BEIO

Boletín de Estadística e Investigación Operativa

Volumen 32
Número 2
Julio 2016

ISSN: 2387-1725

Revista Oficial de la Sociedad de Estadística e Investigación Operativa

BEIO Revista Oficial de la Sociedad de Estadística e Investigación

BEIO (Boletín de Estadística e Investigación Operativa) es una revista que publica cuatrimestralmente artículos de divulgación científica de Estadística y de Investigación Operativa. Los artículos pretenden abordar tópicos relevantes para una gran mayoría de profesionales de la Estadística y de la Investigación Operativa, primando la intención divulgativa sin olvidar el rigor científico en el tratamiento de la materia en cuestión. Las secciones que incluye la revista son: Estadística, Investigación Operativa, Estadística Oficial, Historia y Enseñanza y Opiniones sobre la Profesión.

BEIO nació en 1985 como Boletín Informativo de la SEIO (Sociedad de Estadística e Investigación Operativa). A lo largo de los años ha experimentado una continua evolución. En 1994, aparece publicado el primer artículo científico y desde entonces el número de artículos científicos publicados ha ido creciendo hasta que en 2008 se segregan del Boletín los contenidos relacionados con la parte informativa y comienza a perfilarse como revista de divulgación de la Estadística y de la Investigación Operativa.

Los artículos publicados en BEIO están indexados en Scopus, MathScinet, Biblioteca Digital Española de Matemáticas, Dialnet (Documat), Current Index to Statistics, The Electronic Library of Mathematics (ELibM), COMPLUDOC y Catálogo Cisne Complutense.

La Revista está disponible online en www.seio.es/BEIO.

Normas para el envío de artículos

Los artículos se enviarán por correo electrónico al editor asociado correspondiente o al editor de la Revista. Se escribirán en estilo article de Latex. Cada artículo ha de contener el título, el resumen y las palabras clave en inglés sin traducción al castellano. Desde la página web de la revista se pueden descargar las plantillas tanto en español como en inglés, que los autores deben utilizar para la elaboración de sus artículos.

Edita SEIO
Facultad de CC. Matemáticas
Universidad Complutense de Madrid
Plaza de Ciencias 3, 28040 Madrid

ISSN: 2387-1725

Boletín de Estadística e Investigación Operativa
Vol. 32, No. 2, Julio 2016

Índice

Boletín de Estadística e Investigación Operativa
Vol. 32, No. 2, Julio 2016, pp. 91-95

Editorial

M. Carmen Aguilera Morillo
Departamento de Estadística
Universidad Carlos III de Madrid
✉ mariacarmen.aguilera@uc3m.es

Por análisis de datos funcionales se conoce a un conjunto de técnicas estadísticas desarrolladas con objeto de resolver problemas reales en los que los datos observados son curvas, o funciones en general, que proceden de la observación de una variable aleatoria funcional. El caso más conocido es el de los procesos estocásticos, cuyas realizaciones son funciones dependientes del tiempo. Los primeros resultados obtenidos en este campo datan de 1974, donde J. C. Deville publicó un artículo de gran rigor científico titulado *Méthodes statistiques et numériques de l'analyse harmonique*. Sin embargo, es en 1997 cuando el análisis de datos funcionales adquiere forma como tal, tomando como punto de referencia el libro *Functional Data Analysis* publicado por los profesores J. O. Ramsay y B. W. Silverman.

En los últimos años se ha consolidado como un tema puntero de investigación estadística que está dando lugar a una amplia variedad de publicaciones en revistas de alto impacto, tanto desde un punto de vista teórico, como aplicado. Prueba de ello es el gran número de trabajos publicados en esta área desde 1997, según la base de datos de ISI Web of Knowledge. Para tener una idea más clara al respecto, en la Figura 1 se ha representado la distribución del número de artículos publicados en revistas del JCR desde 1997 hasta la actualidad.

Siguiendo la tendencia internacional, en España también han aumentado en los últimos años tanto el número de investigadores, como el de resultados importantes en este tema, posicionándose entre los 9 países que más artículos han publicado en esta área (ver Figura 2). Otro aspecto a destacar es el gran número de campos de aplicación del análisis de datos funcionales, siendo Medicina el campo donde más se aplica, según muestra la Figura 3. Todo esto justifica el interés por crear un grupo nacional especializado en el análisis de datos funcionales.

