Volumen 22 Número 3 Julio 2006

# Boletín de la Sociedad de Estadística e Investigación Operativa

**Sociedad de Estadística e Investigación Operativa**

**Boletín de la SEIO**
Volumen 22, número 3
JULIO 2006



**SEIO:**
Facultad de CC. Matemáticas, Despacho 502
Universidad Complutense de Madrid
Plaza de Ciencias, 3
28040 Madrid (Ciudad Universitaria)
oficina@seio.es, http://www.seio.es
Tel: (+34) 91 544 91 02

Imprime SEROTEL
P$^o$ de la Castellana, 87.
Dep. Legal: M-13647-1995
ISSN: 1699-8871


**Normas para los envíos de colaboraciones:**

Los artículos se enviarán por correo electrónico al editor asociado correspondiente o al editor del Boletín. No deberán tener una longitud superior a 5 páginas. El formato ha de ser LaTeX, sin macros. En www.seio.es puede descargarse un modelo de artículo.

Las cartas al editor se dirigirán por correo electrónico al editor. La limitación será de 200 palabras.

El resto de colaboraciones y noticias se dirigirán al corresponsal más cercano o directamente al editor. Las referencias bibliográficas y de software se acompañarán de los datos necesarios para su localización y una reseña no superior a 120 palabras. Los resúmenes de tesis se limitarán a 200 palabras y contendrán: título, autor, directores, departamento, universidad y la fecha de lectura. Con relación a congresos y cursos bastará una breve reseña semejante a las publicadas en el Boletín. El formato preferible para estas colaboraciones es MS–Word.

# Índice



# EDITORIAL

**Carmen Alcaide Guindo**
Presidenta del Instituto Nacional de Estadística

La primera referencia a la estadística oficial española data de la creación de la Comisión Estadística del Reino, creada el 3 de noviembre de 1856 por Narváez en tiempos de Isabel II, aunque con anterioridad existen trabajos tan importantes como el Censo de Floridablanca del año 1787.

La Comisión Estadística del Reino pasó a llamarse unos meses más tarde Junta de Estadística, encomendándosele como primer trabajo la elaboración de un censo de población, que tomó como fecha de referencia el 21 de mayo de 1857. En este mismo año se establece que la Estadística sea una disciplina académica.

En 1870 se crea el Instituto Geográfico, el cual, a partir de 1873 pasa a denominarse Instituto Geográfico y Estadístico, asumiendo las tareas estadísticas.

A partir de esa fecha, los servicios oficiales de estadística fueron formando parte de distintos ministerios. Así, en sus principios la estadística estuvo adscrita al Ministerio de Fomento, posteriormente al Ministerio de Trabajo y Previsión Social, y más tarde al Ministerio de la Presidencia. Finalmente se crea el Instituto Nacional de Estadística (INE) por Ley de 31 de diciembre de 1945, con la misión de la elaboración y perfeccionamiento de las estadísticas demográficas, económicas y sociales ya existentes, y la creación de otras nuevas, así como la coordinación con los servicios estadísticos de las áreas provinciales y municipales.

Un cambio sustancial en el INE se produjo con la promulgación de la Ley de la Función Estadística Pública de mayo de 1989, que hizo del Instituto Nacional de Estadística un organismo autónomo, potenciando las nuevas tecnologías, la coordinación con las comunidades autónomas, la necesidad de la elaboración del Plan Estadístico Nacional y las relaciones con la Unión Europea en materia estadística.

Evidentemente, la demanda de información estadística por parte de la sociedad ha cambiado mucho a lo largo de los años. Al principio la estadística se dedicaba principalmente al análisis de la población y su movimiento natural. También se elaboraban censos agrarios y ganaderos, estadísticas industriales e índices del coste de la vida. Por citar alguna fecha, en 1964 el INE se hace cargo plenamente de la elaboración de la contabilidad nacional, que va incorporando la metodología que sigue la Comunidad Económica Europea.

La entrada en la UE y la evolución de la sociedad ha conllevado la necesidad de más información y con más frecuencia. La investigación estadística de todas las actividades económicas ha ido completándose a lo largo de todos estos años. Así, campos como el turismo, el comercio y los servicios en general, el I+D, el medio ambiente y todos aquellos sectores relacionados con la Sociedad de la Información han precisado en la última década de un intenso desarrollo estadístico. En el campo social, especialmente en los últimos años, la demanda de información se ha incrementado considerablemente en el conocimiento de la población inmigrante, el mercado laboral y los sectores de sanidad y educación entre otros. Esto se ha traducido en una multiplicación de encuestas dirigidas a los hogares y a la población.

Es evidente que este fuerte incremento de la producción no hubiera podido abordarse sin las nuevas tecnologías ni la capacidad de procesamiento y almacenamiento de información de que dispone el INE hoy día, unas posibilidades que hace unos años se habrían considerado ilusorias. Asimismo, esta capacidad ha hecho posible la modernización de los sistemas de trabajo y su aplicación a los proyectos estadísticos.

Sólo por nombrar un ejemplo, mencionaría que en el año 2001 se elaboró el censo utilizando cuadernos de recorrido preimpresos y con el nombre de los residentes en España y las variables básicas (fecha de nacimiento, sexo, país de nacionalidad) preimpreso en el cuestionario. Este aspecto, que para muchos pasa desapercibido, permitió que la operación censal no fuera, como en censos anteriores, una operación aislada, sino intrínsecamente relacionada con la base padronal, siendo éstos los dos grandes marcos de población existentes.

En el caso del Censo posibilitó, además, simplificar el proceso de cumplimentación de los cuestionarios (la tabla de composición del hogar iba preimpresa), la mejora de la cobertura y la imputación de la información de los hogares ubicados en viviendas que se conocía que eran principales por la operación censal, pero en la que no fue posible establecer contacto durante la misma, algo que no se había hecho en operaciones censales anteriores. Otro ejemplo de aplicación de nuevas tecnologías también relacionado con el censo 2001 es la utilización, por primera vez, de un Data Warehouse para difundir los resultados del mismo que ha tenido una excepcional acogida entre los usuarios.

Por todo ello, considero que hay un acuerdo unánime en que la componente tecnológica es vital para la Estadística y hay que seguir impulsándola pues, además de favorecer el procesamiento y difusión de la información, puede simplificar la carga de trabajo a los informantes facilitando las respuestas a cuestionarios por Internet o, incluso, permitiendo el transvase directo de información de las bases de gestión de las empresas a bases estadísticas. En esta línea, el INE es miembro de la Asociación XBRL España, la cual promueve el uso de un estándar de intercambio de información (el XBRL -eXtensible Business Reporting Language-), inicialmente financiera, extraída directamente de las bases de las empresas de acuerdo con unas taxonomías determinadas. De hecho el INE participa en una prueba piloto impulsada por la Oficina de Estadística de la Unión Europea (Eurostat).

En lo referente al uso de Internet para facilitar la colaboración de las empresas, el Instituto Nacional de Estadística diseñó en el último trimestre de 2005 un plan para ofrecer a las empresas la posibilidad de respuesta mediante ese medio. Se inició el proyecto con la Encuesta de Índice de Comercio al por Menor, siendo en la actualidad 11 encuestas las que se pueden cumplimentar por Internet. A lo largo de estos meses, el porcentaje de empresas que han utilizado esta vía ha ido creciendo progresivamente, aunque el mismo depende en gran medida de la actividad económica a estudiar y de la propia encuesta, variando el uso de Internet en las distintas encuestas entre el 17 % y el 6 % en los casos menos favorables.

A pesar de estos resultados iniciales, el INE seguirá trabajando en esta línea, ya que supone una facilidad para reducir o, al menos, facilitar la carga de trabajo a las unidades informantes, por un lado, y mejorar la calidad de los resultados eliminando intervenciones humanas en el proceso de tratamiento de los datos y contribuyendo a la reducción de los plazos de elaboración de las estadísticas.

Otro ejemplo del uso de la tecnología para fines estadísticos ha sido la implantación del CATI (Capture Automathique Telephone Interviening) el cual está siendo utilizando con gran éxito en la EPA y en otras encuestas a hogares ya que, además de facilitar la respuesta a los informantes, ha mejorado la homogeneidad de la información.

Como se desprende de estos ejemplos, ante el gran incremento de demandas de información y gracias al apoyo de las tecnologías, el INE responde investigando nuevas vías de captación de datos que garanticen la calidad de los mismos.

Centrándonos en la producción estadística en sí, hay que señalar que la estadística oficial de interés nacional se planifica a través de planes cuatrienales promulgados por medio de Reales Decretos, materializados en programas anuales (también aprobados por Real Decreto) que recogen las operaciones estadísticas que van a ser objeto de producción al año siguiente.

La entrada de España en la Unión Europea ha supuesto que la estadística oficial española en general, y el INE en particular, se hayan visto obligados a producir más cantidad de información, con más frecuencia y con unos patrones de calidad más exigentes.

Es fundamental que, para que un país sea valorado en el contexto internacional, proporcione una información estadística creíble, veraz y comparable. Por lo tanto, la sociedad no requiere solamente que se produzcan más datos; también exige que esos datos sean de calidad en el sentido más amplio de la palabra. De hecho, la definición de calidad en las estadísticas oficiales ha sido motivo de reflexión y análisis en los últimos años.

En relación con el tema de la Calidad Estadística, creo conveniente señalar dos fechas fundamentales: los años 1994 y 2005. En 1994 Naciones Unidas adoptó los principios fundamentales de las estadísticas oficiales, entre los cuales se pueden destacar dos, la **imparcialidad** que deben tener los productores de la estadística oficial y la **fiabilidad** de los datos producidos.

La otra fecha señalada en la definición del concepto de calidad en estadística oficial fue febrero de 2005, fecha en que la Unión Europea aprobó a través Eurostat las recomendaciones de la Comisión contenidas en el documento "Código de buenas prácticas de las estadísticas europeas", el cual recoge en 15 principios los principales aspectos a considerar en las estadísticas europeas.

Es conveniente resaltar que estos 15 principios se agrupan en tres grandes bloques:

- El primer bloque está relacionado con el entorno institucional y recoge los factores institucionales y organizativos que tienen una influencia considerable en la eficacia y credibilidad de la autoridad estadística que los elabora y difunde.

  En este caso, los principios recogidos sobre el entorno institucional son la independencia profesional de los responsables de la elaboración de la estadística oficial; el mandato de recogida de los datos por el que debe existir una exigencia jurídica clara para que las empresas, hogares y el público en general, colaboren en las encuestas; la adecuación de recursos; el compromiso de calidad de las instituciones; la necesidad de garantizar la confidencialidad de la información que proporcionen los informantes a la oficina de estadística, y la imparcialidad y objetividad de las personas que deben elaborar esa información.

- Un segundo bloque hace referencia a los procesos estadísticos, entendiendo por los mismos las normas, orientaciones y buenas prácticas que deben cumplirse en los procesos que llevan a cabo los que elaboran las estadísticas oficiales, necesarios para organizar, recoger, elaborar y difundir la información obtenida. Los principios descritos exigen la necesidad de utilizar una **metodología sólida y unos procedimientos estadísticos adecuados** que no supongan una carga excesiva para los encuestados y que tengan una adecuada relación coste-eficacia.

  Respecto a lo anterior conviene señalar que las exigencias de información de la Unión Europea junto a las mayores demandas de las Comunidades Autónomas, necesarias para definir sus políticas, han llevado a aumentar la carga estadística, principalmente en las empresas, detectándose una necesidad de un mayor aprovechamiento de las fuentes administrativas que pueda facilitar al menos una moderación de esa carga estadística.

- El tercer bloque tiene que ver con la producción estadística y la necesidad de satisfacer las necesidades de los usuarios, entre los que se pueden señalar instituciones europeas, gobiernos, equipos de investigación, empresas y el público en general. Los principios determinados para evaluar la producción estadística son la **pertinencia** de la información, la **precisión** y **fiabilidad** (necesidad de documentar sistemáticamente los errores de muestreo y los ajenos al muestreo), la **oportunidad** y **puntualidad**, la **coherencia** y **comparabilidad** de la información a nivel interno a lo largo del tiempo, o entre regiones o países y, por último, la **accesibilidad** y **claridad** de las estadísticas oficiales, difundiendo las mismas de forma imparcial, con metadatos y orientación de apoyo.

Lo fundamental de estos principios es que no se limitan exclusivamente al ámbito de recomendaciones sino que las instituciones responsables de producir estadísticas para la Unión Europea van a ser evaluadas, siguiendo estos principios, por personas independientes de otras instituciones o países europeos.

Por lo tanto, tal y como se deduce de lo anterior, los elementos a considerar en la estadística oficial son muchos y de muy diversa índole, mereciendo ser destacado el uso de metodologías apropiadas, que nos permitan obtener información más fiable y con menos carga para los informantes.

Así, existen temas en los que debemos poner especial interés en su investigación, como por ejemplo, el de las estimaciones en áreas o dominios pequeños, el aprovechamiento de las fuentes administrativas, la resolución sistemática de los problemas asociados a la confidencialidad de la información, la mecanización de los procesos de imputación y depuración para cualquier encuesta tanto en sus variables cuantitativas como en las cualitativas, la utilización de estimadores asistidos o basados en modelos.

Aunque se podría señalar otras áreas de investigación, solo con las ya mencionadas queda patente que la colaboración entre el INE y la comunidad científica no sólo es necesaria sino imprescindible.

El beneficio de esta colaboración será mutuo. Para el INE, porque ayudará a resolver problemas planteados en la actualidad y asumir nuevos retos, y para la Universidad, porque se le brinda la oportunidad de acceder a una gran variedad de datos del mundo real y poder tomar contacto con problemas desconocidos para ellos, pudiendo contrastar la validez práctica de sus investigaciones teóricas. Así, si bien los lazos entre la Universidad y el Instituto Nacional de Estadística son cada vez mayores, se considera indispensable no descuidar esta cooperación y, si es posible, impulsarla en beneficio de la estadística.

Como prueba de estas relaciones se encuentra la participación del INE como miembro de la SEIO, la asistencia cada vez mayor de los expertos del mundo académico y científico a los foros que el INE tiene con los usuarios de la información, o los convenios de colaboración que se firman entre el INE y determinadas universidades con el objeto de investigar problemas concretos que se han planteado en el proceso de producción de las estadísticas oficiales. También quisiera recordar, que el INE brinda a todas las personas interesadas la posibilidad de trabajar en esta institución, pues de forma regular se incluye en la oferta pública de empleo un número significativo de plazas tanto para el Cuerpo Superior de Estadísticos del Estado, como para el Cuerpo de Diplomados en Estadística del Estado, que les permitirá desarrollar la carrera profesional dentro del INE.

