and only if $\delta_i \leq \Theta$. Then, we say that δ is a Score Function generating R with threshold Θ.

In order to formally pose the problem, we will assume that after manual editing, the true values of a questionnaire are obtained. Thus, we have to consider only the observed and true values. We define the *edited* statistic $X^k(R)$ as the one calculated with the values obtained after editing according to a certain choice. We can write $X^k(R) = \sum \omega_{ij}^k (x_t^{ij} + \tilde{R}_i \varepsilon_t^{ij})$.

The quality of $X^k(R)$ has to be measured according to a certain function. In this paper, we consider only the Squared Error, $(X^k(R) - X^k)^2$. This choice makes easier the theoretical analysis. It remains for future research to adapt the method for other loss functions. The value of the loss function can be written as

$$(X^k(R) - X^k)^2 = \sum_{i,i'} \epsilon_i^k \epsilon_{i'}^k \tilde{R}_i \tilde{R}_{i'}, \tag{9}$$

where $\epsilon_i^k = \sum_j \omega_{ij}^k \varepsilon_t^{ij}$ or, in matrix form, as $(X^k(R) - X^k)^2 = \tilde{R}' E^k \tilde{R}$, with $E^k = \{E_{i,i'}^k\}_{i,i'}$ and $E_{i,i'}^k = \epsilon_i^k \epsilon_{i'}^k$. We can now state the problem of selection as an optimization problem.

$$\begin{aligned} [P_Q] \max_R \ & \mathbf{E}[1^T \tilde{R}] \\ \text{s.t.} \quad & R \in S(\mathcal{G}_t), \mathbf{E}[\tilde{R}^T E^k \tilde{R}] \leq e_k^2, k = 1, \ldots, p, \end{aligned}$$

with $e_k^2 > 0$.

In section 6 we will see the solution to this problem. The vector in the cost function can be replaced for another one in case the editing work were considered different among units (e.g., if we want to reduce the burden for some respondents; this possibility is not dealt with in this paper).

Let us now analyse the expression (9). We can decompose it as

$$(X^k(R) - X^k)^2 = \sum_i (\epsilon_i^k)^2 \tilde{R}_i + \sum_{i \neq i'} \epsilon_i^k \epsilon_{i'}^k \tilde{R}_i \tilde{R}_{i'}. \tag{10}$$

The first term in the RHS of (10) accounts for the individual impact of each error independently of its sign. In the second term the products are negative when the factors have different signs. Therefore, in order to reduce the total error, a strategy will be better if it tends to leave unedited those couples of units with different signs. The nonlinearity of the second term makes the calculations more involved. For that reason, we will also study the problem neglecting the second term.

$$\begin{aligned} &[P_L] \max_R \mathbf{E}[1^T \tilde{R}] \\ &\quad \text{s.t.} \quad R \in S(\mathcal{G}_t), \mathbf{E}[D^k \tilde{R}] \leq e_k^2, k = 1, \ldots, p, \end{aligned}$$

where $D^k = (D_1^k, \ldots, D_N^k)^T$, $D_i^k = (\epsilon_i^k)^2$

This problem is easier than P_Q because the constraints are linear. In section 5 we will see that the solution is given by a certain score function. Since there is no theoretical justification for neglecting the quadratic terms, the SS solution of the linear problem has to be empirically justified by the results obtained with real data.

5. Linear case

In this section, we analyse problem $[P_L]$. The reduction to the full information case yields, by virtue of proposition 2.11, $f^*(r, \omega) = 1^T r$ and $g_2^*(r, \omega) = \Delta^k(\omega) r$, where $\Delta^k = \mathbf{E}[D^k | \mathcal{G}_t]$. From now onwards, we write f and g_2 instead of f^* and g_2^*. Therefore, $[P_L]$ can be stated as a particular case of $[P]$ with

$$\begin{array}{ll} X = R & \mathcal{G} = \mathcal{G}_t \\ g_1(r) = (r^T - \mathbf{1}^T, -r^T)^T & f(r, \omega) = \mathbf{1}^T r \\ g_2(r, \omega) = (\Delta^1(\omega) r - e_1^2, \ldots, \Delta^p(\omega) r - e_p^2)^T & \end{array} \tag{11}$$

In order to avoid heavy notation, the dependence of Δ^k on ω will be implicit in the subsequent analysis. Note that in our application, f does not depend on ω.

Proposition 5.1: *If $\mathbf{E}|\Delta^k|^\beta < +\infty$ for all k with $\beta > 1$ and $\mathcal{G}_t$ is countably generated, then the assumptions of proposition 2.8 hold for the case defined by (11).*

Proof: For fixed r, g_2 is measurable in ω because it is a liner combination of measurable functions. Therefore, assumption 2.1 holds. Since $-f$, g_1 and g_2 are linear, they are convex, so assumption 2.2 holds as well. In order to check assumption 2.3, we set α such that $1/\alpha + 1/\beta = 1$. Now, for $R \in L^\alpha$, $\Delta^k R \in L^1$, so $\mathbf{E}[g_2^k(R(\omega), \omega)]$ is defined, whereas $\mathbf{E}[f(R(\omega), \omega)]$ is defined because $R \in L^\alpha$ implies $R \in L^1$.

The strong lower semicontinuity (in fact, continuity) of g_2 is a consequence of the following inequality

$$\begin{aligned} |\mathbf{E}[g_2(S(\omega), \omega)] - \mathbf{E}[g_2(R(\omega), \omega)]| &= \left| \int [g_2^k(S(\omega), \omega) - g_2^k(R(\omega), \omega)] P(\mathrm{d}\omega) \right| \leq \\ &\leq \int |g_2^k(S(\omega), \omega) - g_2^k(R(\omega), \omega)| P(\mathrm{d}\omega) \leq \|\Delta^k\|_{L^\beta} \|S - R\|_{L^\alpha}, \end{aligned}$$

where we have used Hölder's inequality.

Assumptions 2.4 and 2.5 are obvious. We check 2.7 by setting x_0 as the null vector. □

The deterministic problem in this case can be expressed as

$$[P_D(\lambda,\omega)] \max_r 1^T r - \sum_k \lambda_k(\Delta^k r - e_k^2)$$
$$\text{s.t.} \quad r_i \in [0,1].$$

By applying the Karush–Kuhn–Tucker conditions, we get a solution given by

$$r_i = \begin{cases} 1 \text{ if } \lambda^T \Delta_i < 1 \\ 0 \text{ if } \lambda^T \Delta_i > 1, \end{cases} \tag{12}$$

where $\Delta_i = (\Delta_i^1, \ldots, \Delta_i^p)^T$. The case $\lambda^T \Delta_i = 1$ is a zero-probability event when dealing with quantitative data, given that the distribution of Δ_i should be continuous. This implies,

Proposition 5.2: *The solution to $[P_L]$ is the SS generated by the Score Function $\delta_i = \lambda^T \Delta_i$ with threshold equal to 1.*

We describe in section 7 how to use a model for the practical computation of Δ^k. In order to estimate the dual function $\varphi(\lambda) = \mathbf{E}[L(R_\lambda, \lambda,)]$ we replace expectation for the mean value over a sample as we did in section 3. However, in this case we can obtain a sample of real data instead of a simulated one because we have realizations of the variables for several periods. Thus, we can seek the optimum of $\hat{\varphi}(\lambda) = \frac{1}{h} \sum_{t=t_0}^{t_0+h-1} L_t(r_\lambda^t, \lambda)$.

Summarizing the foregoing discussion, we have proved that,

- The optimum solution to $[P_L]$ is a score-function method whose global SF is a linear combination of the local SF of the different indicators $k = 1, \ldots, p$.
- The local score function of indicator k is given by $\Delta_i^k = \mathbf{E}[(\epsilon_i^k)^2 | \mathcal{G}_t]$.
- The coefficients λ^k of the linear combination are those which maximize $\varphi(\lambda)$.
- The threshold is 1.

6. Quadratic case

The outline of this section is similar to that of the previous one, but the quadratic problem poses some further difficulties, in particular, that the constraints are not convex. Therefore, we will replace them by some convex ones in such a way that under some assumptions the solutions remain the same.

Lemma 6.1: *The following identity holds.*

$$\mathbf{E}[\tilde{R}^T E^k \tilde{R} | \mathcal{G}_t] = R^T \Gamma^k R + (\Delta^k)^T R \tag{13}$$

where $\Gamma^k = \{\Gamma_{ij}^k\}_{ij}$ *and,*

$$\Gamma_{ij}^k = \begin{cases} \mathbf{E}[\epsilon_i^k \epsilon_j^k | \mathcal{G}_t] \text{ if } i \neq j \\ 0 \qquad\quad \text{if } i = j. \end{cases}$$

Proof: Since $(\tilde{R}_i)^2 = \tilde{R}_i$ can write

$$\mathbf{E}[\tilde{R}^T E^k \tilde{R}|\mathcal{G}_t] = \sum_i \mathbf{E}[(\epsilon_i^k)^2 \tilde{R}_i|\mathcal{G}_t] + \sum_{i\neq i'} \mathbf{E}[\epsilon_i^k \epsilon_{i'}^k \tilde{R}_i \tilde{R}_{i'}|\mathcal{G}_t]. \tag{14}$$

If we define $\mathcal{G}_t^* = \mathcal{G}_t \times \sigma(\chi^i) \times \sigma(\chi^{i'})$, by using that $\mathbf{E}[\cdot|\mathcal{G}_t] = \mathbf{E}[\mathbf{E}[\cdot|\mathcal{G}_t^*]|\mathcal{G}_t]$, we can write the right hand side of (14) as

$$\sum_i \mathbf{E}[\mathbf{E}[(\epsilon_i^k)^2 \tilde{R}_i]|\mathcal{G}_t^*]|\mathcal{G}_t] + \sum_{i\neq i'} \mathbf{E}[\mathbf{E}[\epsilon_i^k \epsilon_{i'}^k \tilde{R}_i \tilde{R}_{i'}]|\mathcal{G}_t^*]|\mathcal{G}_t]. \tag{15}$$

Since $\tilde{R}_i$ and $\tilde{R}_{i'}$ are $\mathcal{G}_t^*-$measurable, (15) can be expressed as

$$\sum_i \mathbf{E}[\tilde{R}_i \mathbf{E}[(\epsilon_i^k)^2]|\mathcal{G}_t^*]|\mathcal{G}_t] + \sum_{i\neq i'} \mathbf{E}[\tilde{R}_i \tilde{R}_{i'} \mathbf{E}[\epsilon_i^k \epsilon_{i'}^k]|\mathcal{G}_t^*]|\mathcal{G}_t],$$

but χ^i and $\chi^{i'}$ are independent from $\mathcal{F}$, so $\mathbf{E}[\epsilon_i^k \epsilon_{i'}^k|\mathcal{G}_t^*] = \mathbf{E}[\epsilon_i^k \epsilon_{i'}^k|\mathcal{G}_t] = \Gamma_{ii'}^k$ and $\mathbf{E}[(\epsilon_i^k)^2|\mathcal{G}_t^*] = \mathbf{E}[(\epsilon_i^k)^2|\mathcal{G}_t] = \Delta_i^k$. Now, $\Gamma_{ii'}^k$ and Δ_i^k are $\mathcal{G}_t-$measurable. Finally, using that $E[\tilde{R}_i|\mathcal{G}] = R_i$ and $E[\tilde{R}_i \tilde{R}_{i'}|\mathcal{G}] = R_i R_{i'}$ we get (13). □

Therefore, $[P_Q]$ is a particular case of $[P]$ with

$$\begin{array}{ll} X = R & \mathcal{G} = \mathcal{G}_t \\ g_1(r) = (r^T - \mathbf{1}^T, -r^T)^T & f(r,\omega) = \mathbf{1}^T r \\ g_2(r,\omega) = (r^T \Gamma^1 r + (\Delta^1)^T r - e_1^2, \ldots, r^T \Gamma^p r + (\Delta^p)^T r - e_p^2)^T, & \end{array} \tag{16}$$

where the dependence of the conditional moments on ω is again implicit. Unfortunately, the matrices Γ^k are indefinite and thus the constraints are not convex. We will overcome this difficulty by using the following lemma.

Lemma 6.2: *Let $\bar{g}_2$ be a function such that $\forall R \in S_I(\mathcal{G}), \mathbf{E}[\bar{g}_2(R(\omega),\omega)] = \mathbf{E}[g_2(R(\omega),\omega)]$ and $\forall R \in S(\mathcal{G}), \mathbf{E}[\bar{g}_2(R(\omega),\omega)] \leq \mathbf{E}[g_2(R(\omega),\omega)]$, and let $[P_Q']$ be the problem obtained from $[P_Q]$ replacing g_2 for $\bar{g}_2$. Then, if R is a solution to $[P_Q']$ and $R \in S_I(\mathcal{G})$, then R is a solution to $[P_Q]$.*

Proof: Let us assume that R is a solution to $[P_Q']$ and it is integer. We know that R satisfies $\mathbf{E}[\bar{g}_2(R(\omega),\omega)] \leq e_k^2$ for $k = 1, \ldots, p$. Then, $\mathbf{E}[g_2(R(\omega),\omega)] \leq e_k^2$, so R satisfies the constraints of $[P_Q]$. Let $S \in S(\mathcal{G}_t)$ such that $\mathbf{E}[g_2(S(\omega),\omega)] \leq e_k^2$. Since $\mathbf{E}[\bar{g}_2(R(\omega),\omega)] \leq \mathbf{E}[g_2(R(\omega),\omega)]$ then $\mathbf{E}[\bar{g}_2(R(\omega),\omega)] \leq e_k^2$ and then, $\mathbf{E}[\mathbf{1}^T S] \leq \mathbf{E}[\mathbf{1}^T S]$. □

We may consider for example the two following possibilities.

(i) $\bar{g}_2(r,\omega) = r^T \Sigma^k(\omega) r$, where $\Sigma_{ij}^k = \mathbf{E}[\epsilon_i^k \epsilon_j^k|\mathcal{G}_t]$

(ii) $\bar{g}_2(r,\omega) = r^T M^k(\omega) r + (v^k(\omega))^T r$, where $M_{ij}^k = m_i^k m_j^k$, $m_i^k = \mathbf{E}[\epsilon_i^k|\mathcal{G}_t]$, $v_i^k = \mathbf{V}[\epsilon_i^k|\mathcal{G}_t]$.

The function of (ii) can be used only under the assumption that $\mathbf{E}[\epsilon_i^k \epsilon_j^k|\mathcal{G}_t] = m_i^k m_j^k$ for $i \neq j$ and this will be the one used in our application (section 8). Lemma 6.2 has practical relevance if we check that the solutions of $[P_Q']$ are integer. We will show that this approximately holds in this application.