El grupo Análisis de Datos Funcionales (ADF) es uno de los grupos de trabajo más recientes de la Sociedad de Estadística e Investigación Operativa (SEIO). Este grupo surge por iniciativa de la profesora Ana María Aguilera del Pino, Catedrática del Departamento de Estadística e I. O. de la Universidad de

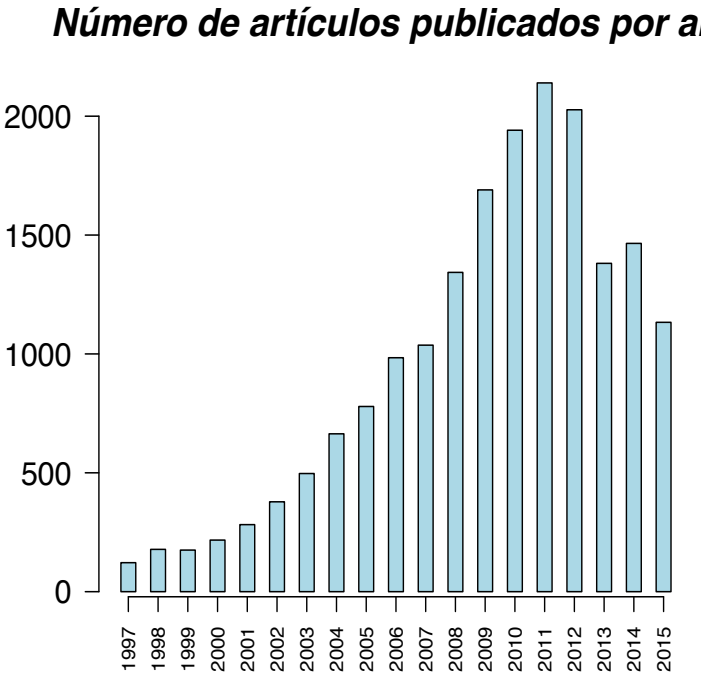


Figura 1: Distribución del número de artículos publicados sobre FDA en revistas indexadas en el JCR, desde 1997 hasta 2015. Fuente: Scopus.

Granada, con el objetivo de crear un grupo especializado de investigadores nacionales que desarrollan su actividad en este campo de la estadística. Esta propuesta fue apoyada por un grupo de doce profesores e investigadores de diversas universidades españolas, tales como, la Universidad Autónoma de Madrid, Universidad de Cantabria, Universidad Carlos III de Madrid, Universidad de Granada, Universitat Politècnica de Catalunya y la Universidad de Santiago de Compostela. Finalmente, la creación del grupo ADF fue aprobada en el Consejo Ejecutivo de la SEIO en Murcia en febrero de 2009, durante la celebración del XXXI Congreso Nacional de Estadística e Investigación Operativa y las V Jornadas de Estadística Pública.

Desde su formación en 2009, el grupo ADF ha tenido como principal objetivo la realización de actividades que fomenten la comunicación y la colaboración entre sus miembros. En un primer periodo comprendido entre 2009 y 2012, y bajo la coordinación de Ana María Aguilera del Pino, tuvo lugar la I reunión de trabajo del grupo en Santander en junio de 2011, en la que aparte de comunicaciones invitadas, sesión de póster y talleres se realizó un encuentro bilateral con el Instituto de Hidráulica Ambiental de la Universidad de Cantabria. La siguiente jornada temática tuvo un carácter internacional y se realizó en Granada. El tema central de esta reunión fué la regresión funcional y contó con una ponencia