Para finalizar, quiero agradecer a la SEIO la oportunidad que me ha brindado de participar en esta revista a través del editorial y manifestar una vez más mi deseo e ilusión en que nuestras instituciones colaboren estrechamente para mejorar la estadística, y en especial la estadística oficial.

Madrid, 21 de junio de 2006.

# 1. ARTÍCULOS DE ESTADÍSTICA

## ESTADÍSTICA NO PARAMÉTRICA: PASADO, PRESENTE Y FUTURO

**Wenceslao González Manteiga y Rosa María Crujeiras Casais**
Departamento de Estatística e Investigación Operativa
Universidade de Santiago de Compostela

**Resumen**

La estadística no paramétrica engloba una serie de técnicas de inferencia cuya característica principal es la ausencia de un modelo paramétrico de distribución subyacente.

Sin pretender revisar todos y cada uno de los conceptos e ideas que conforman este planteamiento, es nuestro objetivo que el lector se familiarice con términos como *técnicas de suavizado*, *Bootstrap*, *verosimilitud empírica* o *datos funcionales*, su porqué, su utilidad, su relevancia en la estadística actual y la posible proyección hacia el futuro.

### 1. De la distribución empírica a las técnicas de suavizado.

Gran parte de la inferencia estadística paramétrica (Bayesiana o no) trabaja bajo el supuesto de disponer de una muestra aleatoria simple (m.a.s.) $\vec{X} = (X_1, \ldots, X_n)$ de una variable poblacional $X$. A esta variable poblacional se le supone una distribución $F_\theta$ con $\theta$ un parámetro desconocido finito dimensional. Los diversos métodos de estimación puntual de $\theta$, la construcción de regiones de confianza o los contrastes de hipótesis para dicho parámetro, fueron abordados a lo largo del siglo XX, y resumidos en su parte esencial, en textos como los *Lehmann* ([14] y [15], en sus ediciones revisadas) o los más recientes [24] y [27], entre otros. Estos textos se utilizan además como referencia en cursos de estadística matemática.

En este contexto, la especificación errónea de la distribución poblacional $F_\theta$, o incluso, su total desconocimiento, propiciaron el desarrollo de la inferencia estadística no paramétrica. No se presupone un modelo paramétrico para la distribución poblacional y simplemente se actúa con la filosofía de dejar hablar a los datos. En el enfoque no paramétrico, la distribución $F$ de la variable poblaciónal no está caracterizada por un parámetro finito dimensional.

Sin la suposición precisa de un modelo paramétrico de distribución poblacional para la variable de interés, nos encontramos en una encrucijada en la que el problema de inferencia puede ser abordado desde dos vertientes: (a) considerar, como ya se comentó, que el espacio de posibles distribuciones $\mathcal{F}$ no se puede parametrizar de forma finito dimensional, o bien (b) examinar alguna distancia $d(F, F_\theta)$ entre la distribución poblacional desconocida y una posible distribución paramétrica para el modelo poblacional.

La vía (a) dio origen a varios procedimientos de estimación, como los M-estimadores, L-estimadores o R-estimadores. Estos estimadores conforman la extensión de los métodos de estimación de máxima verosimilitud, en el caso de los M-estimadores; de los estadísticos de orden, para los L-estimadores; y de los estadísticos de rango, para los R-estimadores. El objetivo en este caso es hacer inferencia sobre una cierta característica $T(F)$, donde $F$ es la distribución poblacional y $T$ un funcional definido sobre el espacio de distribuciones. Por ejemplo, el libro de Serfling, [22], recoge un resumen didácticamente ordenado del material elaborado en las décadas de los sesenta y setenta. Otros textos coetáneos, como [11] o [13] estudiaron el llamado análisis de la robustez que permitía estudiar la sensibilidad de los diversos estimadores a la desviación de un modelo poblacional.

Por otro lado, los llamados tests de bondad de ajuste nacen del enfoque (b) y han evolucionado desde sus inicios con el más básico Kolmogorov-Smirnov, donde $d(\hat{F}, F_{\hat{\theta}}) = \sup |\hat{F}(x) - F_{\hat{\theta}}(x)|$, $\hat{F}$ es un estimador no paramétrico de la función de distribución y $\hat{\theta}$ un estimador para $\theta$, a distintos tests alternativos a este (ver, por ejemplo [8]); como los tests del tipo Cramer Von Mises o de tipo $\chi^2$, por ejemplo. Aunque con diferentes planteamientos, el objetivo es siempre ubicar el modelo de distribución $F$ en una determinada familia paramétrica.

En la mayoría de los procedimientos de estimación o contrastes antes descritos interviene con frecuencia un estimador piloto no paramétrico. En el caso del test de Kolmogorov-Smirnov, este estimador es la la distribución empírica:

$$\hat{F}(x) = F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \le x\}}. \tag{1}$$

Este mismo estimador $F_n$ aparece, por ejemplo, en los métodos de M-estimación, donde un M-estimador es la solución a una ecuación del tipo:

$$\int \Psi(x,t)dF_n(x) = 0, \tag{2}$$

siendo $\Psi$ una función elegida de antemano para el estimador correspondiente.

Una de las principales dificultades en la inferencia no paramétrica surge cuando se pretende estimar cantidades $T(F)$ que dependen de la derivada de $F$, sin ir más lejos:

$$T(F) = F'(x) = f(x) = \lim_{h\to 0} \frac{F(x+h) - F(x-h)}{2h}, \tag{3}$$

la función de densidad asociada a la variable poblacional. La estimación de la densidad dio lugar a las denominadas *técnicas de suavizado*. Es así que un estimador natural de (3) es el que viene dado por

$$\hat{f}_h(x) = \frac{F_n(x+h) - F_n(x-h)}{2h}, \quad \text{con } h \to 0 \tag{4}$$

el llamado estimador de Rosenblatt [21], generalizado por Parzen en [18] a los populares *estimadores tipo núcleo*:

$$\tilde{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{5}$$

donde la función núcleo o *kernel* $K$ es una función de densidad y $h$ se denomina parámetro ventana. El estimador de Rosenblatt se corresponde con la elección de un núcleo uniforme. Una revisión de las técnicas no paramétricas para la estimación de la densidad puede verse en [23], que representa uno de los primeros textos, de una larga lista, dedicados a las técnicas de suavización publicados en los últimos veinte años.

Las técnicas de suavización no paramétrica, que surgen en el contexto de la estimación de la densidad, rápidamente se extendieron al problema de estimación de otras funciones estadísticas notables, como la función de razón de fallo en fiabilidad $r(x) = \dfrac{f(x)}{1 - F(x)}$, la función de regresión $m(x) = E(Y|X = x)$ ligada a un vector $(X,Y) \in \mathbb{R}^{q+1}$ ó la varianza condicional $\sigma^2(x) = Var(Y|X = x)$.

Para el caso de la función de regresión, donde la selección de un modelo paramétrico no adecuado puede llevarnos a conclusiones incorrectas, el primer estimador tipo núcleo es el de Nadaraya-Watson (introducido en [16] y [26]), que viene dado por:

$$\tilde{m}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)}, \tag{6}$$

donde la concentración local de datos propia del estimador de la densidad (5) se sustituye por el promedio local de la variable respuesta.

También existen otros estimadores tipo núcleo, como el de Priestley-Chao o el de Gasser-Muller (véase [12]). Una excelente revisión de estas técnicas de suavización, tanto para el caso de la densidad como para la regresión, puede verse en [25]. En este texto también se incluyen diversos capítulos dedicados al problema de la selección del parámetro $h$ de suavizado.

Los procedimientos de suavización también se abordaron desde espacios funcionales alternativos al de las funciones derivables.

Si la función de regresión puede escribirse como $m = \sum_{k=1}^{\infty} c_k(m)\phi_k$, donde $\{\phi_k\}_{k=1}^{\infty}$ es un sistema ortogonal completo de $\mathcal{L}_2$ y $c_k(m)$, $k = 1, 2, \ldots$ los coeficientes de Fourier de $m$, un estimador suavizado natural viene dado por:

$$\widehat{m^O} = \sum_{k=1}^{\hat{M}} \hat{c}_k \phi_k \tag{7}$$

siendo $\hat{c}_k$ los estimadores de los coeficientes de Fourier y $\hat{M}$ el número de coeficientes estimados (ver [2]).

Otra alternativa es la suavización spline $\widehat{m^S}$, donde dicho estimador viene dado por el argumento que minimiza

$$\min_m \frac{1}{n}\sum_{i=1}^{n}(Y_i - m(X_i))^2 + \lambda \int_0^1 (m^{(p)}(x))^2 dx,$$

y cuya solución es una función del espacio de Sobolev $W_2^p[0,1]$ (ver [10]). El caso particular de $p = 2$ se corresponden con el conocido *spline cúbico*. También de forma alternativa a los estimadores clásicos tipo núcleo, tuvieron éxito los estimadores de tipo lineal local [6] y más recientemente, de manera alternativa a los estimadores basados en desarrollos ortogonales, los estimadores wavelet [2].

**2. El Bootstrap. Todavía actualidad.**

En un artículo de *Annals of Statistics* en 1979, [3], Efron introdujo una metodología que alcanzaría su mayor explendor años después, con el desarrollo computacional: nos referimos al *Bootstrap*.

El Bootstrap es una metodología de inferencia estadística basada en el remuestreo de los datos y cuyo principal inconveniente (aunque no tanto hoy en día) es la necesidad de computación intensiva. Dada una m.a.s. $\vec{X}$ de una variable aleatoria $X$, denotaremos por $\vec{X}^*$ a las distintas remuestras que se puedan obtener a partir de una estimación de la distribución poblacional. Si estas remuestras se generan a partir de la distribución empírica $F_n$, el procedimiento se denomina Bootstrap Naive; si la generación de datos se hace a partir de un estimador núcleo

$$\hat{F}_h(x_0) = \int_{-\infty}^{x_0} \hat{f}_h(x)dx$$

tendremos un Bootstrap Suavizado; también existen las alternativas de Bootstrap Paramétrico, Simetrizado, Bootstrap Bayesiano, entre otras opciones.

Consideremos $R(\vec{X}, F) = T(\vec{X}) - T(F)$, siendo $T(\vec{X})$ un estimador de $T(F)$. Si se tiene como objetivo estimar la distribución de $R(\vec{X}, F)$ dada por $G(t) = \mathbb{P}_F\{R(\vec{X}, F) \le t\}$, el estimador natural Bootstrap es $\hat{G}^{Boot} = \mathbb{P}_{\hat{F}}\{R(\vec{X}^*, \hat{F}) \le t\}$. El estimador Bootstrap $\hat{G}^{Boot}$ raramente tiene expresión explícita y debe ser aproximado por Monte Carlo:

$$\widehat{\hat{G}^{Boot}(t)} = \frac{1}{B}\sum_{j=1}^{B} \mathbf{1}_{\{R(\vec{X}^{*j}, \hat{F}) \le t\}}, \tag{8}$$

siendo $\vec{X}^{*j}$, con $j = 1, \dots, B$, remuestras artificales de la distribución $\hat{F}$, estimador de $F$.

La elección que hemos hecho de $R$ es la más sencilla y permitiría inferir sobre la distribución de $T(\vec{X})$ como estimador de $T(F)$, pudiendo así construirse regiones de confianza o plantear contrastes de hipótesis. La adaptación de la metodología Bootstrap a otros problemas notables, como pueden ser el cálculo de la probabilidad de clasificación incorrecta en el análisis discriminante o la construcción de bandas de confianza para una función de distribución, nos llevarían a distintas elecciones de $R$, que conforman la literatura dedicada a este método en las dos últimas décadas.

Igual que el Barón de Munchausen salió del lago tirando de los cordones de sus botas (en inglés *bootstraps*), del mismo modo Efron presenta una metodología que se apoya en los datos de la muestra observada, para sobre esas observaciones, construir nuevas muestras.

En los últimos quince años han aparecido diversas publicaciones dedicadas a la metodología Bootstrap, como [1] y [5], aunque ya más recientemente, el Bootstrap representa uno o varios capítulos de los textos sobre estadística matemática ([24] o [28]) en los cursos de licenciatura y doctorado.

**3. La verosimilitud empírica: la verosimilitud de finales del siglo XX.**

Volviendo al planteamiento inicial de considerar una m.a.s. $\vec{X}$ de una variable aletoria $X$ con distribución $F_\theta$ y densidad o masa de probabilidad asociada $f_\theta$, si planteamos la función de verosimilitud $\psi(\theta) = \prod_{i=1}^{n} f_\theta(X_i)$, el estimador máximo verosímil $\hat{\theta}$ para el parámetro $\theta$, se obtiene como el argumento que maximiza $\psi(\theta)$. Este método fue estudiado a lo largo del siglo pasado y el principio de verosimilitud sigue presente en textos actuales, como [19], entre otros muchos.

Nuevamente nos encontramos ante el supuesto de una distribución paramétrica para la variable de interés y una vez más, desde el enfoque no paramétrico, se aborda este planteamiento de máxima verosimilitud sin presuponer un modelo poblacional para $X$. De este modo, cuando no se sabe nada sobre el modelo poblacional, uno puede considerar una verosimilitud de tipo no paramétrico: $\Psi(F) = \prod_{i=1}^n (F(X_i) - F(X_i^-)) = \prod_{i=1}^n \omega_i$, $\vec{\omega} = (\omega_1, \dots, \omega_n)$. De forma natural, se deduce que el estimador de máxima verosimilitud no paramétrico se alcanza en $F = F_n$, dada por (1).