Problem $[P'_Q]$ is a case of $[P]$ with

$$\begin{array}{ll} X = R & \mathcal{G} = \mathcal{G}_t \\ g_1(r) = (r^T - \mathbf{1}^T, -r^T)^T & f(r,\omega) = \mathbf{1}^T r \\ \bar{g}_2(r,\omega) = (r^T A^1 r + (b^1)^T r - e_1^2, \ldots, r^T A^p r + (b^p)^T r - e_p^2)^T & \end{array} \quad (17)$$

Where $A^k = \Sigma^k, b^k = 0$ or $A^k = M^k, b^k = v^k$. Since A^k are positive semidefinite, we can state,

Proposition 6.3: *If* $\mathbf{E}|A^k|^\beta, \mathbf{E}|b^k|^\beta < +\infty$ *for all* k *with* $\beta > 1$ *and* $\mathcal{G}_t$ *is countably generated, then the assumptions of proposition 2.8 hold for the case defined by (17).*

Proof: The arguments of proposition 5.1 can be easily adapted to the quadratic case given that the matrices that appear in the definition of g_2 are positive semidefinite. For assumption 2.3 we set α such that $2/\alpha + 1/\beta = 1$. Then, $\mathbf{E}[\bar{g}_2(R(\omega),\omega)]$ is defined. On the other hand,

$$\begin{aligned} |\bar{g}_2^k(r,\omega) - \bar{g}_2^k(s,\omega)| &= |r^T A^k r + (b^k)^T r - s^T A^k s - (b^k)^T s| \leq \\ &\leq |r^T A^k (r-s)| + |s^T A^k (r-s)| + |(b^k)^T (r-s)|. \end{aligned}$$

Then,

$$|\mathbf{E}[\bar{g}_2(R(\omega),\omega)] - \mathbf{E}[\bar{g}_2(S(\omega),\omega)]| \leq \left\{ \left[\|R\|_{L^\alpha} + \|S\|_{L^\alpha} \right] \|A^k\|_{L^\beta} + \|b^k\|_{L^\beta} \right\} \|R - S\|_{L^\alpha}.$$

If $S \to R$ in L^α, the right hand side of the inequality above converges to zero. □

As in the linear case, R_λ is obtained solving a deterministic optimization problem, in this case a quadratic programming problem.

$$[P_D(\lambda,\omega)] \ \max_r \ 1^T r - \sum_k \lambda_k (\bar{g}_2^k(r,\omega) - e_k^2) \quad (18)$$

$$\text{s.t.} \quad r_i \in [0,1]. \quad (19)$$

An important difference with respect to the linear case is that the problem above does not explicitly provide a Score Function generating the SS as when applying the Karush–Kuhn–Tucker conditions in section 5.

We describe in section 7 a practical method to obtain M^k, Σ^k and v^k. $[P_D(\lambda,\omega)]$ was easy to solve in the linear case, but for large sizes (in our case $N > 10,000$), the quadratic programming problem becomes computationally heavy if solved by traditional methods. For $\bar{g}_2$ defined as in (ii), we can take advantage of the low rank of the matrix in the objective function to propose (appendix A) an approximate method to solve it efficiently. In our real data study, we have checked the performance of this method and the results are presented in subsection 8.1.

7. Model-based conditional moments

The practical application of the results in previous sections requires a method to compute the conditional moments of the error with respect to $\mathcal{G}_t$. In this section, we drop the index j to reduce the complexity of the notation, but the results can be adapted to the case of several variables per questionnaire.

Let $\mathcal{H}_t$ be a $\sigma-$field generated by all the information available at time t with the exception of $\tilde{x}_t^i$. Then, $\mathcal{G}_t = \sigma(\tilde{x}_t^i, \mathcal{H}_t)$. Let $\hat{x}_t^i = \tilde{\pi}(x_t^i)$ be a predictor computed using the information in $\mathcal{H}_t$, that is a $\mathcal{H}_t$-measurable random variable optimal in some way decided by the analyst. The prediction error is denoted by $\xi_t^i = \hat{x}_t^i - x_t^i$

We assume that,

Assumption 7.1 ξ_t^i and η_t^i are distributed as a bivariate Gaussian with zero mean, variances ν_i^2 and σ_i^2 and correlation γ_i.

Assumption 7.2 $\varepsilon_t^i = \eta_t^i e_t^i$, where e_t^i is a Bernoulli variable that equals 1 or 0 with probabilities p and $1-p$ and it is independent of ξ_t^i and η_t^i.

Assumption 7.3 ξ_t^i, η_t^i and e_t^i are jointly independent of $\mathcal{H}_t$.

With these assumptions, the conditional moments of the error with respect to $\mathcal{G}_t$ are functions of the sole variable $u_t^i = \hat{x}_t^i - \tilde{x}_t^i$, that is, the difference between the predicted and the observed values. In the next proposition we will also drop i and t in order to simplify notation.

Proposition 7.4: *Under the assumptions 7.1–7.3, it holds*

$$\mathbf{E}[\varepsilon|\mathcal{G}] = \frac{\sigma^2 + \gamma\sigma\nu}{\sigma^2 + \nu^2 + 2\gamma\sigma\nu} u\zeta \tag{20}$$

$$\mathbf{E}[\varepsilon^2|\mathcal{G}] = \left[\frac{\sigma^2\nu^2(1-\gamma^2)}{\sigma^2 + \nu^2 + 2\gamma\sigma\nu} + \left(\frac{\sigma^2 + \gamma\sigma\nu}{\sigma^2 + \nu^2 + 2\gamma\sigma\nu} \right)^2 u^2 \right] \zeta, \tag{21}$$

where,

$$\zeta = \frac{1}{1 + \frac{1-p}{p} \left(\frac{\nu^2}{\sigma^2+\nu^2+2\gamma\sigma\nu} \right)^{-1/2} \exp\{ -\frac{u^2(\sigma^2+2\gamma\sigma\nu)}{2\nu^2(\sigma^2+\nu^2+2\gamma\sigma\nu)} \}}. \tag{22}$$

Proof: First of all, $\mathbf{E}[\varepsilon^\alpha|u] = \mathbf{E}[e\eta^\alpha|u] = \mathbf{E}[\mathbf{E}[e\eta^\alpha|e,u]|u]$. Since e is $\sigma(e,u)-$measurable, $\mathbf{E}[e\eta^\alpha|e,u] = \mathbf{E}[\eta^\alpha|e,u]e$. We can prove that

$$\mathbf{E}[\eta^\alpha|e,u] = \begin{cases} \psi(u) & e = 1 \\ \sigma^2 & e = 0 \end{cases},$$

where ψ is such that $\mathbf{E}[\eta^\alpha|\xi+\eta] = \psi(\xi+\eta)$. Therefore, $\mathbf{E}[\eta^\alpha|e,u]e = \psi(u)e$, and then, $\mathbf{E}[\mathbf{E}[e\eta^\alpha|e,u]|u] = \psi(u)\mathbf{E}[e|u]$.

It remains to compute $\mathbf{E}[e|u]$ and $\psi(u)$ for $\alpha = 1, 2$. For this purpose, we can use the properties of the Gaussian distribution. If (x,y) is a Gaussian-distributed random vector with zero mean and covariance matrix $(\sigma_{ij})_{i,j\in\{x,y\}}$. Then, the conditional distribution $f(y|x)$ is a Gaussian with mean and variance

$$\mathbf{E}[y|x] = \frac{\sigma_{xy}}{\sigma_{xx}} x \qquad \mathbf{V}[y|x] = \sigma_{yy} - \frac{{\sigma_{xy}}^2}{\sigma_{xx}}.$$

Then,

$$\mathbf{E}[y^2|x] = \sigma_{yy} + \frac{{\sigma_{xy}}^2}{\sigma_{xx}} \left(\frac{x^2}{\sigma_{xx}} - 1 \right).$$

Now, we can apply the relations above to $y = \eta$ and $x = u = \eta + \xi$. Then,

$\sigma_{xx} = \sigma^2 + \nu^2 + 2\gamma\sigma\nu$ and $\sigma_{yy} = \sigma^2$, $\sigma_{xy} = \sigma^2 + \gamma\sigma\nu$. Thus, for $\alpha = 1$,

$$\psi(u) = \frac{\sigma^2 + \gamma\sigma\nu}{\sigma^2 + \nu^2 + 2\gamma\sigma\nu} u,$$

and for $\alpha = 2$,

$$\psi(u) = \sigma^2 + \frac{(\sigma^2 + \gamma\sigma\nu)^2}{\sigma^2 + \nu^2 + 2\gamma\sigma\nu} \left(\frac{u^2}{\sigma^2 + \nu^2 + 2\gamma\sigma\nu} - 1 \right).$$

Let us now compute $\zeta = \mathbf{E}[e|u] = P[e = 1|u]$. By an argument similar to Bayes's theorem it can be proved that ζ is equal to $P[e = 1]f(u|e = 1)/f(u)$, where $f(u|e = 1)$ is a zero-mean Gaussian density function with variance $\sigma^2 + \nu^2 + 2\gamma\sigma\nu$ and $f(u)$ is a mixture of two Gaussians with variances $s_1^2 = \sigma^2 + \nu^2 + 2\gamma\sigma\nu$ and $s_2^2 = \nu^2$ and probabilities p and $1 - p$ respectively. Hence,

$$\zeta = p \frac{(2\pi s_1^2)^{-1/2} \exp\{-\frac{u^2}{2s_1^2}\}}{p(2\pi s_1^2)^{-1/2} \exp\{-\frac{u^2}{2s_1^2}\} + (1-p)(2\pi s_2^2)^{-1/2} \exp\{-\frac{u^2}{2s_2}\}}.$$

After simplifying it yields,

$$\zeta = \frac{1}{1 + \frac{1-p}{p} \left(\frac{\nu^2}{\sigma^2+\nu^2+2\gamma\sigma\nu} \right)^{-1/2} \exp\{-\frac{u^2(\sigma^2+2\gamma\sigma\nu)}{2\nu^2(\sigma^2+\nu^2+2\gamma\sigma\nu)}\}}.$$

Finally, since ξ, η and e are independent of $\mathcal{H}_t$, we can conclude by noting that $\mathbf{E}[\varepsilon^2|u] = \mathbf{E}[\varepsilon^2|u, \mathcal{H}_t] = \mathbf{E}[\varepsilon^2|\mathcal{G}_t]$. □

8. Case Study

In this section, we present the results of the application of the methods described in this paper to the data of the Turnover/New Orders Survey. Monthly data from about $N = 13,500$ units are collected. In the moment of the study, data from January 2002 to September 2006 were available (t ranges from 1 to 57). Only two of the variables requested in the questionnaires are considered in our study, namely, Total Turnover and Total New Orders ($q = 2$). The total Turnover of unit j at period t is x_t^{i1} and Total New Orders is x_t^{i2}. These two variables are aggregated separately to obtain the two indicators, so $p = 2$ and $\omega_{i2}^1 = \omega_{i1}^2 = 0$.

We need a model for the data in order to apply proposition 7.4 and obtain the conditional moments. Since the variables are distributed in a strongly asymmetric way, we use their logarithm transform, $y_t^{ij} = \log(x_t^{ij} + m)$, where m is a positive constant adjusted by maximum likelihood ($m \approx 10^5$€). The conditional moments of the original variable can be recovered exactly by using the properties of the log-normal distribution or approximately by using a first-order Taylor expansion, yielding $\mathbf{E}[(\tilde{x}_t^{ij} - x_t^{ij})^2|\mathcal{G}_t] \approx (\tilde{x}_t^{ij} - m)^2 \mathbf{E}[(\tilde{y}_t^{ij} - y_t^{ij})^2|\mathcal{G}_t]$. In our study, we used the approximate version. We found that if $\tilde{x}_t^{ij} - m$ is replaced by an average of the last 12 values of $\tilde{x}_t^{ij}$, the estimate becomes more robust against very small values of $\tilde{x}_t^{ij} - m$.

The model applied to the transformed variables is very simple. We assume that the variables x_t^{ij} are independent across (i, j) and for any pair (i, j), we choose

λ	mean $1^T r$	mean $1^T \lvert r - r_{app} \rvert$
10^2	592.8	0.16
10^4	587.6	1.12
10^6	555.3	1.13
10^8	386.3	0.51
10^9	235.8	0.25
10^{10}	108.7	0.08
10^{11}	44.0	0.07
10^{12}	23.8	0.08

Table 2. Comparison between the exact (r) and approximate (r_{app}) quadratic methods.

among the following simple models.

$$(1 - B)y_t^{ij} = a_t, \tag{23}$$

$$(1 - B^{12})y_t^{ij} = a_t, \tag{24}$$

$$(1 - B^{12})(1 - B)y_t^{ij} = a_t. \tag{25}$$

where B is the backshift operator $Bu_t = u_{t-1}$ and a_t are white noise processes. We obtain the residuals $\hat{a}_t$ and then select the model which produces lesser mean of squared residuals, $\sum \hat{a}_t^2/(T - r)$, where r is the maximum lag in the model. With this model, we compute the prediction $\hat{y}_t^{ij}$ and the prediction standard deviation ν_{ij}. The *a priori* standard deviation of the observation errors and the error probability are considered constant across units (that is possible because of the logarithm transformation). We denote them by σ_j and p_j with $j = 1, 2$ and they are estimated using historical data of the survey.

A database is maintained with the original collected data and subsequent versions after possible corrections due to the editing work. Thus, we consider the first version of the data as *observed* and the last one as *true*. The coefficient γ_i is assumed zero. Once we have computed σ_j, p_j, ν_{ij} and u_t^{ij}, proposition 7.4 can be used to obtain the conditional moments and then, Δ^k, Σ^k and v^k.

8.1. *Accuracy of the Approximate Method to the Quadratic Problem.*

Before assessing the efficiency of the selection, we have used the data to check that the approximate method to solve the quadratic problem does not produce a significant disturbance of the solutions. We have compared the approximate solutions to the ones obtained by a usual quadratic programming approach. For this purpose, we have used the function *quadprog* of the mathematical pack MATLAB. For the whole sample, *quadprog* does not converge in a reasonable time —that is the reason why the approximate method is required— so we have extracted for comparison a random subsample of 5% (roughly over 600 units) and we have solved the problem $[P_Q]$ for a range of values of λ with the exact and approximate methods. In table 2, we present for the different values of λ, the average of the number of units edited using the exact method and a measure of the difference between the two methods. We also have used this data to check the validity of the assumption that the solutions (of the exact method) are integer. For this purpose, we computed $\sum_i \min\{r_i, 1 - r_i\}$, whose value never exceeded 1, while the number of units for which $\min\{r_i, 1 - r_i\} > 10^{-3}$ was at most 2. Therefore, the solution can be considered as approximately integer.

8.2. *Expectation Constraints*

We will now check that the expectation constraints in $[P_L]$ and $[P_Q]$ are effectively satisfied. In order to do this, for $l = 1, \ldots, b$ with $b = 20$ we solve the optimization problem with the variance bounds $e_{1l}^2 = e_{2l}^2 = e_l^2 = [s_0^{((l-1)/(b-1))} s_1^{((b-l)/(b-1))}]^2$. The range of standard deviations goes from $s_0 = 0.025$ to $s_1 = 1$.

The expectation of the dual function is estimated using a $h-$length batch of real data. For any period from October 2005 to September 2006 and for any $l = 1, \ldots, b$, a selection $r(t, l)$ is obtained according the bound e_k^2. The average across t of the remaining squared errors is thus computed as

$$\hat{e}_{kl}^2 = \frac{1}{12} \sum_{t=t_0}^{t_0+11} r(t,l)^T E^k r(t,l).$$

We repeated these calculations for $h = 1, 3, 6$ and 12 both using the linear and the quadratic versions. The results are arranged in tables B1 to B8. For each l we present the average number of units edited, the desired bound and for $k = 1, 2$, the quotient $\hat{e}_{kl}/e_{kl}$. In every case, there is a tendency to underestimate the error when the constraints are smaller and to overestimate it when the bounds are larger. The quadratic method produces better results with respect to the bounds but at the price of editing more units.