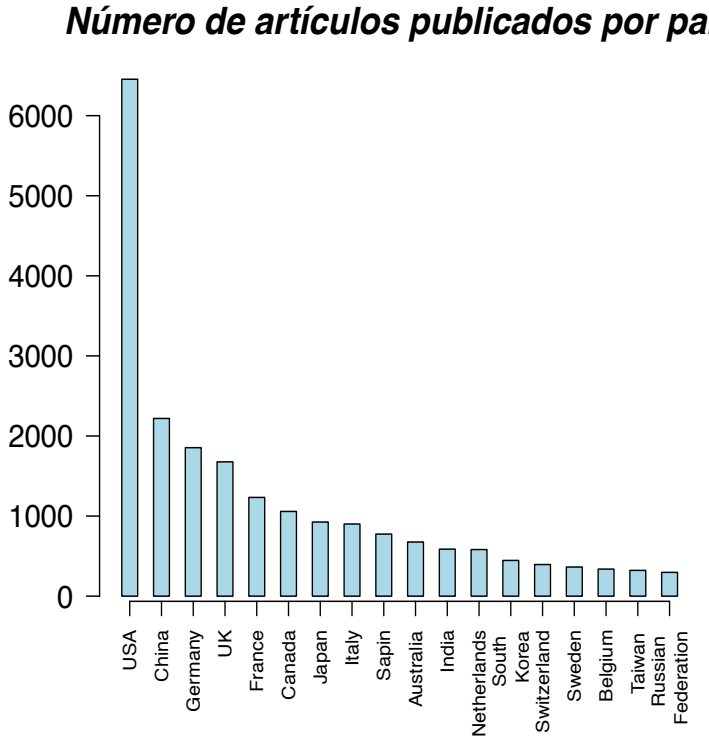


Figura 2: Distribución del número de artículos publicados sobre FDA en revistas indexadas en el JCR, desde 1997 hasta 2015, según el país del autor principal. Fuente: Scopus.

invitada del profesor Hans-Georg Müller (University of California, Davis).

Desde 2013 hasta mayo de 2015 se inicia una segunda etapa del grupo ADF coordinado en esta ocasión por el profesor Pedro Delicado de la Universitat Politècnica de Catalunya. En el marco del programa ECAS (*European Courses in Advanced Statistics*), el grupo ADF participó en el curso sobre *Functional and Complex Structure Data Analysis* celebrado en Castro Urdiales en septiembre de 2013, contando en esta ocasión con la presencia del insigne profesor Peter Hall (University of Melbourne, Australia). En 2014 tuvo lugar en Cádiz la IV reunión del grupo, fomentando sinergias con otras entidades, tales como el Instituto de Ciencias Marinas de Andalucía (ICMAN) del CSIC.

Desde mayo de 2015 hasta la actualidad, el grupo es coordinado por la profesora M. Carmen Aguilera Morillo. Durante esta última etapa, una de las prioridades ha sido dar visibilidad al grupo entre los más jóvenes, consiguiendo la incorporación de investigadores predoctorales y nuevos doctores. Gracias a la ayuda activa de una gran parte del grupo, en noviembre de 2015 se pudo celebrar con éxito en Madrid el *I International Workshop on Advances in Functional Data*, contando en esta ocasión con la intervención del profesor Philip T. Reiss (New York University School of Medicine). Además, dado el interés y la nece-

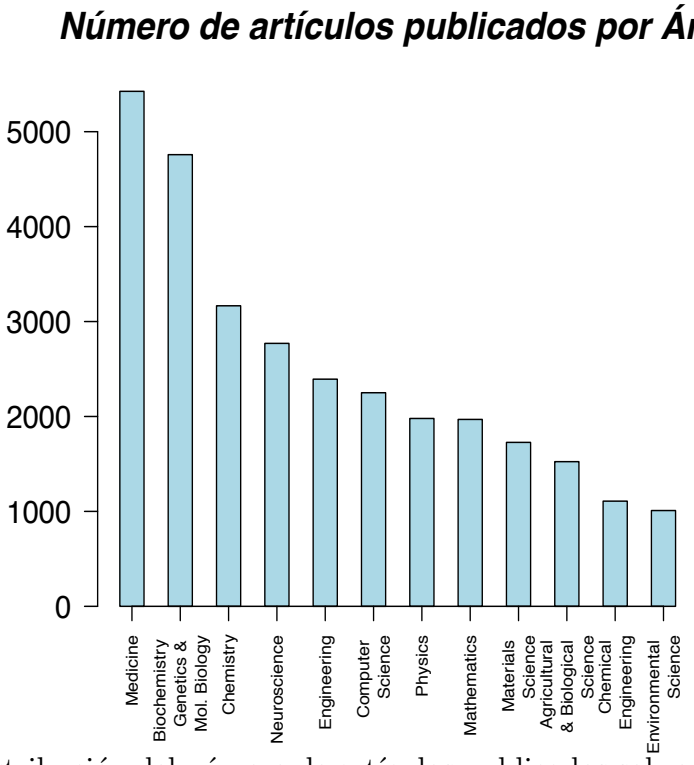


Figura 3: Distribución del número de artículos publicados sobre FDA en revistas indexadas en el JCR, desde 1997 hasta 2015, según el área de conocimiento. Fuente: Scopus.

sidad por conocer nuevas herramientas computacionales para el tratamiento de bases de datos de grandes dimensiones, se programó un taller sobre manejo masivo de datos con R (*Working in R with massive data*), impartido por Carlos Gil Bellosta. Finalmente, resaltar que todas las actividades realizadas por el grupo ADF han sido parcialmente financiadas por la SEIO.