La idea de la *verosimilitud empírica* (o verosimilitud no paramétrica) es reciente, introducida por Owen a finales de los ochenta. Gran parte de sus desarrollos en la década de los noventa se recogen en [17]. La clave del éxito de esta técnica y su creciente desarrollo en los últimos años radica en su flexibilidad y versatilidad para incorporar en la verosimilitud las restricciones que impone el modelo. Así, como ejemplos notables:

1) Si sabemos que $\mu_0 = \int x dF(x)$ es la media (conocida) de $X$, el estimador máximo verosímil para $F$ bajo esta restricción, se obtendría como el argumento que maximiza:

$$\max_{\vec{\omega}} \left\{ \prod_{i=1}^n \omega_i : \sum \omega_i X_i = \mu_0, \omega_i \geq 0, \sum \omega_i = 1 \right\}.$$

2) Si suponemos además que la varianza de $X$ es $\sigma_0^2 = \int (x - \mu_0)^2 dF(x)$; el estimador de máxima verosimilitud vendría dado por:

$$\max_{\vec{\omega}} \left\{ \prod_{i=1}^n \omega_i : \sum \omega X_i = \mu_0, \sum \omega_i (X_i - \mu_0)^2 = \sigma^2, \right.$$
$$\left. \omega_i \geq 0, \sum \omega_i = 1 \right\}.$$

De la misma forma que, para el contraste de $H_0 : \theta \in \Theta_0$ vs. $H_a : \theta \notin \Theta_0$, con $\Theta_0 \subset \Theta$, ligado a la clásica razón de verosimilitudes

$$\lambda = \frac{\sup_{\theta \in \Theta_0} \prod_{i=1}^n f_\theta(X_i)}{\sup_{\theta \in \Theta} \prod_{i=1}^n f_\theta(X_i)} \tag{9}$$

tenemos el test de razón de verosimilitudes, con región crítica $\{-2 \log \lambda > c\}$, lo mismo se puede considerar para la verosimilitud empírica. Supongamos que se pretende contrastar $H_0 : \mu_0 = \int x dF(x)$. El test de razón de verosimilitudes no paramétrico vendría dado por:

$$R = \frac{\max\{\prod_{i=1}^n \omega_i : \sum \omega_i X_i = \mu_0, \omega_i \geq 0, \sum \omega_i = 1\}}{\max\{\prod_{i=1}^n \omega_i : \omega_i \geq 0, \omega_i = 1\}}$$

$$= \max \left\{ \prod_{i=1}^n n\omega_i : \sum \omega_i X_i = \mu_0, \omega_i \geq 0, \sum \omega_i = 1 \right\}$$

Igual que para la razón de verosimilitudes paramétrica, resultados del tipo $-2 \log R \to^d \chi_m^2$ también se obtienen en este contexto, y en el caso particular anterior resulta $m = 1$.

En el reciente libro de Owen antes citado, [17], se presenta la metodología de verosimilitud empírica en su conjunto para el contexto de regresión, para datos incompletos o con censura, entre otras situaciones notables de la inferencia estadística.

**4. La estadística con datos funcionales. La era de los datos de alta dimensión.**

La explosión del llamado *data mining* (también conocido como KDD: Knowledge-Discovery in Databases), está produciendo una variedad de situaciones, en diferentes contextos, en que los datos medidos son realmente curvas. Por ejemplo, el nivel de ozono medido durante un día o un indicador de alta frecuencia de un activo financiero en la bolsa a lo largo de una hora. De esta forma, el dato observado es ahora $\chi = \{X(t) : t \in (t_{min}, t_{max})\}$ y podríamos considerar una muestra de *datos funcionales* $\chi_1, \dots, \chi_n$ que tuviese la misma distribución que $\chi$.

De modo más genérico, una variable aleatoria $\chi$ se dice que es una variable funcional si toma valores en un espacio de dimensión infinita. Con este tipo de planteamiento entramos en la era actual y futura de la *estadística no paramétrica con datos funcionales*. Por ejemplo, supongamos que el soporte de la variable funcional es $L^2(0, 1)$ y consideremos un modelo de regresión del tipo: $Y = \mathfrak{m}(\mathcal{X}) + \varepsilon$, en el que $Y \in \mathbb{R}$ y $\varepsilon$ es una v.a. real de media cero. El objetivo podría ser estimar no paramétricamente el funcional $\mathfrak{m}(\cdot) : L^2(0, 1) \to \mathbb{R}$.

Las técnicas estadísticas para datos reales o vectoriales no se extienden de forma trivial al contexto de datos funcionales. Si tenemos un modelo de regresión no paramétrico $Y = m(X) + \varepsilon$ y consideramos un estimador no paramétrico del tipo Nadaraya-Watson, como en (6) nos puede interesar contrastar la hipótesis nula de que el modelo es lineal $m(X) = \beta^T X$. Esto se podría hacer a través de cualquier estadístico del tipo $d(\hat{m}(X)\cdot\hat{\beta}^T X)$, que establece una distancia entre el estimador no paramétrico y el paramétrico, por ejemplo, una distancia de tipo $L_2$ (ver [9]).

En el supuesto de datos funcionales el modelo lineal es ahora del tipo $\mathfrak{m}(\mathcal{X}) = \int_0^1 \beta(t)\mathcal{X}(t)dt$ y por tanto también de dimensión infinita, ya que lo desconocido $\beta$ es también una función. Podemos construir un estimador no paramétrico del funcional $\mathfrak{m}$, análogo al de Nadaraya-Watson del contexto de regresión, que vendría dado por:

$$\hat{\mathfrak{m}}(\chi) = \frac{\sum_{i=1}^n K\left(\frac{d(\chi,\chi_i)}{h}\right) Y_i}{\sum_{i=1}^n K\left(\frac{d(\chi,\chi_i)}{h}\right)} \tag{10}$$

donde $d$ es una semimétrica asociada al espacio de funciones y $K$ es una función núcleo asimétrica.

En un contexto finito dimensional, en la distribución del estadístico $d(\hat{m}(X), \hat{\beta}^T X)$, la estimación paramétrica puede tener menos peso que la de $\hat{m}$, ya que sus tasas de convergencia hacia la función teórica son más rápidas. Pero ¿qué ocurre si la distancia se mide en el contexto de datos funcionales? Ambos parámetros, $\mathfrak{m}$ y $\beta$ son de dimensión infinita. Problemas similares a este merecen sin duda el interés de la futura metodología estadística.

Siguiendo esta línea de extensión de las técnicas de regresión no paramétrica a los datos funcionales, [7] es una buena referencia. Un planteamiento alternativo para el estudio de datos funcionales, muy próximo a la metodología de la estimación spline, se recoge en [20].

## 5. Algunas aplicaciones.

Sin tratar de ser exhaustivos, finalizamos esta nota comentando algunas aplicaciones donde la inferencia no paramétrica juega un papel importante.

Para aproximar la fórmula de Black-Scholes al comportamiento del mercado, se hace necesario modelizar la volatilidad de manera heterocedástica y por tanto, se requiere la estimación no paramétrica de la varianza.

En los modelos de predicción kriging de la estadística espacial, la suposición de estacionariedad puede resultar restrictiva para llevar a cabo predicciones en problemas de medio-ambiente. Los modelos de predicción han de adaptarse de forma local, y esta adaptación requiere de técnicas de suavización no paramétricas.

Dentro de la estimación de conjuntos, puede plantearse como objetivo estimar el soporte de una cierta variable $X$, con distribuciónn $F$ desconocida. Una estimación del soporte vendría dada por aquellos puntos donde una estimación no paramétrica de la densidad tomase valores positivos. Este planteamiento podría tener aplicación directa en el reconocimiento de imágenes, entre otras situaciones.

En estos ejemplos, junto con los ya comentados en el apartado dedicado a la estadística con datos funcionales, ¿tendría sentido el suponer un modelo paramétrico? La respuesta no es evidente. Pero lo que sí queda claro es que el planteamiento no paramétrico puede aportar una solución.

## 6. Conclusiones.

Como se deduce de este recorrido por la inferencia no paramétrica, las razones del porqué de este modo de hacer inferencia son claras. La metodología no paramétrica es, ante todo, flexible, capaz de modelizar sin suposiciones paramétricas restrictivas y sobre todo, rica y elegante desde el punto de vista matemático. Es esta flexibilidad, junto con el desarrollo computacional, lo que ha permitido el éxito de muchas de estas técnicas en los últimos años. Como ejemplo más claro, no sólo la multitud de artículos y referencias en la literatura, sino también sus numerosas aplicaciones en modelos de interés.

Efron, [4], hace una reflexión sobre los dos modos de hacer estadística, todavía enfrentados. La estadística bayesiana dominó el siglo XIX, mientras que la estadística del siglo XX fue más frecuentista. ¿Hacia dónde va la estadística del siglo XXI?

Como Efron sugiere, los nuevos problemas a los que se enfrenta la estadística, sólo podrán ser resueltos combinando ambas filosofías. Desde nuestra posición no creemos osado afirmar que uno de los nexos de unión entre estos enfoques podría ser la estadística no paramétrica.

**Referencias**


[1] Davison, A. C. and Hinkley, D.V. (1997). *Bootstrap Methods and their Application.* Cambridge University Press.

[2] Efromovich, S. (1999). *Nonparametric Curve Estimation. Methods, Theory and Applications*, Springer.

[3] Efron, B. (1979). Bootstrap methods: another look at the jackniffe. *Ann. Stat.*, **7**, 1-26.

[4] Efron, B. (2004). Bayesians, frequentists and scientists. *Text of the 164th ASA presidential address.*

[5] Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap.* Chapman Hall.

[6] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*, Chapman Hall.

[7] Ferraty, F. and Vieu, P. (2006). *Nonparametric Modelling for Functional Data.* Springer.

[8] Gibbons, J.D. and Chakraborti, Sl. (1992). *Nonparametric Statistical Inference. Third Edition, Revised and Extended*, Marcel Dekker.

[9] González-Manteiga, W. and Cao, R. (1993). Testing the hypothesis of a general linear model using non parametric regression estimation. *Test*, **2**, 161-188.

[10] Gu, C. (2002). *Smoothing Spline Anova Models*, Springer.

[11] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). *Robust Statistics: the Approach Based on Influence Functions*, Wiley.

[12] Härdle,W. (1990). *Applied Nonparametric Regression.* Cambridge.

[13] Huber, P.J. (1981). *Robust Statistics*, Wiley.

[14] Lehmann, E.L. (1991). *Theory of Point Estimation*, The Wadsworth & Brooks.

[15] Lehmann, E.L. (1991). *Testing Statistical Hypothesis*, The Wadsworth & Brooks (2nd edition).

[16] Nadaraya, E.A. (1964). On estimating regression. *Theor. Prob. Appl.*, **10**, 186-190.

[17] Owen, A.B. (2001). *Empirical Likelihood.* Chapman & Hall.

[18] Parzen, E. (1962). On the estimation of a probability density function and the mode. *Ann. Math. Stat.*, **33**, 1065-1076.

[19] Pawtan, Y. (2001). *In all Likelihood: Statistical Modelling and Inference Using Likelihood.* Oxford Science Publications.

[20] Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis.* Springer, 2nd edition.

[21] Rosenblatt, M. (1956). Remarks on some nonparametric esitmates of a density function. *Ann. Math. Stat.*, **27**, 832-837.

[22] Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*, Wiley.

[23] Silverman, B. (1985). *Density Estimation for Statistics and Data Analysis*, Chapman & Hall.

[24] Shao, J. (2003). *Mathematical Statistics*, Springer (2nd edition).

[25] Wand, M. P. and Jones, M.C. (1995). *Kernel Smoothing*, Chapman & Hall.

[26] Watson, G.S. (1964). Smooth regression analysis. *Sankhya, Ser. A*, **26**, 101-116.

[27] Wasserman, L. (2004). *All of Statistics. A Concise Course in Statistical Inference*, Springer.

[28] Wasserman, L. (2006). *All of Nonparametric Statistics*, Springer.

# 2. ARTÍCULOS DE INVESTIGACIÓN OPERATIVA

## ESTABILIDAD EN PROGRAMACIÓN SEMI-INFINITA LINEAL: UN ENFOQUE CUANTITATIVO

**María Josefa Cánovas Cánovas y Juan Parra López**
Dpto. de Estadística, Matemáticas e Informática
Centro de Investigación Operativa
Universidad Miguel Hernández de Elche

### 1. Introducción

El objetivo de este artículo es ofrecer al lector una breve incursión en el campo de la programación lineal con infinitas restricciones, resaltando las diferencias con la programación lineal ordinaria, y adentrándonos fundamentalmente en cuestiones de estabilidad y su cuantificación.

El término Programación Semi-Infinita (PSI) se aplica a problemas de optimización en los que bien el número de variables o bien el de restricciones es finito. Se insiste así en que este concepto abarca como caso particular a los problemas con $n$ variables y $m$ restricciones. Prestaremos especial atención a los problemas de PSI lineal (PSIL). Si llamamos primal a un problema con $n$ variables y (posiblemente) infinitas restricciones, entonces su dual tendrá (posiblemente) infinitas variables y $n$ restricciones.

Los problemas de programación semi-infinita lineal surgen de manera natural al modelar situaciones, o reformular modelos, en diferentes campos de aplicación, como la aproximación funcional, estimación de parámetros, reconocimiento de patrones, políticas medioambientales, etc. El lector interesado no debería perderse la monografía de Miguel A. Goberna y Marco A. López [12] sobre PSIL. Asimismo recomendamos la lectura del reciente artículo de revisión de Marco A. López y Georg Still [15] sobre PSI, en el que se describen fundamentos teóricos (incluyendo condiciones de optimalidad), métodos numéricos (como los de discretización), diversas aplicaciones y una revisión histórica del tema. En concreto se presentan aplicaciones de la PSI a la aproximación de Chebyshev, física matemática, robótica, geometría, optimización bajo incertidumbre (optimización robusta) y economía. En [15] se citan además ciertas aplicaciones a la estadística (diseño de experimentos, regresión, estimación por máxima verosimilitud con restricciones, robustez en estadística bayesiana, estadística actuarial, etc.). Véanse, p.e., [8] y [12].

A modo de ilustración, describiremos muy brevemente un par de aplicaciones de la PSIL: la aproximación de Chebyshev (uniforme) por funciones polinómicas y la optimización robusta aplicada al problema de la cartera.