8.3. *Comparison of Score Functions*

We intend to compare the performance of our method to that of the score-function described in [7], $\delta_i^0 = \omega_i |\tilde{x}^i - \hat{x}^i|$, where $\hat{x}_i$ is a prediction of x according to a certain criterion. The author proposes to use the last value of the same variable in previous periods. We have also considered the score function δ^1 defined as δ^0 but using the forecasts obtained through the models in (23)–(25). Finally, δ^2 is the score function computed using (23)–(25) and proposition 7.4. The global SF is just the sum of the two local ones. We will measure the effectiveness of the score functions by $E_l^j = \sum_n E_l^j(n)$, with

$$E_1^j(n) = \sum_{i \geq n}^{N} (\omega_i^j)^2 (\tilde{x}^{ij} - x^{ij})^2 \qquad E_2^j(n) = \Big[\sum_{i \geq n}^{N} \omega_i^j (\tilde{x}^{ij} - x^{ij}) \Big]^2,$$

where we consider units arranged in descending order according to the corresponding score function. These measures can be interpreted as estimates of the remaining error after editing the n first units. The difference is that $E_1^j(n)$ is the aggregate squared error and $E_2^j(n)$ is the squared aggregate error. Thus, $E_2^j(n)$ is the one that has practical relevance, but we also include the values of $E_2^j(n)$ because in the linear problem $[P_L]$, it is the aggregate squared error which appears in the left side of the expectation constraints. In principle, it could happen that our score function was optimal for the $E_1^j(n)$ but not for $E_2^j(n)$. Nevertheless, the results in table 3 show that δ^2 is better measured both ways.

9. Conclusions.

We have described a theoretical framework to deal with the problem of selective editing, defining the concept of selection strategy. We describe the search for an

	Turnover		Orders	
	E_1	E_2	E_1	E_2
δ^0	0.43	0.44	1.16	1.33
δ^1	0.30	0.38	0.36	0.45
δ^2	0.21	0.26	0.28	0.37

Table 3. Comparison of score functions.

adequate selection strategy as an optimization problem. This problem is a linear optimization problem with quadratic constraints. We show that the score function approach is the solution to the problem with linear constraints. We also show how to solve the quadratic problem.

The score function obtained outperforms a reference SF. Both the linear and the quadratic versions of our method produce selection strategies that satisfy approximately the constraints but for small values of the constraint. The quadratic method seems to be more conservative and then, the bounds are better fulfilled, but more units are edited. On the other hand, the implementation of the linear method is easier and computationally less demanding. This suggests that the quadratic method is more adequate for cases in which the bounds are critical and the linear one for cases in which timeliness is critical.

Acknowledgements

The authors wish thank José Luis Fernández Serrano and Emilio Cerdá for their help and advice.

References

[1] J. Berthelot and M. Latouche, *Improving the efficiency of data collection: A generic respondent follow-up strategy for economic surveys*, J. Bus. Econom. Statist. 11 (1993), pp. 417–424.
[2] D. Bertsimas and M. Sim, *The price of rebustness*, Oper. Res. 52 (2004), pp. 35–53.
[3] P. Billingsley, *Probability and measure*, John Wiley and Sons, New York, 1995.
[4] A.M. Croicu and M.Y. Hussaini, *On the expected optimal value and the optimal expected value*, Appl. Math. Comput. 180 (2006), pp. 330–341.
[5] J. Dupačová and R.J.B. Wets, *Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems*, AOS. 16 (1988), pp. 1517–1549.
[6] L. Granquist, *The new view on editing*, International Statistics Review 65 (1997), pp. 381–387.
[7] D. Hedlin, *Score functions to reduce business survey editing at the U.K. office for national statistics*, Journal of Official Statistics 19 (2003), pp. 177–199.
[8] M. Latouche and J. Berthelot, *Use of a score function to prioritize and limit recontacts in editing business surveys*, Journal of Official Statistics 8 (1992), pp. 389–400.
[9] D. Lawrence and R. McKenzie, *The general application of significance editing*, Journal of Official Statistics 16 (2000), pp. 243–253.
[10] D.G. Luenberger, *Optimization by vector space methods*, John Wiley and Sons, New York, 1969.
[11] S.M. Robinson, *Analysis of sample-path optimization*, Math. Oper. Res. 21 (1996), pp. 513–528.
[12] R.T. Rockafellar and R.J.B. Wets, *Variational analysis*, Springer, New York, 1998.
[13] W. Rudin, *Real and complex analysis*, 3rd ed., McGraw-Hill, New York, 1987.
[14] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on stochastic programming: Modelling and theory*, SIAM, Philadelphia, 2009.
[15] A. Shapiro and T. Homem-de-Mello, *On the rate of convergence of optimal solutions of monte carlo approximations of stochastics programs*, SIAM J. Optim. 11 (2000), pp. 70–86.

Appendix A. Approximate method for the quadratic problem

The Karush–Kuhn–Tucker conditions applied to the problem (18)–(19) with $\bar{g}_2^k(r) = r^T M^k r + (v^k)^T r - e_k^2$ imply that in the optimum, it holds

$$2\sum_k \lambda_k \left(m_i^k (m^k)^T r + v_i^k\right) - 1 = \mu_i^+ - \mu_i^- \;\; \mu_i^+, \mu_i^- \geq 0$$

$$\mu_i^+ (1 - r_i) = 0 \;\; \mu_i^- r_i = 0,$$

where $m^k = (m_1^k, \ldots, m_N^k)^T$. The relations above hold when

$$\begin{array}{l} 2\sum_k \lambda_k \left(m_i^k (m^k)^T r + v_i^k\right) - 1 > 0 \text{ if } r_i = 1 \\ 2\sum_k \lambda_k \left(m_i^k (m^k)^T r + v_i^k\right) - 1 < 0 \text{ if } r_i = 0 \end{array}$$

Let us assume that we know $\alpha = [(m^1)^T r, \ldots, (m^p)^T r]^T = Mr$, where $M = (m^1, \ldots, m^p)^T$. Then, we can built $r(\alpha)$ as

$$r_i = \begin{cases} 1 \text{ if } 2\sum_k \lambda_k \left(m_i^k \alpha_k + v_i^k\right) > 1 \\ 0 \text{ if } 2\sum_k \lambda_k \left(m_i^k \alpha_k + v_i^k\right) < 1 \end{cases}$$

If $\alpha = Mr(\alpha)$, then $r(\alpha)$ is a solution. We can solve approximately the fixed-point problem by minimizing $\|\alpha - Mr(\alpha)\|^2$. In our applications p is typically small, so the dimension of the optimization problem has been strongly reduced.

Appendix B. Tables

Table B1. Error bounds of the linear version (h=1).

e_l	$\hat{e}_{1l}/e_l$	$\hat{e}_{2l}/e_l$	n	e_l	$\hat{e}_{2l}/e_l$	$\hat{e}_{2l}/e_l$	n
0.0250	2.04	3.15	397.0	0.1742	0.85	1.59	29.9
0.0304	1.87	2.72	312.2	0.2116	0.75	1.29	21.1
0.0369	1.61	2.28	245.8	0.2569	0.65	1.05	14.1
0.0448	1.46	1.99	194.8	0.3120	0.61	0.86	9.5
0.0544	1.08	1.54	153.6	0.3788	0.56	0.73	6.1
0.0660	0.90	1.38	119.6	0.4600	0.56	0.69	3.8
0.0801	0.79	1.19	92.5	0.5585	0.61	0.60	2.1
0.0973	0.72	1.02	71.9	0.6782	0.47	0.51	1.3
0.1182	0.74	1.22	54.5	0.8235	0.39	0.42	0.8
0.1435	0.88	1.63	40.9	1.0000	0.32	0.34	0.8

Table B2. Error bounds of the linear version (h=3).

e_l	$\hat{e}_{1l}/e_l$	$\hat{e}_{2l}/e_l$	n	e_l	$\hat{e}_{2l}/e_l$	$\hat{e}_{2l}/e_l$	n
0.0250	2.43	3.74	427.9	0.1742	1.26	1.29	33.7
0.0304	2.45	3.35	338.4	0.2116	1.18	1.18	24.3
0.0369	2.49	3.09	266.7	0.2569	1.00	1.00	17.5
0.0448	2.36	2.39	208.8	0.3120	0.87	0.81	12.0
0.0544	1.86	2.39	165.2	0.3788	0.88	0.70	8.1
0.0660	1.66	1.99	132.6	0.4600	0.79	0.59	5.4
0.0801	1.60	1.68	102.2	0.5585	0.67	0.51	3.1
0.0973	1.50	1.46	79.0	0.6782	0.56	0.44	1.8
0.1182	1.50	1.38	60.9	0.8235	0.43	0.36	1.1
0.1435	1.41	1.24	46.0	1.0000	0.35	0.31	0.7

Table B3. Error bounds of the linear version (h=6).

e_l	$\hat{e}_{1l}/e_l$	$\hat{e}_{2l}/e_l$	n	e_l	$\hat{e}_{2l}/e_l$	$\hat{e}_{2l}/e_l$	n
0.0250	1.89	3.20	414.8	0.1742	0.97	1.31	30.5
0.0304	1.88	2.63	327.8	0.2116	0.82	1.29	20.7
0.0369	2.04	2.43	257.6	0.2569	0.71	1.09	15.0
0.0448	1.83	2.05	202.0	0.3120	0.80	0.85	10.3
0.0544	1.58	1.87	157.7	0.3788	0.74	0.77	6.8
0.0660	1.11	1.59	125.7	0.4600	0.71	0.72	4.7
0.0801	1.10	1.25	95.9	0.5585	0.57	0.57	2.7
0.0973	0.96	1.01	73.2	0.6782	0.47	0.46	1.7
0.1182	0.96	1.18	56.2	0.8235	0.42	0.38	1.2
0.1435	0.96	1.12	41.8	1.0000	0.30	0.31	0.5

Table B4. Error bounds of the linear version (h=12).

e_l	$\hat{e}_{1l}/e_l$	$\hat{e}_{2l}/e_l$	n	e_l	$\hat{e}_{2l}/e_l$	$\hat{e}_{2l}/e_l$	n
0.0250	1.56	3.38	430.4	0.1742	0.76	0.86	31.3
0.0304	1.08	2.63	340.7	0.2116	0.64	1.09	20.0
0.0369	1.02	1.77	268.4	0.2569	0.97	0.91	14.0
0.0448	1.00	1.30	207.9	0.3120	0.79	0.65	7.6
0.0544	0.79	1.26	161.0	0.3788	0.78	0.64	4.4
0.0660	0.72	0.98	129.4	0.4600	0.78	0.65	2.7
0.0801	0.84	0.78	96.4	0.5585	0.64	0.48	1.3
0.0973	1.15	0.60	76.1	0.6782	0.53	0.40	0.9
0.1182	0.96	0.64	57.4	0.8235	0.43	0.33	0.7
0.1435	0.83	0.68	41.1	1.0000	0.36	0.27	0.6

Table B5. Error bounds of the quadratic version (h=1).

e_l	$\hat{e}_{1l}/e_l$	$\hat{e}_{2l}/e_l$	n	e_l	$\hat{e}_{2l}/e_l$	$\hat{e}_{2l}/e_l$	n
0.0250	2.41	4.12	507.6	0.1742	0.99	0.82	265.0
0.0304	1.40	3.01	655.8	0.2116	1.95	0.86	134.8
0.0369	1.60	2.57	540.8	0.2569	0.98	0.86	54.3
0.0448	1.32	2.12	541.3	0.3120	1.35	0.73	36.5
0.0544	0.80	1.71	593.5	0.3788	0.68	0.63	29.2
0.0660	1.06	1.67	469.8	0.4600	0.55	0.52	20.8
0.0801	0.81	1.44	460.6	0.5585	0.46	0.73	19.8
0.0973	0.98	1.21	275.7	0.6782	0.64	0.59	5.9
0.1182	1.19	1.05	150.8	0.8235	0.53	0.52	5.4
0.1435	1.21	0.99	272.6	1.0000	0.44	0.41	1.9

Table B6. Error bounds of the quadratic version (h=3).

e_l	$\hat{e}_{1l}/e_l$	$\hat{e}_{2l}/e_l$	n	e_l	$\hat{e}_{2l}/e_l$	$\hat{e}_{2l}/e_l$	n
0.0250	1.58	2.41	650.3	0.1742	1.24	0.90	85.4
0.0304	1.44	1.63	563.1	0.2116	1.14	0.86	42.3
0.0369	1.27	1.81	589.9	0.2569	0.94	0.73	38.1
0.0448	0.96	1.30	507.3	0.3120	0.80	0.85	26.3
0.0544	0.98	1.66	388.5	0.3788	0.61	0.70	20.6
0.0660	1.45	1.49	298.8	0.4600	0.53	0.49	23.9
0.0801	1.65	1.61	231.4	0.5585	0.59	0.60	15.1
0.0973	1.35	0.82	171.2	0.6782	0.47	0.46	4.3
0.1182	1.03	0.62	152.4	0.8235	0.31	0.37	3.6
0.1435	1.25	0.91	101.5	1.0000	0.31	0.32	2.7

Table B7. Error bounds of the quadratic version (h=6).

e_l	$\hat{e}_{1l}/e_l$	$\hat{e}_{2l}/e_l$	n	e_l	$\hat{e}_{2l}/e_l$	$\hat{e}_{2l}/e_l$	n
0.0250	1.86	2.50	578.4	0.1742	1.24	0.92	77.8
0.0304	1.46	2.02	605.9	0.2116	0.99	0.76	59.8
0.0369	1.42	1.82	401.9	0.2569	0.92	0.67	35.4
0.0448	0.95	1.35	477.3	0.3120	0.71	0.56	27.7
0.0544	0.96	1.17	390.5	0.3788	0.64	0.66	28.0
0.0660	1.41	1.28	278.8	0.4600	0.71	0.59	13.3
0.0801	1.31	1.30	283.4	0.5585	0.56	0.57	6.8
0.0973	0.99	0.84	225.8	0.6782	0.46	0.47	6.8
0.1182	1.32	1.22	140.3	0.8235	0.39	0.38	2.9
0.1435	1.30	0.95	93.1	1.0000	0.32	0.46	1.6

Table B8. Error bounds of the quadratic version (h=12).

e_l	$\hat{e}_{1l}/e_l$	$\hat{e}_{2l}/e_l$	n	e_l	$\hat{e}_{2l}/e_l$	$\hat{e}_{2l}/e_l$	n
0.0250	1.38	1.91	703.4	0.1742	1.19	0.95	83.3
0.0304	1.13	1.77	620.5	0.2116	0.99	0.71	69.3
0.0369	0.93	1.78	595.8	0.2569	0.82	0.57	50.8
0.0448	0.83	1.54	515.7	0.3120	0.73	0.51	46.8
0.0544	0.66	1.38	456.6	0.3788	0.59	0.45	27.8
0.0660	0.84	1.27	422.5	0.4600	0.55	0.51	18.0
0.0801	1.09	1.40	284.6	0.5585	0.58	0.61	10.8
0.0973	1.21	1.05	223.3	0.6782	0.48	0.48	4.5
0.1182	1.51	1.16	120.3	0.8235	0.39	0.40	1.3
0.1435	1.19	0.81	86.1	1.0000	0.32	0.32	1.2

VERSITA

Journal of Official Statistics, Vol. 29, No. 4, 2013, pp. 1–23, http://dx.doi.org/10.2478/jos-2013-0037

An Optimization Approach to Selective Editing

Ignacio Arbués[1], *Pedro Revilla*[1], *and David Salgado*[1]

We set out two generic principles for selective editing, namely the minimization of interactive editing resources and data quality assurance. These principles are translated into a generic optimization problem with two versions. On the one hand, if no cross-sectional information is used in the selection of units, we derive a stochastic optimization problem. On the other hand, if that information is used, we arrive at a combinatorial optimization problem. These problems are substantiated by constructing a so-called observation-prediction model, that is, a multivariate statistical model for the nonsampling measurement errors assisted by an auxiliary model to make predictions. The restrictions of these problems basically set upper bounds upon the modelled measurement errors entering the survey estimators. The bounds are chosen by subject-matter knowledge. Furthermore, we propose a selection efficiency measure to assess any selective editing technique and make a comparison between this approach and some score functions. Special attention is paid to the relationship of this approach with the editing fieldwork conditions, arising issues such as the selection versus the prioritization of units and the connection between the selective and macro editing techniques. This approach neatly links the selection and prioritization of sampling units for editing (micro approach) with considerations upon the survey estimators themselves (macro approach).