En el marco internacional, también existen grupos de trabajo especializados en el análisis de datos funcionales. Entre otros, cabe destacar el grupo de trabajo *fdawg* del Mailman School of Public Health de la Uiversidad de Columbia, donde desarrollan y aplican las técnicas más novedosas sobre ADF en el campo de la medicina y la salud pública. A nivel europeo, el *European Research Consortium for Informatics and Mathematics* (ERCIM) cuenta con una serie de grupos de trabajo muy consolidados, entre los que se encuentra el grupo *Computational and Methodological Statistics*. Este grupo a su vez está formado por grupos especializados en diversos temas de estadística, entre los que se encuentra el grupo *Statistics for Functional Data*, donde participan algunos investigadores españoles. Entre sus principales actividades, destacar las sesiones especializadas sobre ADF que se organizan dentro del congreso internacional del ERCIM, y que anualmente atraen a investigadores de todo el mundo punteros en este campo de

investigación. Otro grupo relevante es el grupo de trabajo STAPH de Toulouse que inició en 2008 la organización de la serie de workshops IWFOS (International Workshop on Functional and Operatorial Statistics) cuyo objetivo es presentar las últimas tendencias en la investigación en estadística funcional a través del intercambio de ideas y la promoción de la colaboración entre investigadores de diferentes países. Los próximo congresos ERCIM y IWFOS se celebrarán en Sevilla del 9 al 11 de diciembre de 2016 y A Coruña del 15 al 17 de Junio de 2017, respectivamente.

Finalmente, me gustaría aprovechar esta publicación para agradecer a los miembros del grupo ADF la buena acogida que tuvo la última reunión, y animarles a mantener y consolidar nuestro grupo a nivel nacional e internacional. Además quisiera agradecer en nombre de todo el grupo la implicación y el excelente trabajo de los anteriores coordinadores, los profesores Ana Aguilera y Pedro Delicado. A título individual, quisiera cerrar este artículo animando a todos los investigadores nacionales interesados en el análisis de datos funcionales a formar parte del futuro del grupo ADF.

Referencias

[1] Deville. J. C. (1974). Méthodes statistiques et numériques de l'analyse harmonique. *Annales de l'INSEE*, **15**, 3-101.

[2] Ramsay, J. O., and Silverman, B. W. (1997). *Functional Data Analysis*, Springer-Verlag, New York (USA).

[3] Grupo de trabajo ADF - Sociedad de Estadística e Investigación Operativa. In: http://fda.seio.es/

[4] fdawg - Columbia University.
In: https://www.mailman.columbia.edu/research/functional-data-analysis-working-group

[5] ERCIM WG on Computational and Methodological Statistics.
In: http://www.cmstatistics.org/

[6] STAPH: Groupe de Travail en Statistique Fonctionnelle et Opératorielle.
In: http://www.math.univ-toulouse.fr/staph/

Boletín de Estadística e Investigación Operativa
Vol. 32, No. 2, Julio 2016, pp. 96-111

Estadística

A Review of the Developments on Integral Priors for Bayesian Model Selection?

Juan A. Cano
Departamento de Estadística e Investigación Operativa, Universidad de Murcia
✉ jacano@um.es

Diego Salmerón
Servicio de Epidemiología, Consejería de Sanidad, IMIB-Arrixaca, Murcia
CIBER Epidemiología y Salud Pública (CIBERESP)
Departamento de Ciencias Sociosanitarias, Universidad de Murcia
✉ dsm@um.es

Abstract

For the sake of objectivity it is a common practice in Bayesian model selection using default priors. However, these priors are usually improper yielding indeterminate Bayes factors that preclude the comparison of the models. Because of this some approaches have been proposed to obtain more refined default prior distributions avoiding the indetermination of their associated Bayes factors. Among these approaches, a special mention is deserved for the intrinsic priors that were introduced in Berger and Pericchi, 1996. Another important development, the expected posterior priors, appeared in Pérez and Berger, 2002. A special mention is also due to the criteria based priors, a summary of which appears in Bayarri *et al.*, 2012.

Here, we mainly focus on the integral priors, that were presented in the germinal paper Cano *et al.*, 2008, comparing them with the priors above mentioned. These integral priors have been further developed in Cano *et al.*, 2007a, and Cano *et al.*, 2007b, where they were used to analyze the random effects model, Cano and Salmerón, 2013, where an extension was introduced to deal with more sophisticated problems, and applied to binomial regression models in Salmerón *et al.*, 2015. Cano *et al.*, 2016, was devoted to present a methodological introduction that could be useful as a user guide for possible practitioners.