Imaginemos que buscamos un polinomio de grado menor o igual que $n$ en la variable $t$, pongamos $p(t) = x_0 + x_1 t + ... + x_n t^n$, que constituya una mejor aproximación uniforme de una función dada $f$ en $[a, b]$. En principio no requerimos ninguna propiedad a $f$. Dicho problema es equivalente al siguiente problema de PSIL en las variables $x_0, x_1, ..., x_n, x_{n+1}$ (donde escribimos Inf en vez de Min):

$$\begin{array}{ll} \text{Inf} & x_{n+1} \\ \text{s. a} & \pm (x_0 + t x_1 + ... + t^n x_n) - x_{n+1} \\ & \leq \pm f(t) \;\; \forall t \in [a, b]. \end{array}$$

Consideremos ahora el problema de la cartera: Queremos invertir un capital $k > 0$ por un periodo de un año en $n$ bienes cuyo rendimiento (por unidad monetaria invertida) al cabo del citado año denotamos por $t_1, ..., t_n$, respectivamente. Ciertamente existe incertidumbre sobre el valor de $t = (t_1, ..., t_n)$. De no ser así, nuestra decisión óptima consistiría en invertir todo el capital en uno de los bienes (si hay varios) con rendimiento máximo. Supongamos que, basándonos en la experiencia y ciertos modelos económicos podemos predecir (con un cierto margen de confianza) que $t$ variará en un subconjunto $T \subset \mathbb{R}^n$; entonces el enfoque pesimista de la situación nos llevará a buscar el beneficio máximo en el peor de los casos, lo que conduce al problema de PSIL (llamado de optimización robusta):

$$\begin{array}{ll} \text{Sup} & v \\ \text{s. a} & t_1x_1+...+t_nx_n-v\geq 0\ \forall(t_1,...,t_n)\in T \\ & x_1+...+x_n=k,\ x_1\geq 0,...,x_n\geq 0. \end{array}$$

La PSI ha despertado el interés de diferentes grupos de investigación desde su inicio, a principios de los años 60, con una serie de artículos (sobre PSIL) de Charnes, Cooper y Kortanek, tanto por los atractivos desarrollos teóricos a los que da lugar como por la variedad de sus aplicaciones. Los citados autores acuñaron el término *semi-infinita*. Desde entonces se han escrito más de mil artículos y diez libros sobre teoría, métodos numéricos y aplicaciones de la PSI. Los antecedentes de esta disciplina se encuentran, entre otros, en la aproximación de Chebyshev, el trabajo de Haar sobre sistemas (semi-infinitos) lineales (1924) y las condiciones de optimalidad de Fritz John (1948). Cabe destacar que la PSI constituye un campo de investigación al que se puede acceder desde diferentes áreas matemáticas, como el análisis, la geometría diferencial, la topología, el cálculo numérico, y por supuesto la investigación operativa.

Son numerosos los investigadores internacionales que han dedicado parte de sus esfuerzos a la PSI (lineal y no lineal). Entre ellos citaremos a Brosowski, Fischer, Jongen y Stein en Alemania, Klatte en Suiza, Still en Holanda, Jiménez, Rückmann y Todorov en México, Vera de Serio en Argentina, Rubinov en Australia, Weber en Turquía, Vaz y Fernandes en Portugal, etc. A nivel nacional, y pidiendo disculpas por adelantado por las involuntarias omisiones, el tema tuvo un desarrollo inicial con Marco A. López y sus entonces doctorandos Miguel A. Goberna, Jesús T. Pastor y poco después Enriqueta Vercher. Esta última ha liderado un grupo en Valencia (Teresa León y Susana Sanmatías) que ha avanzado, entre otros, en el campo de los algoritmos y las aplicaciones. Miguel A. Goberna y Marco A. López lideran el grupo de Alicante (Valentín Jornet, Margarita Rodríguez, Mariola Molina y Mª Dolores Fajardo), mientras que los arriba firmantes, de la mano de Marco A. López, encabezan el grupo de Elche (junto con Fco. Javier Toledo, Francisco J. Gómez-Senent y colaboraciones puntuales de Eva Mª Ortega). Entre las aproximaciones al tema desde áreas afines citaremos a Juan E. Martínez Legaz y Albert Ferrer en Barcelona, Beatriz Hernández en Huelva, y Natividad Llorca y Joaquín Sánchez Soriano en Elche.

Tras la imprescindible dosis de notación, empezaremos resaltando algunas diferencias notables con el campo de la programación lineal (PL) ordinaria. Un problema típico de PSIL puede formularse como

$$\pi:\quad \begin{array}{l} \text{Inf } c'x \\ \text{s. a } a_t'x\geq b_t,\ \ t\in T, \end{array} \tag{1}$$

donde $x\in\mathbb{R}^n$ es el vector de variables de decisión, considerado como matriz columna, $y'$ denota el traspuesto de $y\in\mathbb{R}^n$, $T$ es un conjunto de índices completamente arbitrario (posiblemente infinito), y la función $T\ni t\mapsto\binom{a_t}{b_t}\in\mathbb{R}^n\times\mathbb{R}$ es igualmente arbitraria. Denotamos por $\sigma$ al sistema de restricciones de $\pi$, esto es,

$$\sigma:=\{a_t'x\geq b_t,\ \ t\in T\}\,. \tag{2}$$

El problema $\pi$ puede identificarse con el par $(c,\sigma)$ perteneciente al espacio paramétrico $\Pi=\mathbb{R}^n\times\Theta$, donde $\Theta\equiv(\mathbb{R}^n\times\mathbb{R})^T$ es el espacio paramétrico de los sistemas del tipo (2), identificando $\sigma$ con $\binom{a_t}{b_t}_{t\in T}$. Asociados a $\pi=(c,\sigma)$ consideramos los siguientes elementos:

$F:=\{x\in\mathbb{R}^n:a_t'x\geq b_t \text{ para todo } t\in T\}$, el *conjunto factible* de $\pi$ (o de $\sigma$);

$v:=\inf\{c'x:x\in F\}\in[-\infty,+\infty]$, el *valor óptimo* de $\pi$ ($v=+\infty$ si $F=\emptyset$);

$F^*:=\{x\in F:c'x=v\}$, el *conjunto óptimo* de $\pi$.

Asimismo, denotaremos por $\mathcal{F}$,$\mathcal{F}^*:\Pi\rightrightarrows\mathbb{R}^n$ a las multifunciones (correspondencias) conjunto factible y conjunto óptimo, respectivamente, y por $\vartheta:\Pi\longrightarrow[-\infty,+\infty]$ a la función valor óptimo.

Cuando consideremos diferentes problemas, éstos y sus elementos asociados se distinguirán por medio de sub(super)índices. Así, si $\pi_1\in\Pi$ escribiremos $\pi_1=\left(c^1,\sigma_1\right)$, donde $\sigma_1:=\left\{\left(a_t^1\right)'x\geq b_t^1,\ t\in T\right\}$, y el conjunto factible, valor óptimo y conjunto óptimo de $\pi_1$ se representarán respectivamente por $F_1$, $v_1$ y $F_1^*$.

Consideramos los siguientes subconjuntos relevantes de $\Pi$ y de $\Theta$:

$\Pi_c:=\{\pi=(c,\sigma)\in\Pi:\sigma\in\Theta_c\}$, donde $\Theta_c:=\{\sigma\in\Theta:F\neq\varnothing\}$;

$\Pi_b:=\{\pi=(c,\sigma)\in\Pi:v \text{ es finito}\}$;

$\Pi_s:=\{\pi=(c,\sigma)\in\Pi:F^*\neq\varnothing\}$;

$\Pi_c^d:=\left\{\pi=(c,\sigma)\in\Pi:\pi^d \text{ es consistente}\right\}$.

Obviamente $\Pi_s \subset \Pi_b \subset \Pi_c$. Aquí los subíndices $c$, $b$ y $s$ significan respectivamente *consistente, acotado* y *resoluble*, mientras que el superíndice $d$ hace referencia al problema dual de $\pi$ (en el sentido de Haar), que viene dado por

$$\begin{aligned} \pi^d: \quad & \text{Sup} \sum_{t \in T} \lambda_t b_t \\ & \text{s.a} \sum_{t \in T} \lambda_t a_t = c, \\ & \qquad \lambda \in \mathbb{R}_+^{(T)}, \end{aligned} \tag{3}$$

donde $\mathbb{R}_+^{(T)}$ es el cono convexo de todas las funciones $\lambda : T \to \mathbb{R}_+$ $(:= [0, +\infty[)$ que toman valores positivos solamente en una cantidad finita de índices. El hecho de considerar un espacio tan restrictivo (cuando $T$ es infinito) para las variables duales obedece a la necesidad de garantizar la finitud de las sumas cuando las funciones de coeficientes $t \mapsto \binom{a_t}{b_t}$ son arbitrarias. Si se imponen condiciones adecuadas a dichas funciones de coeficientes, puede ampliarse el espacio de las variables duales.

Como anunciábamos hace una líneas, empezaremos reseñando algunas diferencias notables entre la PSIL y la PL ordinaria. En concreto: en PL ordinaria el conjunto factible es poliéderico, un problema es acotado si y sólo si es resoluble, lo que ocurre si y sólo si es primal-dual consistente, y no existen saltos de dualidad. En PSIL, con $T$ infinito, el conjunto factible puede ser cualquier subconjunto convexo y cerrado de $\mathbb{R}^n$, los conjuntos $\Pi_b$, $\Pi_s$ y $\Pi_c \cap \Pi_c^d$ son distintos dos a dos y pueden existir saltos de dualidad aún cuando $\pi$ y $\pi^d$ sean consistentes. Los siguientes ejemplos ilustran estos comentarios:

**Ejemplo 1.1.** $\pi \in \Pi_s \backslash \Pi_c^d$. *Consideremos el problema, en* $\mathbb{R}^2$,

$$\begin{aligned} \pi: \quad & \text{Inf } x_1 \\ & \text{s.a} \quad x_1 + t\, x_2 \geq 0, \ t \in ]0, +\infty[ \ . \end{aligned}$$

*Se tiene* $F = [0, +\infty[^2$ *y* $F^* = \{0\} \times [0, +\infty[$; *en particular* $\pi \in \Pi_s$. *Sin embargo, el cono convexo generado por* $\{a_t,\ t \in T\}$ *es* $]0, +\infty[^2 \cup \{(0,0)'\}$, *que no contiene a* $c = (1,0)'$ *(véase (3)). Por tanto* $\pi \notin \Pi_c^d$.

**Ejemplo 1.2.** $\pi \in (\Pi_b \cap \Pi_c^d) \backslash \Pi_s$ *y hay salto de dualidad. Consideremos el problema, en* $\mathbb{R}^2$,

$$\begin{aligned} \pi: \quad & \text{Inf } x_1 \\ & \text{s.a} \quad t\, x_1 + \frac{1}{t} x_2 \geq 2, \ t \in ]0, +\infty[ \ , \\ & \qquad x_1 \geq -1, \ t = 0. \end{aligned}$$

*Puede comprobarse fácilmente que*

$$F = \left\{ (x_1, x_2)' \in \mathbb{R}^2 \mid x_2 \geq x_1^{-1}, \ x_1 > 0 \right\},$$

*y por tanto* $\pi$ *es acotado* $(v = 0)$ *pero no resoluble. También se comprueba fácilmente que la única solución factible del problema dual viene dada por* $\lambda_0 = 1$, $\lambda_t = 0$ *si* $t > 0$. *Por lo tanto, el valor del problema dual es* $v^d = \lambda_0 b_0 = -1$.

Nótese que es incluso posible tener salto de dualidad cuando $\pi \in \Pi_s$. Basta añadir al Ejemplo 1.1 la última restricción del Ejemplo 1.2.

## 2. Estabilidad

En la literatura encontramos diferentes criterios de estabilidad para problemas de optimización o sus sistemas de restricciones que, en el caso concreto de la PSIL, resultan ser equivalentes a las condiciones de interioridad $\pi \in int\,(\Pi_c)$ o $\pi \in int\,(\Pi_s)$. Véanse, por ejemplo, [3], [11], [13], [18] y [21]. La topología considerada en $\Pi$ es la de la convergencia uniforme de los vectores de coeficientes, que viene dada a través de la distancia extendida $\delta : \Pi \times \Pi \longrightarrow [0, +\infty]$ definida por

$$\delta(\pi_1, \pi) := \max\left\{ \left\| c^1 - c \right\|, \sup_{t \in T} \left\| \begin{pmatrix} a_t^1 \\ b_t^1 \end{pmatrix} - \begin{pmatrix} a_t \\ b_t \end{pmatrix} \right\| \right\}, \tag{4}$$

donde las normas consideradas en $\mathbb{R}^n$ y en $\mathbb{R}^{n+1}$ son cualesquiera previamente fijadas. En este contexto se engloban, por ejemplo, aquellas situaciones en las que los datos sólo se conocen aproximadamente, o están sujetos a errores de cálculo, en cuyo caso cualquier perturbación uniformemente pequeña de *todos* los coeficientes es admisible. Es evidente que la estabilidad de un problema depende y mucho de la topología considerada. El siguiente ejemplo muestra un problema que es muy estable en nuestro contexto (4) y sin embargo resulta muy inestable cuando se considera inmerso en un contexto paramétrico, en el cual las perturbaciones no recaen directamente sobre los coeficientes, sino sobre el parámetro $\theta$ del que dependen.

**Ejemplo 2.1.** *Consideremos la familia paramétrica de problemas* $\{\pi_\theta : \theta \in \mathbb{R}\}$, *en* $\mathbb{R}^2$, *dada por*

$$\begin{aligned} \pi_\theta: \quad & \text{Inf } x_1 + x_2 \\ & \text{s.a} \quad (t+1)\, x_1 + \left(t^{-1} + 1\right) x_2 \geq \\ & \qquad -1 + \frac{\theta^2 t^2}{1 + \theta^2 t}, \ t \in ]0, +\infty[. \end{aligned}$$

Puede comprobarse fácilmente que, para todo $\theta \neq 0$, se tiene $\mathcal{F}(\pi_\theta) = [1, +\infty[ \times [0, +\infty[$, $\vartheta(\pi_\theta) =$

1 y $\mathcal{F}^*(\pi_\theta) = \{(1,0)'\}$; mientras que $\mathcal{F}(\pi_0) = [0,+\infty[\times[0,+\infty[$, $\vartheta(\pi_0) = 0$ y $\mathcal{F}^*(\pi_0) = \{(0,0)'\}$. De hecho, respecto del conjunto factible, las inclusiones '$\supset$' son inmediatas y las inclusiones inversas se siguen del Lema de Farkas (generalizado): divídase cada inecuación por el correspondiente $t$ y hágase $t \to +\infty$; del mismo modo, multiplíquense las desigualdades por $t$ y tómese $t \to 0$. Así pues, una perturbación arbitrariamente pequeña de $\theta = 0$ produce una reducción drástica del conjunto factible (en términos formales, la multifunción conjunto factible no es semicontinua inferiormente en el parámetro $\theta = 0$), y un alejamiento drástico tanto del valor óptimo como del conjunto óptimo (en ambos casos fallan la semicontinuidad inferior y la superior). Sin embargo, el problema $\pi_0$, visto ahora en nuestro contexto (4), goza localmente de gran estabilidad, dado que $\mathcal{F}(\pi) = \mathcal{F}(\pi_0)$ y $\mathcal{F}^*(\pi) = \mathcal{F}^*(\pi_0)$ para todo $\pi \in \Pi$ tal que $\delta(\pi,\pi_0) < 1$ (con respecto de la norma del supremo en $\mathbb{R}^n$ y en $\mathbb{R}^{n+1}$).