Key words: Selective editing; optimization; observation-prediction model; selection efficiency measure.

1. Introduction

Data editing is a crucial step in the survey statistics production process. It impinges on several dimensions of survey quality such as accuracy, timeliness, response burden or cost effectiveness. This production phase comprises both the detection and treatment of nonsampling errors, mainly of nonresponse and measurement errors. Over time, a typology of errors has been developed, identifying systematic errors, random errors, influential errors, outliers, inliers or missing values, not to mention particular errors within these classes as measurement unit errors or rounding errors. This diversity has given rise to different techniques and algorithms to detect and treat them, such as interactive editing, automatic editing, selective editing, macro editing, and so on (see de Waal et al. 2011 for a comprehensive overview). Nowadays it is widely accepted that no single technique can

[1] D.G. Metodología, Calidad y TIC, Instituto Nacional de Estadística, Paseo de la Castellana, 28071 Madrid, Spain. Emails: ignacio.arbues.lombardia@ine.es, pedro.revilla.novella@ine.es, david.salgado.fernandez@ine.es

Acknowledgments: We acknowledge graphics computing support from S. Saldaña. We are indebted to C. Pérez-Arriero and M. Herrador for their invaluable suggestions regarding the selection efficiency measure. The authors are grateful to T. de Waal, M. Di Zio, U. Guarnera, J. Pannekoek, M. van der Loo and S. Scholtus for comments and suggestions. We express special thanks to L.-C. Zhang for invaluable suggestions to improve the readability of the article.

deal with all kinds of errors. Thus they must be conveniently combined in so-called editing and imputation (E&I henceforth) strategies, specifically designed and fine-tuned for a given survey.

Selective editing focuses upon influential errors so that a selection of influential units is performed to thoroughly treat their errors (mostly with interactive editing), underlining the importance of recognizing and analyzing their source in order to prevent them when the survey is conducted on future occasions (Granquist 1997). In the last two decades, this editing modality has been recognized as a key element in E&I strategies. However, its principles are heuristics. By and large, selective editing comprises four stages (Lawrence and McKenzie 2000), namely (i) the construction of anticipated values $\hat{y}_k$ for each sample unit k according to an editing model; (ii) the construction of local score functions; (iii) the construction of a global score function; and (iv) the choice of cut-off values below which no further unit is selected. In general terms, the rationale is that those questionnaires k with a large discrepancy between the anticipated values $\hat{y}_k$ and the reported values y_k will be selected.

As a first general remark, our proposal can be succinctly described using the recent taxonomy of data editing functions by Pannekoek et al. (2013). They identify six types of editing tasks, called editing functions, according to the accomplishment of either error detection only (as data quality verification or field/record selection) or including error treatment. These six editing functions are (i) rule checking, (ii) compute scores, (iii) field selection, (iv) record selection, (v) amend observations, and (vi) amend unit properties (see Pannekoek et al. 2013 for details). In this context, our proposal is to be understood as a record selection editing function.

We set out two general principles to approach selective editing (Arbués et al. 2012b). In keeping with Latouche and Berthelot (1992), who stated that "*in the development of an effective recontact and follow-up strategy, we have to minimize the amount of resources used without affecting the overall data quality and timeliness of the survey*", we claim that

i) editing must minimize the amount of resources deployed to recontacts, follow-ups and interactive tasks, in general;
ii) data quality must be ensured.

This framework is ample enough to give room to the preceding score function approach, but its rigorous derivation to us seems difficult. In this article we propose a mathematical translation of these principles into a general optimization problem, whose solution is the selection of units. In our formulation, interactive editing resources are tantamount to the number of selected questionnaires, whereas data quality is reduced to the accuracy of estimators. Thus a general optimization approach is to minimize the number of selected units subjected to bounds on loss functions defined for a chosen number of variables of interest. These loss functions may be targeted at the bias, mean squared error (MSE), variance or other measures of the estimation uncertainty. They may be heuristic in nature, such as the so-called pseudo-bias related measures traditionally used for score functions, or they may be explicitly derived under some measurement-error models that are suitable for the data. One example is the contamination model (Di Zio and Guarnera 2013), which is specified in terms of the full distribution of the true data and the conditional distribution of the observations given the true data.

Two versions of the optimization problem are provided, corresponding to the two typical scenarios for the implementation of selective editing. In the first case, selection is carried out unit by unit, such that whether a given unit is selected or not does not depend on the selection of the other units. This mode of execution is suitable for input editing, where in principle the selection can be made in real time on arrival of each questionnaire. We refer to this as the stochastic optimization problem, because the real-time performance of the solution can only be established with respect to hypothetical repetitions of the selection process. In the second case, selection is carried out jointly for all (or a group of) units. This mode of execution is suitable for output (or macro) editing, which takes place at a later stage of the data collection after a sufficient number of observations have become available. We refer to this as the combinatorial optimization problem, where the performance of the solution can be established conditional on the actual sample observations under some specified measurement-error model.

Selection of units does not produce an order of priority by which the units are sorted according to their respective "urgency" to be edited. But prioritization of units is helpful for coping with the contingency of editing fieldwork. It is intrinsically related to selection since it should be possible in some sense to regard the highest prioritized unit as the optimal selection of a single unit, the second highest prioritized unit as the optimal selection of a single unit given that the highest prioritized unit has been selected, and so on. The combinatorial optimization problem can be adapted to yield prioritization. Not only is this a useful variation for practice, but sometimes it is theoretically necessary for obtaining a unique optimization solution, as we shall explain.

To perform a comparison with any other selective editing technique, we propose a selection efficiency measure. The rationale of this measure is to choose as an input the number of units to select and to compare our selection with an averaged random selection of this number of units. The comparison is based on the reduction of the absolute relative pseudo-bias of the survey estimators. In our view, the sooner the influential units are selected (hence the faster the reduction of the absolute relative pseudo-bias), the more efficient the technique will be. We perform a comparison with some score functions in the literature (Latouche and Berthelot 1992) using real data from the Spanish Industrial Turnover Index (ITI) and Industrial New Orders Received Index (INORI) survey.

The article is organized as follows. In section 2 we formulate the generic optimization problem as a mathematical translation of the above two principles. After fixing the notation and setting out the problem in general terms in Subsection 2.1, we show how the choice of the actual information used in this problem drives us either to a stochastic optimization version (Subsection 2.2) or to a combinatorial optimization version (Subsection 2.3). In Section 3 we show the general principles of the construct of any observation-prediction model, as well as a general proposal for continuous variables. In Section 4 we deal with the editing fieldwork and show how to choose the bounds and how to go from the selection to the prioritization of units under the combinatorial optimization approach. In Section 5 a selection efficiency measure is proposed and a comparison with several score functions is carried out using real data from the Spanish ITI and INORI survey. Finally we include an ample discussion in Section 6 in an attempt to assess this proposal in the current framework of selective editing with score functions.

2. The Optimization Problem

Before identifying the variables, the objective function and the restrictions of our optimization problem, we need to introduce the following notation. The sampling design according to which a probability sample s is selected will be denoted by $p(\cdot)$. The sample size will be denoted by n and the corresponding sampling weights by w_{ks}. The sample dependence of the sampling weights implicitly assumes that they do not need to be the design weights. For example, in a ratio estimator of the form $\hat{Y}^{rat} = X \cdot \frac{\hat{Y}^{HT}}{\hat{X}^{HT}}$, where x is a known auxiliary variable from the sampling frame, $X = \sum_{k \in U} x_k$ is a known population total, and $\hat{Y}^{HT} = \sum_{k \in s} \frac{y_k}{\pi_k}$ (analogously for $\hat{X}^{HT}$) stands for the Horvitz-Thompson estimator of the population total $Y = \sum_{k \in U} y_k$, the sampling weights are given by $w_{ks} = \frac{X}{\hat{X}^{HT}} \frac{1}{\pi_k}$, where π_k is the first-order inclusion probability for unit k. More complex situations are embedded under this notation. The true, observed and edited values of a variable $y^{(q)}$, $q = 1, \ldots, Q$ (for ease of notation we drop the superscript (q) hereafter except when strictly necessary), for unit k will be denoted, respectively, by y_k^0, y_k and y_k^*. We assign a binary variable $r_k \in \{0, 1\}$ to each unit k to indicate whether it is selected ($r_k = 0$) or not ($r_k = 1$). The vector $\mathbf{r} = (r_1, \ldots, r_n)^t$ for the whole sample will be referred to as the *selection strategy*. The counterintuitive assignment allows us to relate the preceding three values by the equation $y_k^*(\mathbf{r}) = (1 - r_k) \cdot y_k^0 + r_k \cdot y_k$, where we have made explicit the dependence of the edited values upon the selection strategy. Note that we are implicitly assuming that the editing work drives us from the observed to the true values. If we denote the corresponding measurement error by $\boldsymbol{\epsilon}_k = y_k - y_k^0$, then we can write $y_k^*(\mathbf{r}) = y_k^0 + r_k \boldsymbol{\epsilon}_k$. Note that these edited values are in fact those to be plugged into the survey estimators at this point of the E&I strategy. That is, if we are to estimate the population domain total $Y_{U_d} = \sum_{k \in U_d} y_k^0$ (for ease of notation we will drop the subscript U_d hereafter), then we denote the corresponding chosen estimator by $\hat{Y}^*(\mathbf{r}) = \sum_{k \in s_d} w_{ks} y_k^*(\mathbf{r})$. However, note that this estimator will not be the final estimator after the whole E&I strategy has been executed. Some later procedures such as weight adjustment or outlier treatment may follow. The selection of units proposed herein divides the sample into a critical and a noncritical stream, the treatments of which are decided by the statistician. We will restrict ourselves to population totals and linear estimators. All auxiliary covariates not included in the questionnaire for unit k will be denoted by $\mathbf{x}_k$.

So far the preceding variables are numeric. To use statistical modelling techniques, we promote these numeric variables to random variables according to a model m in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. As usual, this promotion will not be specifically indicated in the notation, except for the selection strategy, so that $\mathbf{R}$ will denote the random selection strategy, and $\mathbf{R}(w) = \mathbf{r}$, with $w \in \Omega$, will be a particular numeric realization called the *selection*. A predicted value of variable y_k according to the chosen model m will be denoted by $\hat{y}_k$. Note that the statistical model m embraces all promoted random variables different from the probability sample s itself. When random variables are used in survey estimators, we write indistinctly $\hat{Y}^0 = \sum_{k \in s_d} w_{ks} y_k^0$, $\hat{Y} = \sum_{k \in s_d} w_{ks} y_k$ and $\hat{Y}^*(\mathbf{R}) = \sum_{k \in s_d} w_{ks} y_k^*(\mathbf{R})$ for the survey estimators targeted at Y. We will denote by $\mathbf{Z}$ the set of random variables actually used by the statistician to select the units in the E&I strategy.

In particular, we will consider two options, namely, either $\mathbf{Z} = \mathbf{Z}^{long} \equiv s$ or $\mathbf{Z}^{long} \equiv \{s, \mathbf{X}\}$ for the stochastic problem (see below for the difference) or $\mathbf{Z} = \mathbf{Z}^{cross} \equiv \{s, \mathbf{X}, \mathbf{Y}\}$

for the combinatorial version. When this cross-sectional information is restricted to unit k, we shall write accordingly $\mathbf{Z}_k^{cross} = \{s, \mathrm{x}, y_k\}$. The use of information is represented as conditioning upon the corresponding random variables. The auxiliary covariates $\mathbf{X}$ are chosen by the statistician according to the chosen statistical model to be used in the problem (see below). They play a similar role to the auxiliary variables in the sampling design or the known auxiliary variables in the weight calibrating process. Indeed, they may coincide partially or totally with these auxiliary variables used in other parts of the estimation process.

2.1. The General Optimization Problem

As stated in the introduction, we want to minimize the number of questionnaires to edit provided that the chosen loss functions of the survey estimators $\hat{Y}^*$ targeted at the population total Y are bounded. To formally set up the optimization problem we need (i) the variables, (ii) the function to optimize, and (iii) the restrictions. Apart from identifying these elements, it is important to show how the available information enters into the formulation of the problem.

The ultimate variables are the selection strategy $\mathbf{r}^T = (r_i, \ldots, r_n)$ for the sample units $s = \{1, \ldots, n\}$, where $r_k = 0$ if the unit k is selected and $r_k = 1$ otherwise. However, since the measurement error $\boldsymbol{\epsilon}_k = y_k - y_k^0$ is conceived to be random in nature conditional on the realized sample s, and given the available information $\mathbf{Z}$ chosen to make the selection of units, this selection can vary depending on the realized $\mathbf{y}$, $\mathbf{y}^0$ and $\mathbf{Z}$. Thus let $\mathbf{R}$ denote the stochastic selection strategy so that (i) $\mathbf{R}(w) = \mathbf{r}$ is a realized selection and (ii) $\mathbb{E}_m[\mathbf{R}|\mathbf{Z}]$ is the vector of probabilities of nonselection under the specific model m given the chosen information $\mathbf{Z}$. The objective function to optimize, given the information $\mathbf{Z}$, is then written as $\mathbb{E}_m[\mathbb{1}^T\mathbf{R}|\mathbf{Z}]$, whose maximization amounts to minimizing the number of selected units.

The constraints derive from the application of a loss function to the survey estimators. Let us concentrate on the two loss functions most used in practice, namely the absolute loss $L = L^{(1)}(a, b) = |a - b|$ or the squared loss $L = L^{(2)}(a, b) = (a - b)^2$. Then it is straightforward to prove (see appendix A) that $\mathbb{E}_m[L^{(r)}(\hat{Y}^*(\mathbf{R}), Y)|\mathbf{Z}] \leq \eta$ warrants $\mathbb{E}_{pm}[L^{(r)}(\hat{Y}^*(\mathbf{R}), Y)] \leq \left(\eta^{1/r} + \mathbb{E}_{pm}^{1/r}\left[L(\hat{Y}^0, Y)\right]\right)^r$, where $O(\cdot)$ stands for the well-known big O. In other words, each constraint controls the loss of accuracy in terms of the chosen loss function L due to nonselected units, up to sampling design variability.