### 3. Cuantificación de la estabilidad: Distancia al mal planteamiento y regularidad métrica.

Dado que las condiciones de interioridad $\pi \in int(\Pi_c)$ y $\pi \in int(\Pi_s)$ constituyen en sí criterios de estabilidad, una forma de cuantificar dicha estabilidad viene dada por la determinación de $\delta(\pi, bd(\Pi_c))$ y $\delta(\pi, bd(\Pi_s))$, donde $bd$ significa frontera. Convendremos que un problema está *mal planteado* (*ill-posed*) respecto de una propiedad (p.e., consistencia primal y/o dual, carácter acotado o resolubilidad) cuando perturbaciones arbitrariamente pequeñas del problema puedan originar tanto problemas verificando dicha propiedad, como otros no haciéndolo; esto es, un problema se dice mal planteado respecto de una propiedad cuando pertenece a la frontera del conjunto de los problemas que gozan de la propiedad en cuestión. En este sentido las expresiones $\delta(\pi, bd(\Pi_c))$, $\delta(\pi, bd(\Pi_b))$, $\delta(\pi, bd(\Pi_s))$, $\delta\left(\pi, bd\left(\Pi_c^d\right)\right)$ y $\delta\left(\pi, bd\left(\Pi_c \cap \Pi_c^d\right)\right)$ representan diferentes nociones de *distancia al mal planteamiento* en PSIL. Dicha distancia es el objeto de estudio de [20], y allí se prueba que $bd(\Pi_b) = bd(\Pi_s) = bd\left(\Pi_c \cap \Pi_c^d\right)$ (véanse los comentarios previos al Ejemplo 1.1). Por tanto, las cinco nociones anteriores de distancia al mal planteamiento se reducen a tres.

El concepto de distancia al mal planteamiento fue introducido por Renegar en [16] en relación con la propiedad de consistencia (tanto primal como dual) en el contexto de la programación lineal cónica (la relación de orden viene definida por un cono) en espacios de Banach, y resulta ser un ingrediente básico en la generalización del número de condición a este contexto, siendo a su vez dicho número de condición un ingrediente importante en el estudio de la complejidad de ciertos algoritmos (véanse, p.e., [10] y [17]).

En nuestro contexto de la PSIL (4), los distintos tipos de distancia al mal planteamiento referidos anteriormente han sido determinados (o en algunos casos acotados inferior y superiormente) en [4] y [5] en términos de los coeficientes del problema (1). Destacamos el hecho de que, en las fórmulas que se indican a continuación, el problema de hallar la correspondiente distancia al mal planteamiento (en el espacio $\Pi$, infinito-dimensional si $T$ es infinito) se traduce en el problema de hallar la distancia, en $\mathbb{R}^n$ y/o $\mathbb{R}^{n+1}$, del origen a la frontera de un conjunto convexo determinado por los coeficientes de $\pi$. En el siguiente teorema nos referimos al espacio $\Pi_\infty := \{\pi \in \Pi : \delta(\pi, bd(\Pi_c)) = +\infty\}$, formado exclusivamente por problemas inconsistentes y caracterizado en [4]. Por $conv(X)$ denotamos a la envoltura convexa del conjunto $X$, y $cl$ significa clausura.

**Teorema 3.1.** *(i)* [4, Teoremas 5 y 6] *Si* $\pi \in \Pi\backslash\Pi_\infty$, *entonces*

$$\delta(\pi, bd(\Pi_c)) = d(0_{n+1}, bd(H)),$$

*donde*

$$H := conv\left(\left\{\begin{pmatrix} a_t \\ b_t \end{pmatrix},\ t \in T\right\}\right) + \mathbb{R}_+\begin{pmatrix} 0_n \\ -1 \end{pmatrix}.$$

*(ii)* [5, Teorema 6] *Para todo* $\pi \in \Pi$, *se tiene*

$$\delta\left(\pi, bd\left(\Pi_c^d\right)\right) = d(0_n, bd(Z)),$$

*donde*

$$Z := conv(\{a_t,\ t \in T;\ -c\}).$$

*(iii)* [5, Teorema 2] *Si* $\pi \in cl(\Pi_s)$, *entonces*

$$\begin{aligned} & \delta(\pi, bd(\Pi_s)) \\ = \quad & \min\{d(0_{n+1}, bd(H)), d(0_n, bd(Z))\}. \end{aligned}$$

Comentamos asimismo que las expresiones anteriores permiten, dado un problema $\pi \in int(\Pi_s)$, y dado $0 < \varepsilon < \delta(\pi, bd(\Pi_s))$, hallar explícitamente una constante de Lipschitz para la función valor óptimo en la bola cerrada de centro $\pi$ y radio $\varepsilon$; esto es, un $L > 0$ tal que

$$|v_1 - v_2| \leq L\delta(\pi_1, \pi_2)$$

para cualesquiera $\pi_1$ y $\pi_2$ en dicha bola. Véase [6].

Otro paso hacia la cuantificación de la estabilidad lo constituye el estudio de diferentes propiedades de tipo Lipschitz para las multifunciones conjunto factible y conjunto óptimo. Informalmente hablando, se trata de medir la tasa de reducción o expansión de dichos conjuntos respecto de la distancia entre los problemas considerados, todo ello en torno a un problema dado.

Una de las propiedades más explotadas, debido a la generalidad y a la vez simplicidad del concepto, es la de *pseudo-Lipschitz* o *propiedad de Aubin*. Vistas desde el enfoque de la multifunción inversa, las propiedades de tipo Lipschitz se traducen en propiedades de regularidad (véase, por ejemplo, [14]), y en concreto la propiedad de Aubin se traduce en la *regularidad métrica*. Para concretar, consideremos una ecuación generalizada del tipo

$$\mathcal{G}(x) \ni y,$$

donde $x$ es la variable, $y$ es el parámetro y $\mathcal{G} : X \rightrightarrows Y$ es una multifunción entre los espacios métricos $X$ e $Y$ (denotando por $d$ ambas distancias asociadas). Esta ecuación generalizada puede ser, por ejemplo, un sistema de ecuaciones del tipo $\{a_t'x \geq b_t,\ t \in T\}$ si

$$\mathcal{G}(x) = (a_t'x)_{t\in T} - \mathbb{R}_+^T \subset \mathbb{R}^T. \tag{5}$$

En este caso el parámetro sería únicamente la función $b \equiv (b_t)_{t\in T}$ del miembro derecho de las restricciones y $\mathcal{G}^{-1}(b)$ es el conjunto factible asociado, suponiendo $(a_t)_{t\in T}$ prefijado.

Decimos que una multifunción genérica $\mathcal{G}$ es métricamente regular en $\bar{x}$ para $\bar{y} \in \mathcal{G}(\bar{x})$ si existen sendos entornos $U$ de $\bar{x}$ y $V$ de $\bar{y}$ y un $k > 0$ tales que

$$d\left(x, \mathcal{G}^{-1}(y)\right) \leq kd(y, \mathcal{G}(x)) \text{ si } (x, y) \in U \times V, \tag{6}$$

donde $\mathcal{G}^{-1}(y) = \{x \in X \mid y \in \mathcal{G}(x)\}$.

Para ilustrar cómo funciona este concepto imaginemos que $\bar{x}$ es una solución de la ecuación $\mathcal{G}(x) \ni \bar{y}$, que $x_p \in U$ e $y_p \in V$ son aproximaciones de $\bar{x}$ e $\bar{y}$ respectivamente, y que $\widetilde{k} > k$. Entonces la ecuación generalizada $\mathcal{G}(x) \ni y_p$ tiene una solución que dista de $x_p$ no más de $\widetilde{k}$ veces $d(y_p, \mathcal{G}(x_p))$, que en el caso de (5) para $y_p = (b_{p,t})_{t\in T}$, es el residuo

$$\sup_{t\in T}(b_{p,t} - a_t'x_p)_+,$$

en general mucho más fácil de calcular o estimar que $d\left(x_p, \mathcal{G}^{-1}(b_p)\right)$. Así, si conocemos la rapidez de convergencia de los residuos a cero, podremos estimar la rapidez de convergencia de las soluciones aproximadas a una solución exacta.

El ínfimo de los $k > 0$ para los que (6) se verifica (para algunos entornos $U$ y $V$) se denomina *módulo de regularidad métrica*, y es una medida cuantitativa de la estabilidad de $\mathcal{G}$ alrededor de $(\bar{x}, \bar{y})$ que resulta de gran importancia en trabajos tanto teóricos como aplicados. Otra medida importante es el *radio de regularidad métrica*, que informalmente podemos interpretar como el tamaño de la mínima perturbación que ha de realizarse sobre la multifunción para que pierda la propiedad de regularidad métrica en el punto correspondiente (véanse [9] y [19]). Ya en nuestro contexto de la PSIL, la regularidad métrica de sistemas de restricciones (ecuaciones e inecuaciones) se ha analizado en [1], donde se proporcionan fórmulas para el módulo y el radio, mientras que el estudio de la regularidad métrica asociada a soluciones óptimas de problemas semi-infinitos convexos es el objetivo de [2].

No quisiéramos terminar sin hacer una sincera invitación a los lectores interesados en este tema apasionante a contactar con nosotros para ampliar información (suya o nuestra) o cotejar puntos de vista. A día de hoy nos hallamos inmersos en la determinación del módulo de regularidad métrica relativo al conjunto óptimo en PSIL (véase [7]), donde se plantean desafiantes problemas abiertos, que están por resolver incluso en el contexto de la programación lineal ordinaria.

## Referencias

[1] Cánovas M.J., Dontchev A.L., López M.A., and Parra J. (2005). Metric regularity of semi-infinite constraint systems. *Math. Program.*, **104B**, 329-346.

[2] Cánovas M.J., Klatte D., López M.A., and Parra J. (2006). Metric regularity in convex semi-infinite optimization under canonical perturbations. Preprint.

[3] Cánovas M.J., López M.A., Parra J., and Todorov M.I. (1999). Stability and well-posedness in linear semi-infinite programming. *SIAM J. Optim.*, **10**, 82-98.

[4] Cánovas M.J., López M.A., Parra J., and Toledo F.J. (2005). Distance to ill-posedness and the consistency value of linear semi-infinite inequality systems. *Math. Program.*, **103A**, 95-126.

[5] Cánovas M.J., López M.A., Parra J., and Toledo F.J. (2006). Distance to solvability/unsolvability in linear optimization. *SIAM J. Optim.* **16**, 629-649.

[6] Cánovas M.J., López M.A., Parra J., and Toledo F.J. (2006). Lipschitz continuity of the optimal value via bounds on the optimal set in linear semi-infinite optimization. *Math. Oper. Res*, to appear.

[7] Cánovas M.J., Gómez-Senent F.J., and Parra J. (2006). On the modulus of metric regularity of the optimal set in linear semi-infinite optimization. Preprint.

[8] Dall'Aglio M. (2001), On some applications of LSIP to Probability and Statistics, in M.A Goberna, M.A. López (Eds.), *Semi-Infinite Programming. Recent advances, Nonconvez Optimization and Its Applications 57*, pp. 237-254, Kluwer Academic Publ., Dordrecht (NL).

[9] Dontchev A.L., Lewis A.S., and Rockafellar R.T. (2003). The radius of metric regularity. *Trans. Amer. Math. Society*, **355**, 493-517.

[10] Freund R.M. and Vera J.R. (1999). Some characterizations and properties of the "distance to ill-posedness" and the condition measure of a conic linear system. *Math. Program.*, **86**, 225-260.

[11] Goberna M.A., López M.A., and Todorov M.I. (1996). Stability theory for linear inequality systems. *SIAM J. Matrix Anal. Appl.,* **17**, 730-743.

[12] Goberna M.A. and López M.A. (1998). *Linear Semi-Infinite Optimization*, John Wiley & Sons, Chichester (UK).

[13] Jongen H.Th., Twilt F., and Weber G.-W. (1992). Semi-infinte optimization: structure and stability of the feasible set. *J. Optim. Theory Appl.*, **72**, 529-552.

[14] Klatte D. and Kummer B. (2002). *Nonsmooth Equations in Optimization: Regularity, Calculus, Methods and Applications*, Kluwer Academic Publ., Dordrecht (NL).

[15] López M.A. and Still G. (2005). Semi-infinite programming, *Eur. J. Oper. Res*, to appear.

[16] Renegar J. (1994). Some perturbation theory for linear programming. *Math. Program.* **65**A, 73-91.

[17] Renegar J. (1995). Linear programming, complexity theory and elementary functional analysis. *Math. Program.*, **70**, 279-351.

[18] Robinson S.M. (1975): Stability theory for systems of inequalities. Part I: linear systems. *SIAM J. Numer. Anal.*, **12**, 754-769.

[19] Rockafellar R.T. and Wets R.J.-B. (1998). *Variational Analysis*, Springer, Berlín.

[20] Toledo F.J. (2003), *Distancia al mal planteamiento en optimización lineal.* Tesis doctoral. Universidad Miguel Hernández de Elche.

[21] Tuy H. (1977). Stability property of a system of inequalities. *Math. Oper. Statist. Series Opt.* **8**, 27-39.

# 3. ARTÍCULOS DE APLICACIÓN

## ANÁLISIS DEL FUNCIONAMIENTO DE UNA OFICINA DE CORREOS

**Olga Costa Estirado**[1] **y David Conesa Guillén**[2]
[1] *Licenciada en Matemáticas y en Ciencias y Técnicas Estadísticas*
[2] *Grup d'Estadística espacial i temporal en Epidemiologia i medi ambient*
*Departament d'Estadística i I.O., Universitat de València*

### 1. Introducción

El objetivo de este trabajo es aplicar la teoría de colas a las líneas de espera originadas en la oficina principal de Correos de la ciudad de Valencia y surge como resultado de la realización de las prácticas profesionales en dicha oficina por parte de la primera autora. El trabajo se estructura en 7 Secciones, siendo esta primera de carácter introductorio. En la Sección 2, comenzamos detallando cómo funciona la oficina, para a continuación, en la Sección 3, introducir los modelos de colas que mejor aproximan dicho funcionamiento. En las Secciones 4 y 5, presentamos los datos y estimamos los parámetros que describen el comportamiento de la oficina, a partir de los cuales obtenemos, en la Sección 6, las primeras conclusiones del estudio a través de las llamadas medidas de eficiencia (tamaño de la cola, tiempo de permanencia en la cola o en la oficina y utilización de los servidores). Para finalizar, en la última Sección comentamos posibles futuras líneas de trabajo.

### 2. Funcionamiento de la oficina

De lunes a viernes, la oficina de correos permanece abierta de 08:30 a 20:30 con un horario ininterrumpido de atención al público. Cuando un cliente llega a la oficina, se encuentra en la entrada con dos ordenadores en los que debe seleccionar el servicio que desea realizar. Una vez seleccionado, el ordenador le asigna un ticket con una letra y un número: la letra se refiere al servicio solicitado y el número al orden de llegada (para ese servicio).

Los clientes son atendidos en las distintas ventanillas cuando su número de ticket aparece en la pantalla de alguna de ellas. Para establecer el orden y la forma de atención, los responsables de la oficina los clasifican según el servicio que solicitan en los cuatro tipos siguientes:

- *Enviar*: Incluye a los clientes que solicitan servicios referentes al envío de correspondencia (envío de cartas, giros y telegramas, etc.).
- *Recoger*: Son aquellos que acuden a la oficina a recoger cartas y giros postales y también los que requieren gestiones relacionadas con los apartados de correos.
- *Filatelia*: Agrupa a los clientes que desean adquirir productos de Filatelia (sellos, sobres, etc.).
- *Banco*: Son los clientes que acuden a la oficina a realizar gestiones bancarias.

En concreto, los clientes son atendidos en orden de llegada, independientemente de la letra del ticket pero sí dependiendo del tipo al que pertenecen y, una vez atendidos, abandonan la oficina.

Respecto a los mostradores, en la oficina hay 15 que atienden a los 4 tipos de clientes. Hay uno específico para los clientes tipo *Filatelia* y otro para los clientes tipo *Banco*, y el resto, 13, se emplean unos para atender al tipo *Recoger* y otros para el tipo *Enviar*. La política actual de la oficina para la asignación de estos mostradores es la que muestra el Cuadro 1.

| HORARIO | Nº DE SERVIDORES |
|---|---|
| 08:30-09:30 | 3 mostradores que atienden conjuntamente *Enviar* y *Recoger* |
| 09:30-14:30 | 2 para *Recoger* y 11 para *Enviar* |
| 14:30-18:00 | 1 para *Recoger* y 3 para *Enviar* |
| 18:00-20:30 | 1 para *Recoger* y 6 para *Enviar* |

Cuadro 1: *Asignación actual de ventanillas.*

## 3. Los modelos de colas

Una vez presentado el funcionamiento de la oficina, pasamos a describir nuestra primera aproximación al problema, que se basa en la utilización de modelos básicos de Teoría de Colas. En concreto, hemos considerado 4 colas independientes que hacen referencia a los 4 tipos de clientes.

Cada una de estas colas se puede aproximar según un modelo $M|M|k$ de acuerdo con la notación de Kendall (1951) [2], donde la primera letra hace referencia a la distribución de los tiempos entre llegadas, la segunda a la distribución de los tiempos de servicio y la $k$ al número de mostradores asignados a cada línea de espera. En particular, la $M$ indica que tanto los tiempos entre llegadas como los de atención siguen una distribución exponencial (ver Medhi (2002) [4] para una descripción más detallada de estos modelos).

Conviene destacar que, a priori, se puede intuir que las llegadas a la oficina se producirán según un proceso de Poisson de parámetro desconocido $\lambda$, tasa media de llegadas de clientes a la oficina, lo que garantizaría la exponencialidad de los tiempos entre llegadas. Lo que no está tan claro es que los tiempos de atención o de servicio sigan también una distribución exponencial con media $\frac{1}{\mu}$, ya que parece más lógico pensar que estos tiempos se mantendrán más o menos constantes a lo largo del día. La justificación del uso de los modelos $M|M|k$ la veremos en la Sección 5.

A continuación describimos el resto de elementos que determinan cada una de las colas:

Clientes: Consideramos como clientes a los individuos/usuarios que acuden a la oficina para realizar alguno de los servicios ofrecidos.

Servidores: Son los empleados de la oficina que están detrás de cada uno de los mostradores de atención al público. De acuerdo con la política de la oficina, los clientes del tipo *Filatelia* y *Banco* son atendidos por un único servidor ($k = 1$) mientras que los de *Enviar* y *Recoger* son atendidos por $k$ servidores (con $k$ variable).

Disciplina de la cola: Cada una de las colas sigue una disciplina FIFO (First In First Out), es decir, el primero que llega es el primero que es atendido.

Tras introducir los modelos de colas que aproximan el funcionamiento de la oficina, pasamos a comprobar, en las dos secciones siguientes, que realmente tanto las llegadas como los servicios son procesos de Poisson y a estimar los parámetros de dichos procesos.

## 4. Los datos.

La manera habitual de estimar los parámetros de los modelos de colas es a través de tiempos entre llegadas consecutivas y de tiempos de servicio (Bath et al. (1997) [1] presentan de manera detallada cómo realizar inferencia estadística en modelos de colas).

En nuestro caso, y debido a que la oficina está totalmente informatizada, nos es posible conocer en cualquier momento del día cual es su situación, es decir, conocer el estado de cada mostrador (cerrado, atendiendo o pausado), el número de ticket al que están atendiendo y también, entre otras cosas, el número de clientes por servicio esperando a ser atendidos. Así pues, aunque no tenemos información directa sobre los tiempos entre llegadas ni sobre los tiempos de servicio, sí que disponemos de datos sobre los tickets expedidos, en concreto, sobre su hora de expedición, su letra, su número y los tiempos de inicio y fin de atención.

Para utilizar esta información, hemos considerado que un ticket expedido equivale a una llegada a la oficina. ¿Motivo? Generalmente, en los ordenadores de la entrada no se acumulan los clientes, por lo tanto, cuando un cliente llega a la oficina saca el ticket instantáneamente y en este caso la hora de expedición del ticket coincide con la hora de llegada del cliente. De esta manera, el tiempo entre expediciones de dos tickets consecutivos puede considerarse como el tiempo entre dos llegadas consecutivas.

Respecto a los tiempos de servicio, contamos con datos de los instantes de inicio y finalización de la atención. Como hora de inicio de atención consta la hora en la que el ticket ha sido llamado al mostrador correspondiente y como hora de fin de atención la hora en la que, en ese mismo mostrador, se ha llamado al siguiente número. Por tanto, hemos calculado el tiempo de servicio (o de atención) como el tiempo que transcurre desde que un cliente es llamado al mostrador hasta que se llama al siguiente (en el mismo mostrador).

Para concluir, conviene destacar que disponemos de una base de datos con 130.000 registros correspondientes a los tickets expedidos desde el día 03-01-2005 hasta el día 03-05-2005, ambos inclusive. Como el rango de fechas es muy amplio y deseamos encontrar un modelo que represente un día estándar en la oficina de correos, hemos eliminado los registros con fechas conflictivas.

## 5. Inferencia sobre los parámetros del modelo

En apartados anteriores ya hemos avanzado que tanto los tiempos entre llegadas consecutivas como los tiempos de servicio siguen una distribución exponencial. En este apartado, presentaremos el análisis de los datos que justifica dicha afirmación.

Como las llegadas a la oficina dependen en gran medida de la hora, como primera aproximación al problema consideraremos franjas horarias donde tengamos valores constantes de los parámetros a estimar: $\lambda$, tasa media de llegadas de clientes a la oficina, y $\mu$, tasa media de servicio por mostrador, para cada uno de los modelos.

Analizaremos de forma distinta los datos de tiempos entre llegadas y los de tiempos de servicio. Los primeros nos servirán para encontrar las franjas horarias donde la afluencia de llegadas a la oficina es similar y en cada una de ellas obtendremos una estimación del parámetro $\lambda$. Con los tiempos de servicio nos limitaremos a estimar $\mu$ en las franjas obtenidas a partir del análisis anterior.

Insistimos en que este procedimiento nos da una primera aproximación al problema real. Habitualmente, el parámetro $\lambda$ no es constante sino que varía a lo largo del tiempo (procesos de Poisson no estacionarios). Considerar parámetros variables $\lambda(t)$ y $\lambda(t)$ en lugar de parámetros constantes supondría, sin duda, una mejor aproximación al funcionamiento de la oficina, pero no es el objetivo de este trabajo, aunque sí que lo consideramos como su extensión natural.

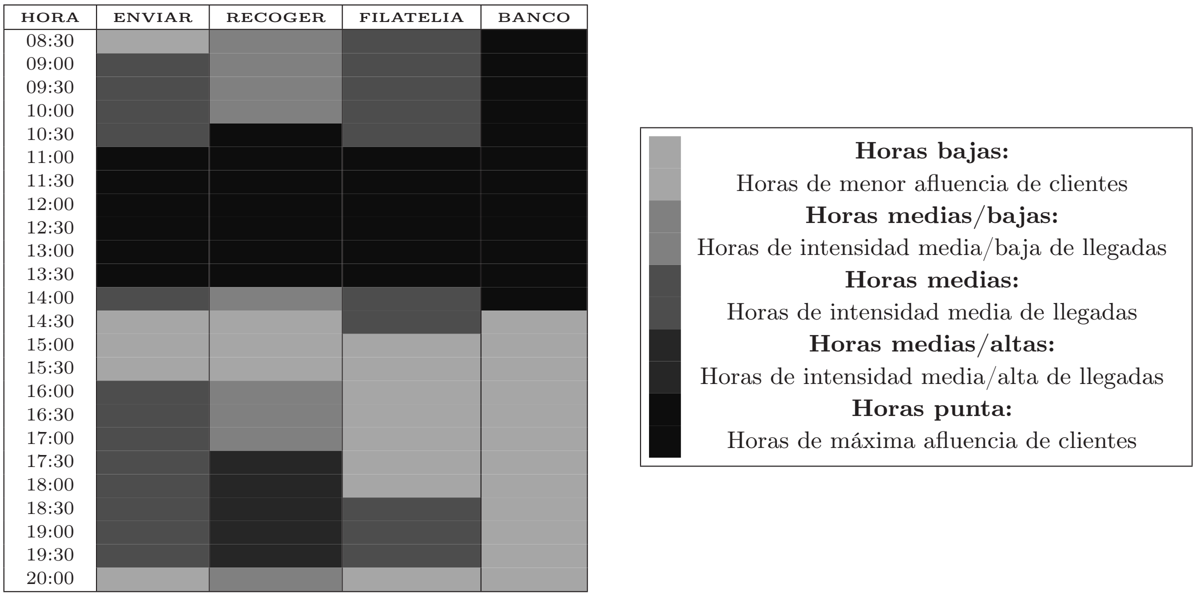
### 5.1. Tiempos entre llegadas

Tal y como hemos comentado anteriormente, hemos calculado los tiempos entre llegadas como la diferencia entre la hora de llegada de un cliente y la hora de llegada del cliente anterior.

El procedimiento que hemos seguido para detectar las franjas horarias se basa en una técnica propuesta por Law y Kelton (2000) [3]. En primer lugar hemos dividido el tiempo en intervalos horarios de 30 minutos desde las 08:30 hasta las 20:30. En cada uno de los intervalos hemos analizado, utilizando el programa SPSS (2001) [5], los datos de cada cola, separándolos además por día de la semana (L, M, X, J, V) para comprobar si influye el día en los tiempos entre llegadas.

Así, para cada uno de estos intervalos hemos representado los datos en un histograma y hemos realizado el test de bondad de ajuste de Kolgomorov y Smirnov al nivel $\alpha$ =0.05 contrastando la hipótesis nula de que los datos de cada intervalo son exponenciales. Conviene destacar que todos los tests no muestran evidencia en contra de dicha hipótesis, por lo que asumimos que los datos de cada grupo provienen de una distribución exponencial. Además observamos que no existen diferencias significativas entre los días de la semana.

Como estimador del parámetro $\lambda$ de la exponencial, utilizamos el estimador máximo verosímil, cuya expresión en este caso es la inversa de la media muestral. Agrupando los datos de los intervalos con medias muestrales similares, conseguimos reducir el número de parámetros. Confirmamos la exponencialidad de los datos aplicando de nuevo el test de bondad de ajuste en cada uno de los grupos resultantes. Siguiendo con este procedimiento, obtenemos grupos con medias muestrales bastante diferenciadas que son los que finalmente nos dan las franjas horarias que buscábamos y que hemos representado en el Cuadro 2.




Cuadro 2: *Franjas horarias para las 4 colas.*

Para no extender innecesariamente este trabajo, en la siguiente tabla mostramos únicamente las estimaciones del modelo de colas *Enviar* (u. t. = minuto).

| Parámetro | $\lambda_{H.bajas}$ | $\lambda_{H.medias}$ | $\lambda_{H.punta}$ |
|---|---|---|---|
| Estimación | 0.739 | 1.264 | 1.962 |

Podemos observar que realmente hay diferencias importantes en las llegadas diarias a la oficina, en las horas bajas hay 0.7 llegadas por minuto mientras que en las horas punta esta tasa es superior, aproximadamente 2 clientes por minuto.

### 5.2. Tiempos de servicio

Como hemos comentado anteriormente, los tiempos de servicio los analizamos en las franjas horarias obtenidas a partir del análisis de los tiempos entre llegadas. Conviene señalar que hemos eliminado los tiempos superiores a 20 minutos ya que, según los responsables de la oficina, estos tiempos se deben a un error en la toma de los datos (una posible causa podría ser que algún servidor hubiera abandonado momentáneamente un mostrador sin dejarlo pausado).

Hemos representado los datos mediante histogramas y realizado el test de bondad de ajuste de Kolgomorov y Smirnov a un nivel $\alpha$ =0.05, encontrando que no podemos rechazar la hipótesis de que los datos provengan de distribuciones exponenciales.

Al igual que antes, en la tabla siguiente presentamos sólo las estimaciones de los parámetros de la cola *Enviar* (u. t. = minuto).

| Parámetro | $\lambda_{H.bajas}$ | $\lambda_{H.medias}$ | $\lambda_{H.punta}$ |
|---|---|---|---|
| Estimación | 0.234 | 0.254 | 0.254 |

En ella observamos que la tasa de servicio es prácticamente idéntica durante toda la jornada laboral siendo sensiblemente menor durante las horas bajas (esto implica que, en estas horas, los servidores tardan más tiempo en atender a los clientes). Podríamos haber considerado únicamente una franja para los tiempos de servicio, pero hemos decidido mantener las originales por coherencia con la estructura de las llegadas.