For these loss functions, each constraint can be always written as a bound on a quadratic form, denoted by $\mathbb{E}_m[\mathbf{R}^T\Delta\mathbf{R}|\mathbf{Z}]$ (see appendix A). Particular forms suitable for the stochastic and combinatorial problems will be explained in Subsection 2.2 and 2.3. The $n \times n$ matrix Δ specifies the potential losses at the unit level. Measures of bias and/or MSE seem natural in practice and they stem from the choice of the absolute or the squared loss function respectively. These measures can be heuristic in nature, such as the pseudo-bias for traditional score functions, or explicitly derived under some appropriate measurement-error model. In particular, non-zero off-diagonal terms of Δ allow for cross-unit terms to be included in the "overall" loss.

The choice of the matrix Δ is naturally linked to the choice of the loss function L, hence the term loss matrix (see appendix A for details). Thus, if Δ is diagonal with entries $|w_{ks}\boldsymbol{\epsilon}_k|$,

then we are choosing the absolute loss so that $\mathbb{E}_m[L(\hat{Y}^*(\mathbf{R}), \hat{Y}^0)|\mathbf{Z}]$ is also bounded by η (up to sampling design factors). This is targeted at the bias. Similarly, if $\Delta_{kl} = w_{ks}w_{ls}\epsilon_k\epsilon_l$, then we are choosing the squared loss so that $\mathbb{E}_m[(L(\hat{Y}^*(\mathbf{R}), \hat{Y}^0)|\mathbf{Z}]$ is also equally bounded. In turn, this is targeted at the mean squared error. In both cases, model-based techniques using data from the current time period can be applied in the combinatorial version, whereas in the stochastic version we are obliged to resort to auxiliary information from other periods.

For instance, the (local) score for a given y-variable is usually conceived as the product of a "risk" component and an "influence" component. A generic measure can be given using a model-based approach. Let $p_k = P\left(y_k^0 \neq y_k|y_k\right)$, that is, the posterior probability that the true value is different from the observed one. Let $\tilde{\mu}_k = \mathbb{E}_m\left(y_k^0|y_k, y_k^0 \neq y_k\right)$, that is, the conditional expectation of the true value given that it is different from the observed one. Then, we have

$$\mathbb{E}_m\left(y_k^0|y_k\right) = (1 - p_k)y_k + p_k\tilde{\mu}_k \quad \text{and} \quad \delta_k = y_k - \mathbb{E}_m\left(y_k^0|y_k\right) = p_k(y_k - \tilde{\mu}_k)$$

It follows that $w_k\delta_k$ can be used to construct the local score of unit k with respect to y, which is the product of "risk" measured by p_k and "influence" measured by $w_k(y_k - \tilde{\mu}_k)$, where w_k can be the sample weight, for example. Di Zio and Guarnera (2013) derive such a measure under the contamination model, which is suitable for the combinatorial problem. For the stochastic problem, where scoring does not use observations other than the unit at hand, $\tilde{\mu}_k$ cannot be evaluated for the current sample data and instead information from preceding realizations of this survey or similar surveys must be used. It is customary to replace it with some reference value, such as y_k from a previous time point, giving rise to a pseudo-bias. Nor can the "risk" component be assessed properly, and some heuristics measure might be used, such as in the SELEKT approach of SCB (Statistics Sweden 2011). The auxiliary information, which we exploit in the observation-prediction model (see Section 3), is fundamental.

The main difference between both versions arises when considering their actual application. The stochastic problem, supplemented by the assumption that ignores the cross-unit terms, allows the construction of score functions to be applied independently to each unit. The supplementary assumption amounts to considering these cross-terms more or less constant over time, hence playing no significative role in the selection. Conversely, the combinatorial problem needs a sufficient number of observations available to carry out the selection jointly for all units.

Taking into account the possibility of multiple constraints, we now arrive at the following general optimization problem:

$$\begin{aligned} [P_0] \quad & \max \ \mathbb{E}_m[\mathbf{1}^T\mathbf{R}|\mathbf{Z}] \\ & \text{s.t.} \quad \mathbb{E}_m[\mathbf{R}^T\Delta^{(q)}\mathbf{R}|\mathbf{Z}] \leq \eta_q, \quad q = 1, 2, \ldots, Q, \\ & \mathbf{R} \in \Omega_0 \end{aligned}$$

where Ω_0 denotes the admissible outcome space of $\mathbf{R}$, and q refers to the different constraints. Manipulation of Ω_0 creates extra flexibility for adoption. For instance, the problem can be recast for selection conditional on the units that have already been selected, by restricting Ω_0 such that certain R_k s are fixed at 0. The different constraints

may arise from the fact that there are multiple y-variables of interest, or the constraints may be directed at the different population domains even when there is only a single y-variable. In particular, the loss matrices $\Delta^{(1)}, \ldots, \Delta^{(Q)}$ may all be derived under a single multivariate model for the joint data, even when the bounds are marginally specified for each target quantity on its own.

Variations of the optimization problem stated above are possible, by either adopting a different function for optimization and/or different forms of constraints. For instance, maximization may be changed to minimization as long as suitable alterations of the selection variables and the loss functions are provided. Alternatively, one may for example use $w_k\delta_k$ in Δ but state the constraint as $\mathbb{E}_m[|\mathbf{R}^T\Delta\mathbf{R}|\,|\mathbf{Z}] \leq \eta$. We do not explicitly consider such variations of the problem in this article, but note that (i) it is possible to adapt the solutions presented below, should such variations be desirable in practice, and (ii) the expounded optimization approach can be carried out in the same spirit.

2.2. *The Stochastic Optimization Problem*

As stated above, the main assumption in this version of problem P_0 is neglecting the cross-unit terms in each constraint. Then these constraints can be rewritten as $\mathbb{E}_m[\mathbf{R}^T\Delta\mathbf{R}|\mathbf{Z}] = \mathbb{E}_m[\mathbf{R}^T\mathrm{diag}(\Delta)|\mathbf{Z}]$. Furthermore, the distinction between $\mathbf{Z}^{long} = s$ and $\mathbf{Z}^{long} \equiv \{s, \mathbf{X}\}$ is a matter of choice. In the former case, the restrictions are required to be fulfilled only on average for all realizations of the survey, whereas in the latter case they are imposed on the current realization, given the realizations of preceding time periods. The deduced stochastic optimization problem is solved in Arbués et al. (2012a) by using the duality principle, the sample average approximation and the interchangeability principle. The solution resulting from this linear problem is given in terms of matrices $\boldsymbol{M}^{(q)} = \mathbb{E}_m[\Delta^{(q)}|\mathbf{Z}^{cross}]$. This dependence on $\mathbf{Z}^{cross}$ may seem misleading, but only momentarily. Since this selection scheme is to be applied unit by unit upon receipt of each questionnaire and no cross-sectional information except that regarding each unit k separately will be used effectively, the formal conditioning upon $\mathbf{Z}^{cross}$ reduces effectively to conditioning upon the information $\mathbf{Z}_k^{cross} = \{s, \mathbf{x}, \mathbf{y}_k\}$ of each unit. Thus we write $\boldsymbol{M}^{(q)} = \mathbb{E}_m[\Delta|\mathbf{Z}^{cross}] = \mathrm{diag}\left(\mathbb{E}_m\left[\Delta_{kk}^{(q)}|\mathbf{Z}_k^{cross}\right]\right) = \mathrm{diag}\left(M_{kk}^{(q)}\right)$ On the other hand, in order to obtain the optimal Lagrange multipliers λ_q^* involved in the dual problem, a historic double-data set with raw and edited values is necessary. Putting it all together we arrive at the final solution, which only requires the diagonal entries of the matrices $\boldsymbol{M}^{(q)}$:

$$R_k = \begin{cases} 1 & \text{if} \quad \sum_{q=1}^{Q} \lambda_q^* M_{kk}^{(q)} \leq 1, \\ 0 & \text{if} \quad \sum_{q=1}^{Q} \lambda_q^* M_{kk}^{(q)} > 1. \end{cases} \tag{1}$$

Note that since the scheme is "trained" on the historic data, the evaluation of $M_{kk}^{(q)}$ given the observations in the current sample necessarily yields a pseudo-measure, regardless of the definition of the loss matrices.

This provides a score function for unit-by-unit selection. In the special case of $Q = 1$, unit k is selected provided $M_{kk} > 1/\lambda^*$, so that M_{kk} can be regarded as a single score and $1/\lambda^*$ as the threshold value. Equivalently, one may consider $\lambda^* M_{kk}$ as a "standardized"

score, in the sense that the threshold value is generically set to 1. The latter extends in a straightforward manner to the setting with multiple constraints, where each $\lambda_q^* M_{kk}^{(q)}$ is a standardized local score, and $\sum_{q=1}^{Q} \lambda_q^* M_{kk}^{(q)}$ is the standardized global score, with the generic global threshold value 1.

The global scoring derives from the linear structure of the dual problem and few variations are allowed without a substantial modification of problem P_0. As an exception, if a global score is initially envisaged as the weighted sum of local scores, then one may incorporate each weight into the constraint that generates the corresponding standardized local score to begin with.

The stochastic problem thus clarifies the fact that the performance of unit-by-unit selection can only be established over hypothetical repetitions of the selection process. At the end of each selection process, we have the realized selection strategy $\mathbf{r}$, and the realized loss $\sum_{k=1}^{n} r_k M_{kk}^{(q)}$, which can either be higher or lower than the specified bound η_q, for $q = 1, \ldots, Q$. Upon any hypothetical repetition of the selection process, however, y_k and y_k^0 will vary, and so will the corresponding $M_{kk}^{(q)}$ and r_k. It is over such hypothetical repetitions that the constraint $\mathbb{E}_m[\mathbf{R}\Delta^{(q)}\mathbf{R}|\mathbf{Z}] \leq \eta_q$ can possibly be satisfied, but not for each particular realization of the selection process.

2.3. The Combinatorial Optimization Problem

The combinatorial problem deals with the selection among all (or a group of) units. Cross-unit terms are now allowed and the information actually used is that given by the sample s, the auxiliary covariates $\mathbf{X}$ and the variables of interest $\mathbf{Y}$, that is by $\mathbf{Z} = \mathbf{Z}^{cross}$. Notice that all this information is available only after all questionnaires have been collected, thus it is only applicable as a form of output editing. It is easily proved that each constraint reduces to $\mathbb{E}_m[\mathbf{R}^T\Delta^{(q)}\mathbf{R}|\mathbf{Z}^{cross}] = \mathbf{r}^T\mathbf{M}^{(q)}\mathbf{r}$, where $\mathbf{M}^{(q)} = \mathbb{E}_m[\Delta^{(q)}|\mathbf{Z}^{cross}]$, which can now be possibly evaluated under some measurement-error model. Consequently, it becomes possible to establish the performance of the realized selection strategy directly. The optimization problem can be rephrased as

$$
\begin{aligned}
[P_{co}(\mathbf{M}, \eta, \Omega_0)] \quad & \max \mathbf{1}^T\mathbf{r} \\
& \text{s.t.} \quad \mathbf{r}^T\mathbf{M}^{(q)}\mathbf{r} \leq \eta_q, \quad q = 1, 2, \ldots, Q, \\
& \mathbf{r} \in \Omega_0
\end{aligned}
$$

Note that a more direct derivation can be obtained by not promoting the selection strategy vector $\mathbf{r}$ to a random vector $\mathbf{R}$ when modelling the measurement errors.

This combinatorial problem is solved in two different forms using two greedy algorithms, which run in $n^4 \cdot Q$ and $n^3 \cdot Q$ times, respectively. The solution of both algorithms is not exact a priori but suboptimal with a good degree of approximation. The faster algorithm is noticeably less precise than the slower one. This lack of precision entails a small amount of overediting in practice, that is, more units than those optimally obtained will be selected. The fourth and third power dependence on n may appear discouraging for practical applications. However, firstly, the input size P in problem P_{CO} is actually $P = O(n^2)$, thus the algorithms run in $O(P^2)$ and $O(P^{3/2})$, which are acceptable speeds for combinatorial problems. Secondly, in practice the problem is intended to be

applied not to entire samples but to their breakdowns into publication cells, which are the figures upon which precision is called for (see Section 6). These heuristic algorithms locally search the optimum in each iteration until the current solution satisfies all the restrictions. To do this we introduce infeasibility functions $h_i(\mathbf{r})$ for each algorithm $i = 1$, 2 (see Salgado et al. 2012 for details) indicating whether a solution satisfies all the restrictions ($h(\mathbf{r}) = 0$) or not ($h(\mathbf{r}) > 0$). Both algorithms start from the initial solution $\mathbf{r} = 1$ and in each iteration select the next unit in a locally optimal way until all restrictions are satisfied. The infeasibility functions will also be used later when constructing the prioritization of units.

Finally, we can regard both versions as related to two different approaches to the problem of optimization under uncertainty (see e.g., Wets 2002). The combinatorial version is consistent with the wait-and-see approach, since it puts off all decisions until all the information is available. The stochastic version is, at least partially, a here-and-now approach, since the decision about the procedure or rule of selection (although not the selection itself) is made before the data collection.

3. The Observation-Prediction Model

To substantiate the constraints in both versions of the optimization problem, we need to compute the loss matrices $\mathbf{M}^{(q)} = \mathbb{E}_m[\Delta^{(q)}|\mathbf{Z}^{cross}]$ and to choose the bounds η_q. We now show how to undertake the former whereas the latter is dealt with in the next section.

To compute the loss matrices we make use of the standard model-based techniques, but not in a conventional way. Let us digress very briefly. When facing the editing tasks and, in particular, the selection of units, one resorts to the very best auxiliary information available at that precise moment. With full generality, this will comprise (i) the reported values of the variables of analysis $\mathbf{y}_k^{(t)}$ for the present ($t = T$) and preceding ($t < T$) time periods, (ii) the true values of these variables $\mathbf{y}^{(0,\,t)}$ for those edited units in the past $t < T$, (iii) and the values of auxiliary covariates $\mathbf{x}_k^{(t)}$ for all time periods. In the notation of preceding sections, we have $\mathbf{y}_k = \mathbf{y}_k^{(T)}$, $\mathbf{y}_k^0 = \mathbf{y}_k^{(0,T)}$ and $\mathbf{x}_k = \mathbf{y}_k^{(t_1)}, \mathbf{y}_k^{(0,t_1)}, \mathbf{x}_k^{(t_2)}$, with $t_1 < T$ and $t_2 \leq T$. Note that some of these values can be coincidentally equal (e.g., when the measurement error is null) and that $\mathbf{y}_k^0$ is only known after accomplishing the editing work. But this is not everything. We also know (at least we can know) a point prediction $\hat{y}_k$ for each y-variable based on these auxiliary variables. For instance, we can make use of a time series model $\left\{y_k^{(0,t)}\right\}_{t<T}$ to make a point prediction $\hat{y}_k^{(T)}$. Different choices arise depending on the amount and type of auxiliary information. These predictions will enter into the selection problem as auxiliary covariates, so that $\mathbf{x}_k = \mathbf{y}_k^{(t_1)}, \mathbf{y}_k^{(0,t_1)}, \hat{\mathbf{y}}_k^{(T)}, \mathbf{x}_k^{(t_2)}$, with $t_1 < T$ and $t_2 \leq T$.

Let us denote by m^* the auxiliary model used to make the predictions $\hat{\mathbf{y}}_k$, not to be confused with the measurement error model m considered throughout this paper. This measurement error model m is given as usual in terms of (i) the conditional distribution of the predicted values $\mathbf{y}$ upon the true values $\mathbf{y}^0$, and (ii) the distribution of $\mathbf{y}^0$ conditional on the available auxiliary information $\mathbf{X}$. To be specific, for a y-variable we will assume $y_k = y_k^0 + \epsilon_k^{obs}$ and $y_k^0 = \hat{y}_k + \epsilon_k^{pred}$. In other words, we are using the predicted value computed according to the auxiliary model m^* as an exogenous variable for the model regarding y^0. In this sense we refer to this proposal as an observation-prediction model.

Generalizing these ideas, let us consider

i) an observation model $\mathbb{P}_{obs|0}(\mathbf{y}|\mathbf{y}^0)$, that is, a conditional probability distribution for the observed values $\mathbf{y}$ given the true values $\mathbf{y}^0$;
ii) a prediction model $\mathbb{P}_{0|pred}(\mathbf{y}^0|\hat{\mathbf{y}})$, that is, a conditional probability distribution for the true values $\mathbf{y}$ given the predicted values $\mathbf{y}^0$ according to an auxiliary model m^*.

Now let us denote by $\mathbb{P}_{obs|pred}$ the probability distribution of $\mathbf{y}$ conditional on the predicted values $\hat{\mathbf{y}}$ and by $\mathbb{P}_{0|obs,pred}$ the probability distribution of the true values $\mathbf{y}^0$ conditional on the observed values $\mathbf{y}^{obs}$ and the predicted values $\hat{\mathbf{y}}$. Then by Bayes' theorem or a generalization thereof, we can write

$$\mathbb{P}_{0|obs,pred} = \frac{\mathbb{P}_{obs|0} \times \mathbb{P}_{0|pred}}{\mathbb{P}_{obs|pred}} \tag{2}$$

The product must be understood in a suitable generalized form when the distributions are completely general. As usual, if the probability distributions are absolutely continuous with density functions $f.(\cdot)$, the equation (2) can be easily recognized as

$$f_{0|obs,pred}(\mathbf{y}^0) = \frac{f_{obs|0}(\mathbf{y}|\mathbf{y}^0, \hat{\mathbf{y}}) f_0(\mathbf{y}^0|\hat{\mathbf{y}})}{\int_{\mathbb{R}^Q} f_{obs|0}(\mathbf{y}|\mathbf{y}^0, \hat{\mathbf{y}}) f_0(\mathbf{y}^0|\hat{\mathbf{y}}) \mathrm{d}\mathbf{y}^0}.$$

The discrete case also boils down to applying Bayes' theorem. Once we have the distribution $\mathbb{P}_{0|obs,pred}$, the loss matrices can be computed as

$$\mathbf{M}^{(q)} = \mathbb{E}_{0|obs,pred}[\Delta^{(q)}|S, \mathbf{Y}, \hat{\mathbf{Y}}]. \tag{3}$$

To illustrate this proposal, let us consider the following generic example with a continuous variable y. Let define the us define the observation model $y_k^{obs} = y_k^0 + \epsilon_k^{obs}$ and the prediction model $y_k^{obs} = \hat{y}_k + \epsilon_k^{pred}$, with the following specifications:

1. $\epsilon_k^{obs} = \delta_k^{obs} e_k$.
2. $e_k \simeq Be(p_k)$, where $p_k \in (0, 1)$.
3. $\left(\epsilon_k^{pred}, \delta_k^{obs}\right) \simeq N\left(\mathbf{0}, \begin{pmatrix} \nu_k^2 & \rho_k \sigma_k \nu_k \\ \rho_k \sigma_k \nu_k & \sigma_k^2 \end{pmatrix}\right)$.
4. ϵ_k^{pred}, δ_k^{obs} and e_k are jointly independent of $\mathbf{Z}_k^{cross}$.
5. e_k is independent of ϵ_k^{pred} and δ_k^{obs}.

These are equivalent to stating that unit k has a probability $1 - p_k$ of reporting a value without measurement error $\left(y_k = y_k^0\right)$ and, when reporting an erroneous value, the measurement error distributes as a normal random variable with zero mean and variance σ_k^2. On the other hand, the prediction error distributes as a normal random variable with zero mean and variance ν_k^2. Both errors distribute jointly as a bivariate normal random variable with correlation ρ_k. Reporting an erroneous value is independent of both types of errors.

For the time being let us assume that the parameters $\theta = \left(p_k, \sigma_k^2, \nu_k^2, \rho_k\right)^T$ are known. Let us focus on the squared loss function. Then it is easy to prove (Arbués et al. 2012a) that

$$\mathbb{E}_m\left[\left(y_k - y_k^0\right)|s_k, y_k, \hat{y}_k\right] = \nu_k \cdot \frac{\sigma_k^2 + \rho_k\sigma_k\nu_k}{\sigma_k^2 + \nu_k^2 + 2\rho_k\sigma_k\nu_k} \cdot \left(\frac{y_k - \hat{y}_k}{\nu_k}\right) \cdot \zeta_k\left(\frac{y_k - \hat{y}_k}{\nu_k}\right), \tag{4}$$

$$\mathbb{E}_m\left[\left(y_k - y_k^0\right)^2|s_k, y_k, \hat{y}_k\right] = \nu_k^2 \cdot \left(\frac{\sigma_k^2 + \rho_k\sigma_k\nu_k}{\sigma_k^2 + \nu_k^2 + 2\rho_k\sigma_k\nu_k}\right)^2 \cdot$$

$$\left[\frac{\sigma_k^2\left(1-\rho_k^2\right)\left(\sigma_k^2 + \nu_k^2 + 2\rho_k\sigma_k\nu_k\right)}{\left(\sigma_k^2 + \rho_k\sigma_k\nu_k\right)^2} + \left(\frac{y_k - \hat{y}_k}{\nu_k}\right)^2\right] \cdot \zeta_k\left(\frac{y_k - \hat{y}_k}{\nu_k}\right),$$

$$\mathbb{E}_m\left[\left(y_k - y_k^0\right)\left(y_l - y_l^0\right)|s_k, y_k, \hat{y}_k\right] = \mathbb{E}_m\left[\left(y_k - y_k^0\right)|s_k, y_k, \hat{y}_k\right]\mathbb{E}_m\left[\left(y_l - y_l^0\right)|s_k, y_k, \hat{y}_k\right],$$

$$k \neq l,$$

where

$$\zeta_k(x) = \frac{1}{1 + \frac{1-p_k}{p_k}\left(\frac{\nu_k^2}{\sigma_k^2 + \nu_k^2 + 2\rho_k\sigma_k\nu_k}\right)^{-1/2} \exp\left(-\frac{1}{2}\frac{\sigma_k^2 + 2\rho_k\sigma_k\nu_k}{\sigma_k^2 + \nu_k^2 + 2\rho_k\sigma_k\nu_k}x^2\right)}.$$

Should we choose the absolute loss function, then, under the same hypotheses, we would have (see appendix A):

$$\mathbb{E}_m\left[|y_k - y_k^0||s_k, y_k, \hat{y}_k\right] = \sqrt{\frac{2}{\pi}} \cdot \nu_k \cdot {}_1F_1\left(-\frac{1}{2}; \frac{1}{2}; -\frac{(y_k - \hat{y}_k)^2}{2\nu_k^2}\right) \cdot \zeta_k\left(\frac{y_k - \hat{y}_k}{\nu_k}\right), \tag{5}$$

where ${}_1F_1(a; b; x)$ denotes the confluent hypergeometric function of the first kind.

The estimation of the parameters θ depends on the scenario. For the stochastic problem, as before, we are obliged to use some reference values or heuristic measures. Once more we resort to the auxiliary information. Our choice depends very much on the amount and type of auxiliary information. From the historic double-data sets comprising τ past time periods (e.g., a fixed panel) we can compute

$$\hat{p}_k = \frac{1}{\tau}\sum_{t=1}^{\tau} I_{y_k^{(t)} \neq y_k^{(0,t)}},$$

$$\hat{\sigma}_k^2 = \frac{1}{\tau - 1}\sum_{t=1}^{\tau}\left(\epsilon_k^{(t)} - \bar{\epsilon}_k\right)^2,$$

$$\hat{\nu}_k^2 = \frac{1}{\tau - 1}\sum_{t=1}^{\tau}\left(\epsilon_k^{(t)} - \bar{\epsilon}_k\right)^2, \text{ where } \epsilon_k^{(t)} = \hat{y}_k^{(t)} - y_k^{(0,t)},$$

$$\hat{\rho}_k = \frac{1}{\tau - 1}\sum_{t=1}^{\tau}\left(\epsilon_k^{(t)} - \bar{\epsilon}_k\right)\left(\epsilon_k^{(t)} - \bar{\epsilon}_k\right).$$

In case of rotating panels or sampling designs with too short a continuity in the sample for a number of units, we are forced to make simplifying assumptions such as partitioning

the sample $s = \cup_{i=1}^{I} s_i$ and positing $\boldsymbol{\theta}_k = \boldsymbol{\theta}_i$ if $k \in s_i$. We can also adopt these assumptions for some of the parameters. The extreme case would $\boldsymbol{\theta}_k = \boldsymbol{\theta} = (p, \sigma^2, \nu^2, \rho)^T$ for all $k \in s$, which can be further supplemented with extra hypotheses such as $\rho = 0$.

On the other hand, for the combinatorial problem we do have (almost) the complete current sample so that we can make use of these data, although with important limitations. It is clear that it is impossible to estimate each $\boldsymbol{\theta}_k$ using only the current sample. We are obliged to make some simplifying assumptions, as above. In practice, however, it is advisable to use not only data from the current time period ($t = T$), but also from preceding periods ($t < T$). The stationarity across time periods of the response mechanism supports this course of action.

Alternatively, the contamination model by Di Zio and Guarnera (2013) is a relevant example of a model-based technique which uses exclusively data from the current period (except for the covariates for the model) to estimate the model parameters. The usage of statistical models to make the selection of units allows us to cherish the hope of extending this approach to qualitative and semicontinuous variables, thus paving the way for the use of selective editing in household surveys.

4. Fieldwork: Selection and Prioritization of Units

The problem is not completely specified until we choose the bounds η_q to formulate the optimization problem completely. The bound η on a given constraint $\mathbb{E}_m[\mathbf{R}^t \Delta \mathbf{R} | \mathbf{Z}] \leq \eta$ can be set either absolutely or relatively in terms of a chosen figure of merit or reference value. This can be, for example, the a priori variance used in the sampling design phase so that the constraint establishes a bound for the loss of accuracy as a fraction of the desired precision. The decision will necessarily involve some subject-matter knowledge.

So far, the formulation of the selective editing problem as an optimization problem is complete, providing a *selection* of units expressed by the solution $\mathbf{r}$. However, in practice having a selection of units must be confronted with the actual conditions of fieldwork. In particular, both controllability and availability of resources, such as person hours for example, are important issues in this respect. Given a particular selection, either we may run out of resources and cannot edit all selected units or we may finish the editing field work ahead of time and thus miss the opportunity to achieve better accuracy. In this sense it seems natural to have at our disposal a set of selections to optimize the actual use of resources. We achieve this by having a *prioritization* of units. Next we show how to prioritize units in the optimization approach. In section 6 we discuss in more detail this issue of the selection/prioritization of units in relation with the fieldwork.

From the preceding sections it is clear that it does not make sense to prioritize units in the stochastic formulation. On the other hand, to prioritize units in the combinatorial version we propose combining different selections by choosing a sequence of appropriate values as bounds. The basic idea is to choose large initial bounds which drive us to select no unit, then to decrease the bounds until one unit is selected and to flag this unit for future selections. Then we again decrease the bounds until a new unit is selected and flagged for future selections. The procedure is repeated until all units have been flagged.

Let $f^{[k]} \subset s = \{1, \ldots, n\}$ denote the set of flagged units at iteration k and $\Omega_0^{[k]}$ the outcome space of the combinatorial problem at iteration k. For any given strategy vector $\mathbf{r}$

we denote by $\mathcal{I}^{-1}(\mathbf{r})$ the set of strategy vectors $\bar{\mathbf{r}}$ obtained from $\mathbf{r}$ transforming exactly a component 1 into 0. For example, $\mathcal{I}^{-1}((1,1,0)^T) = \{(0,1,0)^T, (1,0,0)^T\}$. Let h denote the infeasibility function used in the greedy algorithms (see section 2.3).

The algorithm of prioritization reads as follows:

1. Set $f^{[0]} = \emptyset$, $\Omega_0^{[0]} = \{0,1\}^{\times n}$, $\mathbf{s}^{[0]} =$ and $\boldsymbol{\eta}^{[0]} = (\mathbf{s}^{[0]T}M^{(1)}\mathbf{s}^{[0]}, \ldots, \mathbf{s}^{[0]T}M^{(Q)}\mathbf{s}^{[0]})^T$.
2. FOR $k = 0$ TO $k = n$
   i. Set $\mathbf{s}^{[k+1]} = \arg\min_{\mathbf{s}\in\mathcal{I}^{-1}(\mathbf{s}^{[k]})}(h(\mathbf{s}))$. In case of multiple $\mathbf{s}^{[k+1]}$ choose one at random.
   ii. Set $l^* \in s$ such that $s_{l^*}^{[k+1]} \neq s_{l^*}^{[k]}$.
   iii. Set $f^{[k+1]} = f^{[k]} \cup \{l^*\}, \Omega_0^{[k+1]} = \Omega_0^{[k]} - \{\mathbf{s}^{[k]}\}$ and $\boldsymbol{\eta}^{[k+1]} = (\mathbf{s}^{[k+1]T}M^{(1)}\mathbf{s}^{[k+1]}, \ldots, \mathbf{s}^{[k+1]T}M^{(Q)}\mathbf{s}^{[k+1]})^T$.
3. FOR $k = 0$ TO $k = n$
   i. Set $\mathbf{r}^{[k]} = \arg\max[P_{co}(\mathbf{M}, \boldsymbol{\eta}, \Omega_0^{[k]})]$.
4. Set $\boldsymbol{s} = \sum_{k=0}^{n}\mathbf{r}_k$.

The vector $\mathbf{s}$ provides the prioritization: unit k must be edited in the s_kth place. Notice that steps 1 and 2 provide a sequence of bounds $\boldsymbol{\eta}^{[k]}$ and a sequence of outcome sets $\Omega_0^{[k]}$ which are used in step 3 to solve $n+1$ concatenated combinatorial problems. Two comments are in place here. On the one hand, in practice, Step 3 indeed reduces to the first point in Step 2 since $\mathbf{r}^{[k]} = \mathbf{s}^{[k]}$ because h is the infeasibility function of the optimization algorithm.