## 6. Cálculo de las medidas de eficiencia

Tal y como hemos visto, los parámetros nos dan una visión de cómo funciona el sistema. De todas maneras, en Teoría de Colas, cuando el sistema se encuentra en equilibrio (es decir, cuando su comportamiento se hace independiente del tiempo), es mucho más habitual el utilizar las llamadas medidas de eficiencia (tales como el tiempo de espera o el número de clientes en cola) para describir de una manera mucho más intuitiva cuál es su comportamiento. Estas medidas nos pueden ser de utilidad, por ejemplo, para decidir aumentar el número de ventanillas abiertas en ciertas horas del día y a qué tipo de clientes asignárselas.

En nuestro caso, comprobamos qué colas se encuentran en esa situación verificando para ello que se cumple la condición de equilibrio, que, en este tipo de modelos, viene determinada por el hecho de que la intensidad de tráfico debe ser menor que la unidad. Si alguna de las colas analizadas no verifica dicha condición significa que, con el tiempo, puede llegar a congestionarse. En particular, para los modelos $M|M|1$ (*Filatelia* y *Banco*) y $M|M|k$ con $k > 1$ (*Enviar* y *Recoger*) esta condición viene expresada respectivamente por:

$$\rho = \frac{\lambda}{\mu} < 1 \qquad \text{y} \qquad \rho = \frac{\lambda}{k\mu} < 1$$

donde $\rho$ es la intensidad de tráfico (tasa media de llegadas/tasa media de servicio por servidor ocupado).

Tal y como podemos observar en el Cuadro 3, es suficiente un servidor para la cola *Filatelia* y otro para la cola *Banco* para atender a los clientes de estos servicios sin que estas colas lleguen a congestionarse de manera excesiva. En otras palabras, se verifica la condición de equilibrio en todas las situaciones.

| | FILATELIA | | | BANCO | |
|---|---|---|---|---|---|
| | H. bajas | H. medias | H. punta | Mañana | Tarde |
| Tasa de llegadas ($\hat{\lambda}$) | 0.057 | 0.115 | 0.229 | 0.028 | 0.015 |
| Tasa de servicio ($\hat{\mu}$) | 0.379 | 0.290 | 0.290 | 0.359 | 0.399 |
| Intensidad de tráfico ($\hat{\rho}$) | 0.151 | 0.396 | 0.792 | 0.078 | 0.037 |

Cuadro 3: *Valores de $\hat{\lambda}$, $\hat{\mu}$ y $\hat{\rho}$ para los modelos Filatelia y Banco (u. t. = minuto).*

En el Cuadro 4 presentamos algunas medidas de eficiencia para estos dos modelos. Los resultados obtenidos confirman que efectivamente una ventanilla para cada cola es suficiente para garantizar un buen servicio. Además, observamos que el mostrador de la cola *Banco* está la mayor parte del tiempo desocupado, mientras que el de *Filatelia* funciona bastante bien durante todo el día excepto en las horas punta donde el tiempo esperado en la cola es aproximadamente de 13 minutos. Los responsables de la oficina podrían plantearse en este caso destinar un mostrador más para esta cola si consideraran que este tiempo de espera es excesivo.

| | FILATELIA | | | BANCO | |
|---|---|---|---|---|---|
| | H. bajas | H. medias | H. punta | Mañana | Tarde |
| Utilización | 15.09 % | 39.56 % | 79.18 % | 7.76 % | 3.70 % |
| Probabilidad de que el sistema esté vacío | 0.849 | 0.604 | 0.208 | 0.922 | 0.963 |
| Longitud esperada de la cola | 0.027 | 0.259 | 3.011 | 0.007 | 0.001 |
| Número esperado en el sistema | 0.178 | 0.655 | 3.802 | 0.084 | 0.038 |
| Tiempo esperado en la cola | 0.469 | 2.258 | 13.118 | 0.234 | 0.096 |
| Tiempo total esperado en el sistema | 3.108 | 5.708 | 16.568 | 3.022 | 2.600 |
| Probabilidad de que un cliente espere | 0.151 | 0.396 | 0.792 | 0.078 | 0.037 |

Cuadro 4: *Medidas de eficiencia para las colas Filatelia y Banco.*

Para finalizar, analizamos la condición de equilibrio en los modelos de las colas *Enviar* y *Recoger*. Por cuestiones de espacio, no vamos a mostrar todas las comprobaciones, pero son varios los casos en los que no se cumple esta condición. Como en estos casos el sistema no alcanza el equilibrio, se hace necesaria la utilización de herramientas de simulación para describir las medidas de eficiencia del sistema. La simulación de la oficina es una futura línea de trabajo y no se encuentra recogida en nuestra primera aproximación al funcionamiento de la oficina.

## 7. Posibles extensiones

A partir de ahora, nuestro trabajo se centra en la validación de la modelización planteada, para lo cual necesitaremos herramientas de simulación. De esta manera comprobaremos si los modelos que hemos elegido se ajustan al funcionamiento actual de la oficina, a la vez que nos ayudarán a predecir su comportamiento.

Otra de las futuras líneas de trabajo que nos planteamos es la aproximación del funcionamiento de la oficina mediante modelos $M(t)|M(t)|k$, que constituyen la extensión natural de los modelos considerados en este trabajo.

## Agradecimientos

Los autores agradecen a Manuel Molina Fernández la cuidadosa lectura de este trabajo y los comentarios realizados que han llevado a esta versión final. Los autores también agradecen a la oficina principal de Correos de la ciudad de Valencia la colaboración prestada en la realización de este trabajo. El segundo autor agradece la financiación del proyecto de investigación MTM 2004-03290 cofinanciado por el Ministerio de Educación y Ciencia y fondos FEDER.

## Referencias

[1] Bhat U. N., Miller G. K. and Rao S. S. (1997). Statistical Analysis of Queueing Systems. *Frontiers in Queueing*, editado por J. H. Dshalalow.

[2] Kendall D. G. (1951). Some problems in the theory of queues. *DJ. Roy. Statist. Soc. Ser. B*, **13**, 151-185.

[3] Law M. and Kelton W. D. (2000). *Simulation modeling and analysis*, Mc Graw Hill, 3ª edición.

[4] Medhi J. (2002). *Stochastic Models in Queueing Theory*, Academic Press, 2ª edición.

[5] SPSS for Windows, Rel. 11.0.1 (2001). Chicago: SPSS Inc.

# 4. ESTADÍSTICA OFICIAL

## ENCUESTA SOBRE LAS PERSONAS SIN HOGAR (EPSH-2005)

**Pedro Ruíz Salvador**
Instituto Nacional de Estadística

**Resumen**

Actualmente no se dispone de estadísticas oficiales sobre las personas sin hogar ya que los principales estudios han sido realizados desde el ámbito no oficial. Sin embargo, en los últimos años, la preocupación sobre la exclusión social desde los gobiernos y la propia sociedad civil, tanto a nivel nacional como a nivel europeo, ha ocasionado la activación de programas específicos que requieren la elaboración de la información estadística necesaria para la toma de decisiones. En este contexto, el Instituto Nacional de Estadística (INE) ha realizado una encuesta dirigida a las personas sin hogar durante el mes de febrero de 2005 cuya metodología y resultados son la base de este artículo.

### 1. Antecedentes

En los últimos años, tanto desde instancias gubernamentales como desde la propia sociedad civil ha aumentado el grado de sensibilidad y preocupación por la cohesión de nuestras sociedades, especialmente por aquellos procesos de exclusión social que generan una segregación creciente en las mismas.

Así, el propio Tratado de Amsterdam estableció como uno de los seis objetivos de la Política Social Europea el de combatir la exclusión social (art.136). El Programa de Acción para la Inclusión Social de la Unión Europea (2001-2005) y el homónimo español que es parte integrante, alientan a la elaboración de la información estadística necesaria para poderlo llevar a cabo. Ello exige la elaboración de indicadores y estadísticas sobre la exclusión social y, en particular, sobre las personas sin hogar como manifestación extrema de la misma.

En este contexto, Eurostat constituyó en octubre de 2001 un Grupo de Trabajo con el fin de establecer un marco metodológico para la investigación sobre las personas sin hogar en el ámbito de la UE. De acuerdo con las reflexiones realizadas, los objetivos ideales a alcanzar serían:

a) La medición periódica de la carencia de alojamiento.

b) El estudio de las características de las personas carentes de alojamiento mediante encuestas directas.

Con esta finalidad, se considera conveniente llevar a cabo diferentes tipos de investigaciones, entre las que podemos señalar:

- Operaciones para mejorar el conocimiento de los centros de alojamiento.
- Encuestar mediante entrevista directa a los usuarios de los centros de alojamiento.

### 2. Objetivos

En el invierno del 2004 el INE llevó a cabo la Encuesta a las Personas sin Hogar (Centros) que estaba dirigida a los centros que prestan servicio a las personas sin hogar y que permitió conocer las características de los mismos (prestaciones ofrecidas, población atendida, orientación, fuentes de financiación, recursos humanos y financieros, período de actividad habitual, capacidad y ocupación) además de proporcionar una estimación del número medio de usuarios de la red de centros en 2002 y a 5 de noviembre de 2003.

Con el fin de seguir profundizando en el conocimiento de las personas sin hogar, considerando las reflexiones realizadas en el seno de la Unión Europea, y que en nuestro sistema estadístico no existe información al respecto, el INE ha realizado una encuesta para estudiar las características de los usuarios de los centros que prestan servicios de alojamiento y/o de restauración denominada *Encuesta a las Personas sin Hogar – Personas (EPSH-2005).*

La encuesta tiene por finalidad conocer el perfil sociodemográfico, las condiciones de vida y las dificultades de acceso al alojamiento de las personas del colectivo a nivel nacional y, si fuera posible, también por comunidades autónomas para las variables más relevantes.

## 3. Características generales objeto de estudio

Las características objeto de estudio son las siguientes:

- *Características sociodemográficas* (sexo, edad, lugar de nacimiento, tiempo de residencia en España, empadronamiento y nacionalidad).
- *Frecuentación de servicios* (tipo de alojamiento usado, frecuencia de uso y hábitos en el uso del mismo).
- *Condiciones de vida* ( comportamiento respecto al alojamiento de las distintas personas: uso de teléfono, internet, ...).
- *Alojamiento: Antecedentes y búsqueda* (causas por las que se está sin alojamiento y desde cuando, disponibilidad de las personas para cambiar su situación de alojamiento y búsqueda de alojamiento).
- *Actividad, empleo y paro* (relación con la actividad tanto desde una perspectiva retrospectiva como en la actualidad).
- *Situación económica* (diferentes fuentes de ingresos, cuantía, principales gastos y endeudamiento).
- *Formación* (nivel de estudios terminados, edad de abandono de los mismos, dificultades de lectoescritura y cálculo).
- *Salud* (valoración subjetiva del estado de salud, sueño, relación de la persona con el sistema de salud, aproximación objetiva al estado de salud, adicción a medicamentos, alcohol y drogas).
- *Familia* (vínculos y antecedentes).
- *Utilización de los servicios sociales* (utilización de los centros de día, relaciones con trabajadores sociales,valoración de los servicios sociales, percepción de la renta mínima de inserción).
- *Relación con la justicia* (haber sido víctima de un delito o agresión, haber sido denunciado o detenido, asistencia jurídica recibida, y en su caso, si fueron condenados).

## 4. Ámbito de la encuesta

El ámbito abarcado por la encuesta se desglosa en los tres apartados siguientes:

### 4.1. Ámbito poblacional

La encuesta va dirigida a las personas sin hogar, de 18 y más años, que viven en municipios de más de 20.000 habitantes. Aceptamos como definición de *persona sin hogar* la establecida como base de trabajo provisionalmente en el Grupo de trabajo de Eurostat, con ligeras modificaciones para su adaptación a la realidad española.

En general podemos decir que se incluye las personas que duermen temporalmente en la calle, en edificios no habitables, en alojamientos de emergencia o colectivos proporcionados por la ayuda pública o privada, en pensiones o casas de huéspedes o en casas ocupadas. Y se excluyen aquellas cuyos alojamientos son los hospitales o similares, las prisiones, las residencias de estudiantes, los orfelinatos, los cuarteles, los barcos amarrados, las casas rodantes u hoteles.

Sin embargo, aunque *la población teórica* es la indicada (a falta de un acuerdo sobre la definición de persona sin hogar en el contexto de la Unión Europea), la población que efectivamente va a ser investigada es aquella que acude a los centros que ofrecen servicios de alojamiento y/o de restauración.

### 4.2. Ámbito territorial

La encuesta se realiza en los municipios mayores de 20.000 habitantes en toda España (excepto en el País Vasco en donde se extiende a todos los municipios al realizarse la investigación mediante convenio con el Instituto Vasco de Estadística).

### 4.3. Ámbito temporal

A semejanza de otras experiencias internacionales recientes, se ha optado por extender las entrevistas a lo largo de un mes y, teniendo en cuenta la naturaleza de nuestro estudio, se ha considerado que dicho período debe ser en invierno, habiendo establecido que fuera en el mes de febrero de 2005.

## 5. Diseño muestral

Como marco de unidades primarias se ha empleado el directorio de centros, que se deriva de la Encuesta sobre Personas sin Hogar(Centros) realizada previamente por el INE.

Se ha utilizado un muestreo en dos etapas, estratificado en las unidades de primera etapa.

Las unidades de primera etapa la constituyen los centros ubicados en municipios mayores de 20.000 h. que prestan sus servicios a la población objeto de estudio. Los centros se estratifican de acuerdo a las variables más significativas recogidas en el marco como son: tipo de servicios que prestan y número de servicios.

Las unidades de segunda etapa la constituyen los servicios prestados por los centros de atención, tanto de alojamiento como de restauración, que se ofrecen a este tipo de población. A través de la demanda de estos servicios se va a estimar el número total de personas usuarias de los centros como aproximación al concepto de persona sin hogar.

De acuerdo con la experiencia internacional y teniendo en cuenta la elevada falta de respuesta que se podía presentar, así como la imposibilidad de establecer un criterio uniforme para la selección de las personas en los centros que reciben los servicios prestados, se optó por establecer el tamaño de ésta en 4.000 servicios.

Para determinar el tamaño muestral de unidades de primera etapa (centros) se han tenido en cuenta los siguientes aspectos:

- El número de entrevistas diario se fijó en tres, teniendo en cuenta que el tiempo medio de cumplimentación del cuestionario es de 30 minutos.
- De acuerdo con la estructura de los centros (hay pocos centros grandes y muchos muy pequeños), se visitarán 1, 2, 3, 4 y 5 días, distribuyendo las visitas en una o varias semanas a lo largo de cuatro semanasdel mes de febrero.