On the other hand, this invites us to reconsider the role of the infeasibility function in the prioritization of units: This depends on the choice of h. Should we choose, instead of the original infeasibility function $h_1(\mathbf{r}) = \sum_{q=1}^{Q}\left(\mathbf{r}^t M_{kl}^{(q)}\mathbf{r} - m_q^2\right)^+$ of algorithm 1, the function $h(\mathbf{r}) = \sum_{q=1}^{Q} w_q\left(\mathbf{r}^t M_{kl}^{(q)}\mathbf{r} - m_q^2\right)^+$, where $w_q \geq 0$ are positive weights expressing the different priority given to the accuracy of each variable $y^{[q)}$, we would arrive at a different prioritization. This can also be viewed more geometrically. To produce a sequence of bounds we begin by having no selected units, that is, by $\boldsymbol{\eta}_0 = (\mathbb{1}^t M^{(1)}\mathbb{1}, \ldots, \mathbb{1}^t M^{(Q)}\mathbb{1})^t$, and we need to produce a sequence of points in $\mathbb{R}^Q$ such that its final point is $\mathbf{0}$. There exist infinitely many possibilities (see Figure 1). In this context, the prioritization amounts to choosing a path from $\boldsymbol{\eta}_0$ to $\mathbf{0}$. This path expresses the priority which the statistician gives to the accuracy of the different estimators along the process of prioritization of units. The original infeasibility functions of the algorithms confer the same relevance on every estimator $\hat{Y}^{(q)}$.

5. A Selection Efficiency Measure: Comparison with the Score Function Approach

To make a comparison of the selection undertaken under any approach, we propose the following selection efficiency measure for an estimator $\hat{Y}$. Beforehand, we need a double data set with raw and edited values according to a gold standard so that when a unit is selected, its raw values are substituted by their corresponding edited counterparts, considered true. We will denote by $\hat{Y}^{sel}(n_{ed})$ the estimator obtained when n_{ed} questionnaires have been selected according to a selective editing technique *sel* and edited correspondingly. Note that $\hat{Y}^{sel}(n_{ed} = n) = \hat{Y}^0$. As a figure of merit for the

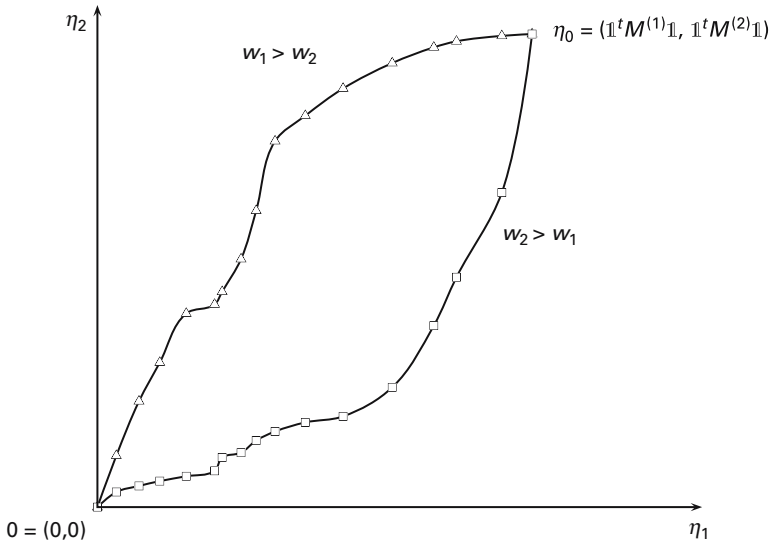

Fig. 1. Example of two different sequences of bounds with $Q = 2$ arising from different choices of the weights w_q.

selection of units we will focus upon the absolute relative pseudo-bias of an estimator $\hat{Y}$, given by $\widetilde{\mathrm{ARB}}\ (\hat{Y}^{sel}(n_{ed})) = \left|\frac{\hat{Y}^{sel}(n_{ed}) - \hat{Y}^0}{\hat{Y}^0}\right|$.

The rationale of the proeosed measure is the comparison with a random selection of units. The idea is to compare $\widetilde{\mathrm{ARB}}(\hat{Y}^{sel}(n_{ed}))$ for a selective editing technique *sel* with $\widetilde{\mathrm{ARB}}(n_{ed}) \equiv \widetilde{\mathrm{ARB}}(\mathbb{E}[\hat{Y}^{ran}(n_{ed})])$, where *ran* stands for an equal-probability selection and $\mathbb{E}$ is the expectation with respect to this random selection. It is immediate to show that $\widetilde{\mathrm{ARB}}_0\,(n_{ed}) = \left(1 - \frac{n_{ed}}{n}\right)\widetilde{\mathrm{ARB}}(\hat{Y}^{ran}(0))$. Let us denote by $\gamma_0(n_{ed})$ and $\gamma^{sel}(n_{ed})$ the straight and polygonal lines with vertices $\gamma_0(n_{ed}) \simeq \{(0, \widetilde{\mathrm{ARB}}_0\,(0)), (n_{ed}, \widetilde{\mathrm{ARB}}_0)(n_{ed}))\}$ and $\gamma^{sel}(n_{ed}) \simeq \{(0, \widetilde{\mathrm{ARB}}_0\,(\hat{Y}^{sel}(0))), (1, \widetilde{\mathrm{ARB}}_0\,(\hat{Y}^{sel}(1))), \ldots, (n_{ed}, \widetilde{\mathrm{ARB}}_0\,(\hat{Y}^{sel}(n_{ed})))\}$, respectively. Let us also denote by $A_\gamma(n_{ed})$ the signed area of the surface between the curve γ and the horizontal axis to the left of the vertical line at n_{ed} (see Figure 2). The area is agreed to be positive if the polygonal line lies below the straight line and is otherwise negative. We propose the following definition for the efficiency of the technique *sel*:

$$\epsilon^{sel}(n_{ed}) \equiv (A_{\gamma_0}(n_{ed}) - A_{\gamma^{sel}}(n_{ed}))/A_{\gamma_0}(n_{ed}) = 1 - \frac{A_{\gamma^{sel}}(n_{ed})}{A_{\gamma^0}(n_{ed})}.$$

Note that this measure depends on the number of units to select. This allows us to recognize those techniques which prioritize the most influential units first. A typical situation is depicted in Figure 2.

We have carried out a comparison of the preceding proposal of prioritization of units with that obtained from some score functions in the literature. In order to avoid possible interferences with missing data and units recently added to the sample, we have used a rectangular subset of the sample data of the Spanish ITI and INORI surveys (INE Spain 2010). For clarity's sake we shall concentrate on one particular score function, illustrate the corresponding results and make some comments regarding the similar behavior of all of them. We have used a slightly enhanced version of the RATIO function of Latouche and

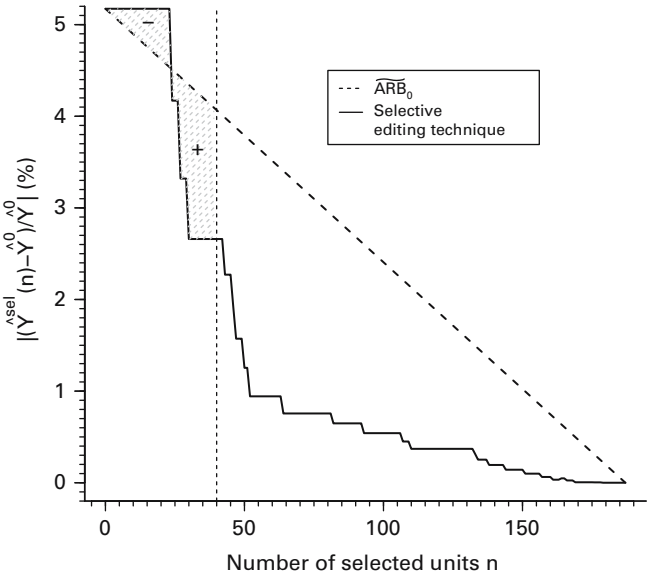

Fig. 2. Absolute relative pseudo-bias vs. number of selected units

Berthelot (1992). Let $r_k^{(t)} = \frac{y_k^{(t)}}{y_k^{(t-1)}}$ and define

$$\bar{r}_k^{(t)} = \begin{cases} \left| \dfrac{r_k^{(t)}}{\text{median}_k\left(r_k^{(t)}\right)} - 1 \right| & \text{if } r_k^{(t)} > \text{median}_k\left(r_k^{(t)}\right), \\ \left| 1 - \dfrac{r_k^{(t)}}{\text{median}_k\left(r_k^{(t)}\right)} \right| & \text{otherwise.} \end{cases}$$

Also define $g_k^{(t)} = w_{ks} \times \bar{r}_k^{(t)} \times \sqrt{\max\left(y_k^{(t)}, y_k^{(t-1)}\right)}$ and then the local score $s_k^{(t)} = \frac{|g_k^{(t)} - \text{median}_k\left(g_k^{(t)}\right)|}{\text{IQR}_k\left(g_k^{(t)}\right)}$, where IQR stands for the interquartile range. For $q = 1, \ldots, Q$ variables, these combine in the global score function defined as RATIO2$(k,t) = S_k^{(t)} = \sum_{q=1}^{Q} s_k^{(q,t)}$. The enhancement arises due to the fact that only data from the time period $t-1$ is used and not from $t-2$, as in the original proposal. Thus this function RATIO2 can only be used as a form of output editing after all data have been collected (as the combinatorial approach, which we are making the comparison with).

Regarding the prioritization of units computed under the combinatorial approach, firstly we must specify the auxiliary model m^* to find the predicted values $\hat{y}_k$. For each unit we have fitted three alternative time series models $\xi_1 : (1-B)_{z_t} = a_t$, $\xi_2 : (1-B^{12})_{z_t} = a_t$ and $\xi_3 : (1-B)(1-B^{12})_{z_t} = a_t$, where B stands for the backshift operator, $zt = \log\left(m + y_t^0\right)$ (m being a nuisance parameter estimated by maximum likelihood) and a_t denotes white noise. Each predicted value $\hat{y}_k$ is computed according to the corresponding best model ξ^* (in terms of the minimal estimated mean squared error). Since the sample is a fixed panel selected by cut-off, the sampling weights w_{ks} are all equal to 1.

Next, we have applied the generic univariate observation-prediction model illustrated in Section 3 to the logarithmic transforms of the turnover and the new orders received

independently. The common error probability $p_k = p$ and observation variance $\sigma_k^2 = \sigma^2$ have been estimated from the past three months using a double-data set. The prediction variance ν_k^2 has been computed according to the corresponding chosen best model ξ^* for each unit. As loss matrices, we have chosen both the squared and the absolute loss function with entries given by Equations (4) and (5), respectively.

Finally, to make the comparison with a random selection of units, we have computed the absolute relative pseudo-bias for 50 equal-probability random selections. We have calculated the mean and first and third quartiles of the corresponding distribution. This provides a confidence-like interval for each number of selected units (see Figure 3). The motivation is to provide an insight not only against the average random selection but also of its distribution.

We have carried out this comparison for 23 NACE Rev. 2 divisions and subdivisions (aggregations of groups according to subject-matter knowledge). Firstly, RATIO2 showed a better performance than the rest of score functions (RATIO, DIFF, FLAG ITI, FLAG INORI; see Latouche and Berthelot 1992). In 15 cases the absolute loss yielded the most efficient prioritization, with 9 of these cases having the RATIO2 score function as more efficient than the squared loss choice (Figure 3 illustrates this behavior). However in 5 cases it is the squared loss function that outperforms the other two choices, in which the absolute loss also did better than the score function. In the remaining 3 cases, RATIO2 slightly overcame the absolute loss, which in turn performed better than the squared loss.

Thus, in general, the absolute loss is more efficient than the squared loss in terms of the pseudo-bias, as expected. This also happens with the score function RATIO2, since it is also targeted at the bias. In general, the absolute loss is also more efficient than the score functions. However, in actual production conditions, both missing data and respondents newly added to the sample must be taken into account. In these cases, in the optimization approach the prediction values $\hat{y}_k$ must be imputed or fixed under some supplementary scheme, since the considered time series models fail to produce these values. As an elementary test, we assigned $\hat{y}_k = y_k$ in these cases in order for them not to be selected at first positions. The general result was a slight deterioration of the performance of the score functions for all values of n_{ed}, while in the optimization approach, the behavior was as good as before for the most influential units ($n_{ed} = 1, 2, \ldots$), but noticeably poorer for the

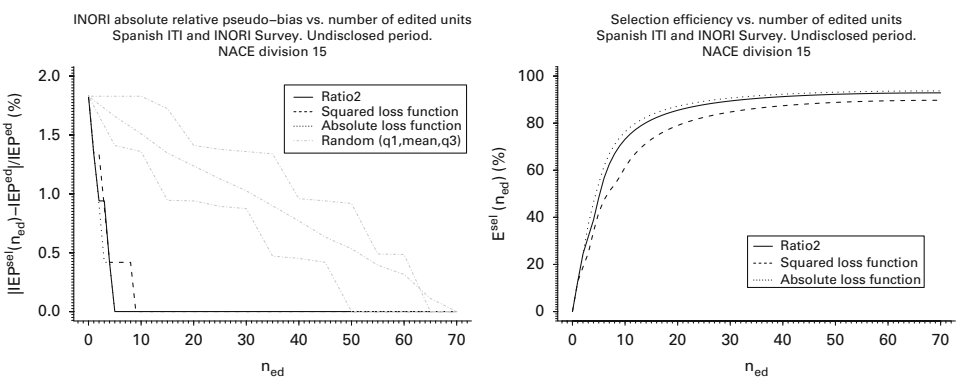

Fig. 3. Absolute relative pseudo-bias and editing efficiency vs. number of selected units

last units ($n_{ed} \geq n/2$). We have not considered these issues in the preceding comparison, since they belong to sophistications of the observation-prediction model.

In our opinion it is important to note that the above results have been obtained with crude time series models and extremely simplified assumptions, and they do not incorporate any subject-matter knowledge. Thus there is more room to elaborate further on them (using better parameters, building multivariate models, etc.). In this line of thought the most attractive point will arise if working models can be built for discrete or semicontinuous variables, paving the way for the use of selective editing techniques also in household surveys. The possibility of using well-established tools such as time series models or statistical models in general, reinforces the statistical defensibility of the data editing work.

6. Discussion and Concluding Remarks

Once we have detailed the methodological proposal, we now proceed to discuss several issues regarding this optimization approach from different perspectives. As two immediate objections, a cautious reader can point out the limitation to linear estimators and the polynomial running time of the algorithms. Firstly, the limitation to linear estimators, which contrasts with the common use in practice of some nonlinear estimators as such as ratio estimators or regression estimators, can be easily overcome as follows. In practice most non-linear estimators $\hat{Y}_{U_d}^{nl}$ are functions of linear estimators $\hat{Y}_{U_d}^{nl} = f\left(\hat{Y}_{U_d}^{(1)}, \ldots, \hat{Y}_{U_d}^{(M)}\right)$. Then instead of considering the corresponding restriction for the MSE of $\hat{Y}_{U_d}^{nl}$, we consider a restriction for each linear estimator $\hat{Y}_{U_d}^{(m)}$, $m = 1, \ldots, M$. The rationale amounts to expecting an accurate nonlinear estimator if each linear estimator is accurate. Moreover, a bounded growth in the number of restrictions is expected, since nonlinear estimators are usually built from different combinations of survey variables, whose number is fixed by the questionnaire. Secondly, the polynomial running time of the selection algorithms is not a practical concern, at least in Spanish sampling sizes standards, as we will now explain. On the one hand, the estimation problem in a finite population U is essentially a multivariate problem seeking accurate and numerically consistent estimations in given partitions of the population U. These partitions are fixed according to the breakdown established by the statistical dissemination plan of each survey. Thus the selection or prioritization should be applied to each of these publication cells, since no lack of accuracy is rightfully allowed in any published figure. On the other hand, we have applied this approach to the Spanish ITI and INORI survey as a pilot experience at INE Spain (details will be published elsewhere). In these monthly short-term business statistics, the sampling size amounts to around 12,000 industrial establishments broken down into 37 publication cells with sizes ranging up to 1,500 units at most. The prioritization of units in all cells took a total of three hours on a desktop PC, which is a reasonable working time.

As a deeper concern, one can inquire why the roles of the two basic principles of our formulation are not interchanged, that is, why data quality is not optimized (minimizing the loss function) restricting the amount of resources used (number of questionnaires to recontact). We give two reasons to support our proposal. From a broad perspective, in a statistical office it appears desirable to minimize the cost of each survey in order to optimize

resources to face and embrace as many other surveys in the statistical production as possible. In our view, this is a natural decision given the increasing demand for information from stakeholders. From a more methodological standpoint, the multivariate feature of the problem again arises. If we interchanged the roles of both principles, we would need to minimize the loss function of the different variable estimators corresponding to each publication cell restricted to the number of questionnaires to be recontacted. As a matter of fact this is a multiobjective optimization problem, which ineludibly needs some decisions to compute a solution (see e.g., Marler and Arora 2004). In this respect, our position in official statistics production is to minimize the number of decisions taken by the survey conductor, which is clearly expressed in the following citation by Hansen et al. (1983): "[. . .] it seems desirable, to the extent feasible, to avoid estimates or inferences that need to be defended as judgments of the analysts conducting the survey".

As a matter of fact, the question of the number of decisions is a first relevant point to establish a comparison with the score function approach. Nowadays the score function approach is undisputedly the favored technique for selecting influential units in the editing production phase. Thus it provides the framework assessing advantages and disadvantages of any other technique. Furthermore, in our opinion, a comparison will help us reveal fundamental aspects of the editing production phase irrespective of the particular techniques. Regarding the number of decisions let us recall that the score function approach comprises four main decisions to determine a selection of units (Lawrence and McKenzie 2000), namely (i) an editing model to construct the anticipated values, (ii) each local score function, (iii) a global score function, and (iv) a cut-off value. On the one hand, in the optimization approach the first three decisions are jointly substituted and integrated into a single step: the construction of the observation-prediction model or an alternative statistical error-modelling technique, and the subsequent formulation of the optimization constraints. Furthermore, in our view, this integration renders this selection procedure more natural within the statistical language, in contrast to a score function, which can seem extraneous. In this sense, let us point out that the construction of an observation-prediction model is a multivariate exercise, so the integration of the choices of both local and global score functions comes naturally together with the construction of the statistical model. On the other hand, the choice of the cut-off value is now substituted for the choice of the bounds in the optimization problem. In the score function approach, this value must be chosen normally using data from previous realizations of the survey and using a heuristic or empirical connection between this value and the chosen loss function of the survey estimators. In the optimization approach, the choice of the bounds makes use of a priori values of variances (or some other similar measure) as in the survey design stage and shows a neater connection with the loss function, thus fitting again more naturally into the whole survey statistics production process. Indeed, we have shown how the prioritization of units under the score function approach can be reproduced and slightly overcome with a very simple model. Furthermore, although admittedly still too far, this proposal points toward enlarging the traditional sampling strategy $(\mathcal{D}, T)$ comprising the sampling design $\mathcal{D}$ and the construction of the estimator T (see e.g., Hedayat and Sinha 1991) with a selection strategy R, so that we would have a triplet $(\mathcal{D}, R, T)$. This follows the spirit of the total survey design.

The selection/prioritization issue goes hand in hand with the double version of the optimization approach. This issue arises mainly from resource availability and

controllability, mainly of timeliness and person-hours in the editing fieldwork. When having a selection of units in practice we face two situations: Either we run out of resources to accomplish the interactive editing of all selected units or we end up ahead of time and then we miss the opportunity to gain more accuracy. Now, since editing near the source is a must for this production phase, it is advisable to have a real-time selection mechanism on each questionnaire, as pointed out in the introduction, independently of the rest of the sample. Conversely, on later stages it is preferable to prioritize units to edit (interactively) the most influential first. In this line of thought, the stochastic approach suits the selection whereas the combinatorial approach suits the prioritization. Furthermore, since both approaches derive from a common general framework focused on the exploitation of auxiliary information, we envisage a more complex, although unified, editing process. Let us parameterize the auxiliary information used in the editing work in terms of its longitudinal, cross-sectional and multivariate dimensions. By longitudinal we mean the value of variables for each unit in previous time periods. By cross-sectional we refer to the information stemming from the sample at the current period. Finally, by multivariate we mean the information arising from the multi-dimensional character of the survey (always several variables are investigated). If we focus on the longitudinal and cross-sectional dimensions of the auxiliary information, Figure 4 represents the transition from micro-selective to macro editing as the data collection stage is completed. In our view, these two editing techniques appear as the head and tail of a time-continuous process driven by the evolution of the data collection. We envisage that intermediate techniques combining both the longitudinal and available cross-sectional information as a time-continuous process during the data collection will be of practical usefulness.

Regarding the optimization approach, we want to point out that both versions fit naturally as the head and tail of this time-continuous editing process, so that the stochastic version corresponds to exploiting longitudinal information as in traditional selective editing techniques, whereas the combinatorial version arises as a macro editing technique focusing upon the cross-sectional information. In contrast, the score function approach and traditional macro editing techniques can hardly be seen under the same methodological

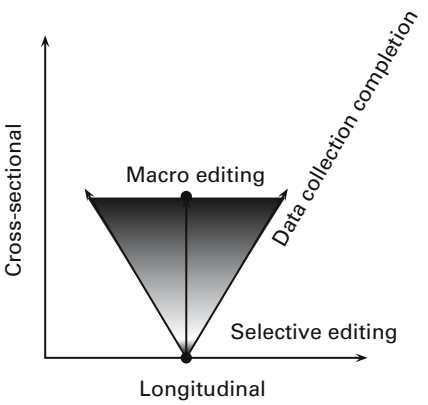

Fig. 4. Schematic representation of the transition from micro-selective to macro editing as data collection is completed. As data collection is completed, more cross-sectional information is available

principles. It remains open for future work to find a more general formulation for this proposed time-continuous process embedding both optimization versions.

A complementary comparison can be made with the automatic data editing techniques based on the Fellegi- Holt methodology, in particular with the different approaches to the error localization problem, which also make an extensive use of optimization techniques (see de Waal et al. 2011). The common points reduce to the fact that mathematical optimization appears as a natural translation of the proposed data editing principles. Conversely, the Fellegi-Holt methodology focuses upon each questionnaire, seeking to minimize the number of items to change satisfying all edits, whereas in this approach we focus upon the whole sample, seeking to minimize the number of units to be recontacted satisfying restrictions upon the loss functions using a statistical model instead of edits.

To conclude, as immediate future prospects, we have recently begun to analyze the inclusion of these techniques in the current E&I strategies in most business surveys in INE Spain. A pilot experience with the ITI and INORI survey fosters our hope to reduce current recontact rates and consequently both editing costs and the response burden at our office. R packages and SAS macros implementing this optimization approach are under intense development and being tested in these pilot experiences. Apart from this, more methodological research is needed to find generic multivariate models fitting the observation-prediction model and to generalize them to both qualitative and semicontinuous variables. In this context, multivariate models already present in the literature for data editing (Di Zio and Guarnera 2013) appear as a fruitful alternative. In addition, we already have a first adaptation of the preceding greedy algorithms to be applied to surveys with self-weighting samples and qualitative variables. We are collaborating with experts from the Spanish National Health Survey to produce an observation-prediction model adapted to these variables.

A. Mathematical Appendix

We include some mathematical proofs. Firstly we prove how the constraints imply a control on the loss of accuracy. In particular, if $L = L^{(r)}$ denotes the absolute ($r = 1$) or squared loss ($r = 2$) function, we prove that $\mathbb{E}_m[L(\hat{Y}^*(\mathbf{R})\hat{Y}^0)|\mathbf{Z}] \leq \eta$ (where $\mathbf{Z} = \mathbf{Z}^{st}$ or $\mathbf{Z}^{cross}$) implies $\mathbb{E}_{pm}[L(\hat{Y}^*(\mathbf{R}), Y)] \leq \left(\eta^{1/r} + \mathbb{E}_{pm}^{1/r}[L(\hat{Y}^0, Y)]\right)^r$. It is straightforward to prove that $d(A,B) = \mathbb{E}_{pm}^{1/r}[L^{(r)}(A,B))]$ is a metric. Then, by the triangle inequality, we have

$$d(\hat{Y}^*(\mathrm{R}), Y) \leq d(\hat{Y}^*(\mathrm{R}), \hat{Y}^0) + d(\hat{Y}^0, Y).$$

Now, using properties of the conditional expectation, we can write

$$d^r(\hat{Y}^*(\mathrm{R}), \hat{Y}^0) = \mathbb{E}_{pm}[\mathbb{E}_m[L(\hat{Y}^*(\mathbf{R}), \hat{Y}^0|\mathbf{Z}]] \leq \eta,$$

where $\mathbf{Z} = \mathbf{Z}^{st}$ or $\mathbf{Z}^{cross}$. The result follows immediately.

Secondly we show the connection between the loss matrices and the loss function. In the absolute loss case, we have $\mathbb{E}_m[|\hat{Y}^*(\mathbf{R}) - \hat{Y}^0|\,|\mathbf{Z}] = \mathbb{E}_m\left[\left|\sum_{k\in s} R_k w_{ks}\epsilon_k\right|\right]|z] \leq \left[\sum_{k\in s} R_k^2 |w_{ks}\epsilon_k|\,|\mathbf{Z}\right]$, since $R_k^2 = R_k$. Thus we can write $\mathbb{E}_m[|\hat{Y}^*(\mathbf{R}) - \hat{Y}^0|\,|\mathbf{Z}] = \mathbb{E}_m[\mathbf{R}^T\Delta\mathbf{R}|\mathbf{Z}]$, where Δ is diagonal with entries $\Delta_{kk} = |w_{ks}\epsilon_k|$. In the squared loss case, in turn we have $\mathbb{E}_m[(\hat{Y}^*(\mathbf{R}) - \hat{Y}^0)^2|\mathbf{Z}] =$

$\mathbb{E}_m\left[\sum_{k\in s}\sum_{k\in s} R_k R_l w_{ks}\epsilon_k w_{ls}\epsilon_l|\mathbf{Z}\right]$. Thus, we an also write $\mathbb{E}_m[(\hat{Y}^*(\mathbf{R}) - \hat{Y}^0)^2|\mathbf{Z}] = \mathbb{E}_m[\mathbf{R}^T\Delta\mathbf{R}|\mathbf{Z}]$, where $\Delta_{kl} = w_{ks}\epsilon_k \cdot w_{ls}\epsilon_l$.

The conditional moments (4) and (5) are found along similar lines. Under the hypotheses assumed in section 3 regarding the observation-prediction model, it follows that $y_k - \hat{y}_k = \epsilon_k^{obs} + \epsilon_k^{pred}$ and $\mathbb{E}_m\left[\left(y_k - y_k^0\right)^r|s_k, y_k, \hat{y}_k\right] = \mathbb{E}_m\left[\delta_k^{(obs)r}|s_k, y_k, \hat{y}_k\right]$. $\mathbb{E}_m[e_k|s_k, y_k, \hat{y}_k]$, with $r = 1, 2$. Conditioning on $s_k, y_k, \hat{y}_k$ amounts to conditioning on $s_k, \epsilon_k^{obs}, \hat{y}_k$, thus we can rewrite these conditional expectations as $\mathbb{E}_m[\cdot|s_k, y_k - \hat{y}_k, \hat{y}_k]$. Now the second term is computed using Bayes' theorem, so that $\mathbb{E}_m[e_k|s_k, y_k - \hat{y}_k, \hat{y}_k] = \zeta k\left(\frac{y_k - \hat{y}_k}{\nu_k}\right)$. For the first term, we notice that the random vector $\left(\delta_k^{obs}, \delta_k^{obs} + \epsilon_k^{pred}\right)^T$ is normally distributed with expectation $\mu = 0$ and variance $\Sigma = \begin{pmatrix} \sigma_k^2 & \sigma_k^2 + \rho_k\sigma_k\nu_k \\ \sigma_k^2 + \rho_k\sigma_k\nu_k & \sigma_k^2 \end{pmatrix}$. The conditional moments follow then from standard properties of the multivariate normal distribution.

7. References

Arbués, I., González, M., and Revilla, P. (2012a). A Class of Stochastic Optimization Problems with Application to Selective Data Editing. Optimization, 61, 265–286.

Arbués, I., Revilla, P., and Salgado, D. (2012b). Optimization as a Theoretical Framework to Selective Editing. UNECE Work Session on Statistical Data Editing, WP1, 1–10.

de Waal, T., Pannekoek, J., and Scholtus, S. (2011). Handbook of Statistical Data Editing and Imputation. New York: Wiley.

Di Zio, M., and Guarnera, U. (2013). A Contamination Model for Selective Editing. Q2 Journal of Official Statistics, xx, xxx-xxx.

EDIMBUS (2007). Recommended Practices for Editing and Imputation in Cross-sectional Business Surveys. ISTAT, CBS, SFSO, EUROSTAT. Available at http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPM_EDIMBUS.pdf. (Accessed Q3 January 10, 2013).

Granquist, L. (1997). The New View on Editing. International Statistical Review, 65, 381–387.

Hansen, M.H., Madow, W.G., and Tepping, B.J. (1983). An Evaluation of Model-dependent and Probability Sampling Inferences in Sample Surveys. Journal of the American Statistical Association, 78, 776–793.

Hedayat, A.S. and Sinha, B.K. (1991). Design and Inference in Finite Population Sampling. New York: Wiley.

INE Spain (2010). Industrial Turnover Indices. Industrial New Orders Received Indices. Base 2005. CNAE-09. Methodological Manual. Available at http://www.ine.es/en/metodologia/t05/t0530053_en.pdf. (Accessed January 10, 2013).

Latouche, M. and Berthelot, J.M. (1992). Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys. Journal of Official Statistics, 8, 389–400.

Lawrence, D. and McKenzie, R. (2000). The General Application of Significance Editing. Journal of Official Statistics, 16, 243–253.

Marler, R.T. and Arora, J.S. (2004). Survey of Multi-objective Optimization Methods for Engineering. Structural and Multidisciplinary Optimization, 26, 369–395.

Pannekoek, J., Scholtus, S., and van der Loo, M. (2013). Automated and Manual Data Editing: a View on Process Design and Methodology. Journal of Official Statistics, xx, Q2 xxx-xxx.

Salgado, D., Arbués, I., and Esteban, M.E. (2012). Two Greedy Algorithms for a Binary Quadratically Constrained Linear Program in Survey Data Editing. INE Spain Working Paper 02/12. Available at http://www.ine.es. (Accessed January 10, 2013).

Särndal, C.-E., Swensson, B., and Wretman, J.H. (1992). Model Assisted Survey Q3 Sampling. New York: Springer.

Statistics Sweden (2011). User's Guide to SELEKT 1.1 A Generic Toolbox for Selective Data Editing. Statistics Sweden document, February 17, 2011.

Wets, R.-B. (2002). Stochastic Programming Models: Wait-and-see Versus Here-and-now. Institute for Mathematics and Its Applications, 128.

Received February 2013
Accepted September 2013