La distribución de la muestra de servicios entre estratos ha sido proporcional al tamaño de cada uno de ellos según el número de servicios y, dentro de los estratos, se ha utilizado el mismo criterio para distribuir la muestra entre los distintos centros.

La selección de las unidades de primera etapa (centros) se ha realizado con probabilidad proporcional al número de servicios que prestan y para ello, previamente en cada estrato, los centros se han ordenado según la provincia a la que pertenecen excluyendo aquellos centros que prestaron menos de cuatro servicios.

Dentro de cada centro seleccionado, la muestra de servicios se ha seleccionado con igual probabilidad. De acuerdo a las distintas situaciones, y teniendo en cuenta que las personas son las unidades de investigación seleccionadas a través de los servicios que demandan, la forma de actuar en la selección de la muestra de servicios ha sido la siguiente:

- En aquellos centros que dispensan el servicio de alojamiento y en el supuesto de que se pudiera disponer de una lista de las personas que son acogidas en una noche, el entrevistador selecciona la muestra de servicios a partir de la misma mediante un muestreo sistemático.

  En el caso en el que el centro, por cualquier circunstancia, no facilite o no disponga de tal lista, se elegirán los individuos según su orden de llegada o de salida, según el momento de la entrevista. Si la llegada o el abandono del centro de los allí acogidos se produce de forma simultánea por varios de ellos impidiendo poner en práctica los mecanismos anteriores, el entrevistador seleccionará de la forma que considere aleatoria los individuos a entrevistar, por ejemplo seleccionando a uno en cada grupo de los que entran o salen simultáneamente hasta completar el tamaño muestral.

- En los centros que se ofrecen servicio de restauración la elección de los entrevistados se hizo por orden de llegada al mismo para recibir dicha prestación.

  Si la afluencia masiva de personas al centro impide llevar a cabo tal mecanismo de selección, el entrevistador recurrirá también a elegir de la forma más aleatoria posible a los entrevistados. Por ejemplo, seleccionando durante la comida y procurando que esperen al final para realizar la entrevista o haciendo la selección según llegada en el momento en que el entrevistador esté disponible.

### 6. Recogida y tratamiento de la información

Considerando el ámbito temporal y el tamaño de la muestra, se ha estimado más eficiente la división del territorio en 18 zonas de recogida, tomando en consideración para la agrupación de provincias la carga de trabajo de las mismas y su proximidad. De esta forma los trabajos de campo (recogida de la información, depuración de los datos e inspección y control del trabajo) se centralizarón en la delegación cabecera de zona.

En primer lugar, se remitió una carta de presentación de la encuesta a los centros solicitando su colaboración. A continuación, el Inspector de la encuesta se puso en contacto, a través de entrevista personal, con cada uno de los centros seleccionados para informar del objetivo de la encuesta y conocer de antemano si disponían de alguna lista de las personas que asisten a diario al centro, para poder hacer una selección previa de los informantes. También aprovechó para concertar una cita para el entrevistador y un lugar (despacho o similar) donde se pudiera llevar a cabo la entrevista con las menores molestias posibles para el centro y sus usuarios.

Los entrevistadores, en su primera visita al centro, fueron acompañados por el Inspector de la encuesta o por el Inspector de entrevistadores, para proceder a su presentación. Asi también el Inspector se aseguró de la disponibilidad de un lugar apropiado para celebrar las entrevistas, de la selección aleatoria de los entrevistados y de que durante la realización de las entrevistas no estuviera presente el asistente social o cualquier otra persona perteneciente al centro que pudiera influir en las respuestas.

De acuerdo con las conclusiones de la experiencia piloto realizada previamente, el personal que participó en la encuesta, tanto en calidad de entrevistadores como de inspectores de entrevistadores, fueron trabajadores sociales con experiencia en el trato con personas en riesgo de exclusión social y, en la medida de lo posible, con conocimientos de idiomas (francés, inglés, árabe, ruso).

La grabación de datos de los cuestionarios recogidos se ha realizado mediante técnicas de escaneado y reconocimiento óptico de caracteres. El fichero de grabación generado ha sido objeto de un proceso de depuración con el fin de detectar las posibles inconsistencias en la información primaria para dar lugar al fichero base de explotación que ha sido tratado mediante la aplicación estadística SPSS para el análisis de datos y la elaboración de las tablas previstas en el plan de explotación de la encuesta.

### 7. Algunos resultados

La población sin hogar atendida en los centros se estima en 21.900 personas en todo el territorio y está compuesta por el 82,7 % de varones frente al 17,3 % de mujeres, lo que muestra una notable masculinización del fenómeno, en contraste con la distribución por sexo de la población en general.

El colectivo en cuestión es relativamente joven ya que el 72,7 % tiene menos de 45 años, siendo la edad media de las personas sin hogar de 37,9 años. El 79,1 % de esta población está empadronada en algún municipio.

Atendiendo a la nacionalidad, el 51,8 % son españoles y el 48,2 % extranjeros. Entre los extranjeros, el grupo mayoritario es el de los africanos, seguido por los europeos y los americanos.

De acuerdo con las categorías consideradas para estudiar la situación familiar, sólo el 17,4 % de las personas mantiene una unión estable (casado o pareja de hecho) y, en cuanto a la descendencia, el 46 % de la población sin hogar tiene hijos aunque solamente una décima parte vive con ellos.

El 70,2 % de las personas sin hogar pernocta todas las noches en el mismo lugar, lo que podría indicar la existencia de referencias en el alojamiento. El 45,6 % de la población sin hogar se aloja al margen de la red asistencial existente (en pisos ocupados, espacios públicos o alojamientos de fortuna). Según el tiempo transcurrido sin alojamiento, el 37,5 % declara llevar más de tres años sin alojamiento propio.

El ingreso medio de las personas sin hogar se estima en 302 euros al mes. En cuanto a la principal fuente de ingresos, el 19,9 % de las personas sin hogar vive de su salario, el 7,4 % de la venta de objetos y la prestación de servicios, el 14,2 % vive del dinero de gente de la calle y otro 16,4 % de la familia y los amigos. El 17,5 % vive de prestaciones públicas y el resto no tiene ingresos o no sabe/no contesta.

En lo relativo a los estudios terminados, el 64,8 % de la población ha alcanzado un nivel de educación secundaria, el 15,3 % estudios primarios o inferiores y el 13,2 % estudios superiores (universitarios o no), mientras que el 6,7 % de las personas se declara sin estudios.

En lo que concierne a la apreciación subjetiva de su estado de salud, el 15,6 % declara tener mala o muy mala salud, mientras que el 52,7 % dice tener buena o muy buena salud. El 30 % de la población sin hogar es abstemia y nunca ha consumido drogas.

El 60,6 % del colectivo opina que los servicios sociales le ayudaron poco o nada en su situación.

Una información más detallada es proporcionada en la página web del INE http://www.ine.es/inebase/.

# 5. ESTUDIOS MONOGRÁFICOS Y OPINIONES SOBRE LA PROFESIÓN

## EN MEMORIA DE LÁZARO CÁNOVAS MARTÍNEZ

**Alfredo Marín Pérez**
Departamento de Estadística e Investigación Operativa
Universidad de Murcia

Lázaro Cánovas Martínez, profesor titular y director del Departamento de Estadística e Investigación Operativa de la Universidad de Murcia, falleció repentinamente en la madrugada del día 15 de mayo. Tenía 38 años, y deja esposa y dos hijas de 11 y 8 años. A nadie que lo conociera bien le pudo extrañar que en ese momento estuviera sentado frente a su ordenador, trabajando en una aplicación informática para las Pruebas de Acceso a la Universidad, de cuyo tribunal iba a ser secretario. A Lázaro le apasionaba la programación, y necesitaba poco estímulo externo para sentarse a desarrollar una idea para el departamento, para su investigación, para las pruebas de acceso o para quien se lo pidiera. En estos días que han transcurrido desde su muerte hemos añorado al compañero, al amigo, al director, y nos hemos sentido huérfanos cuando hemos tenido dificultades con un programa, con una máquina, con linux, cuando hemos necesitado esa ayuda que él nunca negaba a nadie.

Lázaro estudió Matemáticas en Murcia (1990) y se incorporó rápidamente al Departamento. Defendió su tesis doctoral “El problema del k-centro en $\mathbb{R}^n$ con normas $l_{pb}$ estrictas” en 1995. En su relativamente breve carrera investigadora trabajó en este tema y en Programación Entera, métodos para la triangulación de poliedros, sistemas de votación y otros. Ha publicado unos 20 artículos, presentado unas 50 ponencias en congresos y codirigido dos tesis doctorales. Sé que si le hubiéramos pedido que nos relatara sus mejores momentos profesionales habría citado la muy reciente defensa de la tesis doctoral de Sergio García, su colaboración -truncada en los albores- con Martine Labbé, el inolvidable año en que nos unimos a un fantástico grupo de amigos e investigadores en el proyecto Saderyl, el momento de la creación de nuestro propio grupo de investigación, al que nos gustaba llamar Z3 en la intimidad y fuera de ella, los días vividos en el EURO Summer Institute de 1995, la publicación en Mathematical Programming de aquel artículo del que tan orgulloso se sentía, o incluso reviviría los momentos de nervios sufridos en el curso de nuestra gran aventura práctica, cuando las fuerzas vivas de Caravaca introducían las claves para el sistema de votación que habíamos desarrollado con otros dos formidables amigos y que servía para decidir el concurso de los Caballos del Vino.

Como profesor, Lázaro era muy querido por los alumnos, fiel cumplidor de sus obligaciones y defensor a ultranza de las prácticas. Tampoco hizo ascos a la gestión. Además de la dirección del departamento y de la colaboración en las pruebas de acceso, era actualmente miembro del Claustro y fue en su momento miembro del Consejo Académico de Investigación Operativa de la SEIO. En todos los foros era enemigo de enfrentamientos personales, y los pocos, poquísimos, que alguna vez pudieron enemistarse con él, no consta que oyeran de su boca una mala palabra. Le recordamos todos sonriente, afable. En las relaciones sociales, Lázaro se sentía como pez en el agua. Gustaba de divulgar sus últimos conocimientos en materia de dispositivos informáticos, de vino y exquisiteces gastronómicas. Si eras su invitado en Murcia no dudaba en llevarte al lugar óptimo en el que pudieras adquirir ese producto estrella de la gastronomía murciana, por lejos que estuviera, y ya puestos en camino te llevaría a ese restaurante con vistas al mar o a la huerta que él sabía que no decepcionaba.

Con todos estos datos, ¿es de dudar que dejara un sinnúmero de amigos y un descomunal vacío en la vida de todos? Tendrá que pasar mucho tiempo antes de que, al recorrer el pasillo del departamento, dejemos de buscar con los ojos la luz proveniente de su despacho que nos daba la opción de entrar a compartir un momento de alegría o disolver una preocupación.

## Corresponsales

Luz Braña Rey
luzmari@ine.es
Instituto Nacional de Estadística

Francisco Callealta Barroso
franciscoj.callealta@uah.es
Universidad de Alcalá de Henares

Fernando Reche Lorite
freche@ual.es
Universidad de Almería

Ana Justel
ana.justel@uam.es
Universidad Autónoma de Madrid

Jordi Ocaña
jocana@ub.edu
Universitat de Barcelona

Luis Antonio Sarabia Peinador
lsarabia@ubu.es
Universidad de Burgos

Juan Luis González Caballero
juanluis.gonzalez@uca.es
Universidad de Cádiz

Araceli Tuero
tueroma@unican.es
Universidad de Cantabria

Isabel Molina Peralta
imolina@est-econ.uc3m.es
Universidad Carlos III de Madrid

Licesio Rodríguez Aragón
L.RodriguezAragon@uclm.es
Universidad de Castilla-La Mancha

Susana Muñoz López
smunoz@estad.ucm.es
Universidad Complutense de Madrid

José María Caridad y Ocerín
ccjm@uco.es
Universidad de Córdoba

José Antonio Vilar Fernández
eijoseba@udc.es
Universidade da Coruña

Miguel González Velasco
mvelasco@unex.es
Universidad de Extremadura

Vera Pawlowsky-Glahn
vera.pawlowsky@ima.udg.es
Universitat de Girona

Rocío Raya Miranda
rraya@ugr.es
Universidad de Granada

Beatriz Hernández Jiménez
beatriz.hernandez@dmat.uhu.es
Universidad de Huelva

Emilio Lozano Aguilera
elozano@ujaen.es
Universidad de Jaén

David Alcaide López de Pablo
dalcaide@ull.es
Universidad de la Laguna

María Eva Vallejo Pascual
ddeevp@unileon.es
Universidad de León

Carles Capdevila Marques
ccm@matematica.udl.es
Universitat de Lleida

Carmen Morcillo Aixelá
aixela@uma.es
Universidad de Málaga

Marc Almiñana Alemany
marc@umh.es
Universidad Miguel Hernandez

Lázaro Cánovas
lcanovas@um.es
Universidad de Murcia

Susana Montes Rodríguez
montes@uniovi.es
Universidad de Oviedo

Dolores Romero Morales
Dolores.Romero-Morales@sbs.ox.ac.uk
University of Oxford

Maria del Pilar Moreno Navarro
mpmornav@upo.es
Universidad Pablo de Olavide

Pilar Muñoz
pilar.munyoz@upc.edu
Universitat Politécnica de Catalunya

Javier Alcaraz Soria
jalcaraz@eio.upv.es
Universitat Politécnica de Valencia

Ana Fernández Militino
militino@unavarra.es
Universidad Pública de Navarra

Antonio Alonso Ayuso
antonio.alonso@urjc.es
Universidad Rey Juan Carlos

Juan Carlos Fillat Ballesteros
juan-carlos.fillat@dmc.unirioja.es
Universidad de la Rioja

María Teresa Santos
Martín maysam@gugu.usal.es
Universidad de Salamanca

María José Lombardía Cortiña
mjoselc@usc.es
Universidade de Santiago de Compostela

Antonio Beato Moreno
beato@us.es
Universidad de Sevilla

José Domingo Bermúdez Edo
Jose.D.Bermudez@uv.es
Universitat de Valencia

María Cruz Valsero Blanco
mcruz@eio.uva.es
Universidad de Valladolid

Alberto Rodríguez Casal
arodriguez@uvigo.es
Universidade de Vigo

Fernando Plo
fplo@unizar.es
Universidad de Zaragoza