This review paper is intended to be a presentation of the state of the art regarding the developments of the integral priors. One of the main advantages of this methodology is that it can be applied to compare both

nested and non-nested models. Another one is that integral priors are invariant σ-finite measures for two parallel Markov chains which simulation very often can be carried out easily and therefore these Markov chains can be used to approximate the corresponding Bayes factor.

Keywords: Objective Bayesian model selection, Intrinsic priors, Expected posterior priors, Criteria based priors, Integral priors.
AMS Subject classifications: 62F15.

1. Preview

Selecting prior distributions for Bayesian estimation and model selection issues is an old problem for which some solutions have been proposed in the last decades. Default priors have preferably been used but usually they are improper prior distributions $\pi^N(\theta) = ch(\theta)$, where $h(\theta)$ is a function whose integral diverges, and the constant $c > 0$ is arbitrary. In estimation this is not a problem since the posterior does not depend on c. However, in model selection problems when we have two models, $M_i : \mathbf{x} \sim f_i(\mathbf{x} \mid \theta_i)$, $i = 1, 2$, and default priors $\pi_i^N(\theta_i) = c_i h_i(\theta_i)$, $i = 1, 2$, the Bayes factor

$$B_{21}^N(\mathbf{x}) = \frac{m_2^N(\mathbf{x})}{m_1^N(\mathbf{x})} = \frac{c_2 \int f_2(\mathbf{x} \mid \theta_2) h_2(\theta_2) d\theta_2}{c_1 \int f_1(\mathbf{x} \mid \theta_1) h_1(\theta_1) d\theta_1},$$

depends on the arbitrary ratio c_2/c_1. Since we are dealing with model selection a problem arises that needs to be solved: the indetermination of the ratio c_2/c_1. The proposed procedure consists in adjusting these default priors to produce priors that avoid the indetermination problem.

A solution for this problem was introduced in Berger and Pericchi, 1996, it consists of using intrinsic priors that are solutions to a system of two functional equations. Nevertheless, very often intrinsic priors are not unique, for instance when model M_1 is nested in model M_2, the system of functional equations reduces to a single equation with two *incognita* that usually has many solutions. Likewise, in the non-nested case the class of intrinsic priors may be very large. For instance, in Cano *et al.*, 2004, it is shown that any couple of equal priors are intrinsic when comparing the double exponential versus the normal location models.

The expected posterior priors were stated in Pérez and Berger, 2002, as another solution for the problem of the indetermination. In this article the authors propose objective priors defined as expected posteriors under some common predictive marginal $m^*(x)$ suitably chosen, that is,

$$\pi_i^*(\theta_i) = \int \pi_i^N(\theta_i \mid x) m^*(x) dx, \tag{1.1}$$

where x is an imaginary minimal training sample. Note that a minimal training

sample is a sample of minimal size for which the posterior $\pi_i^N(\theta_i \mid x)$ is proper, $i = 1, 2$. The main concern when using this methodology is the choice of $m^*(x)$ which is difficult to assess, mostly when comparing non-nested models.

One more alternative solution consists of producing priors satisfying some type of objective criteria. Among the employed criteria some of them are related with consistency in different ways, for instance, chosing the true model as the sample size goes to infinity that is called model selection consistency, see Bayarri *et al.*, 2012, for other consistency criteria. Another, important criterion is exact predictive matching, that is fulfilled when the priors for the two models under comparison $\{\pi_1(\theta_1), \pi_2(\theta_2)\}$ are such that their corresponding predictive marginals $m_1(x)$ and $m_2(x)$ are equal for every imaginary minimal training sample x .

In section 2 we introduce integral priors mentioning some of their good properties and very quickly we go down to its core showing how they work in the theoretical and in the practical way. In section 3 we review a simple case, that is testing a normal mean with known variance. The case was dealt with in Cano *et al.*,2016, where it was treated like a testing problem (not an estimation problem) using improper priors and therefore giving rise to the problem of undefined Bayes factors. Here we review how numerical computation can be carried out through this simple example and that the approximated Bayes factors obtained are very good approximations, in fact this example can be considered as a guide to use integral priors to be applied later to solve more complex problems. In addition the nice property of the integral prior for the complex model of concentrating mass in the null when comparing nested models is exhibited.

In section 4 we review more complex applications that have been carried out using the integral priors methodology, like testing in the scenarios of a Cauchy distribution and binomial regression models.

Section 5 is devoted to summarize the application of integral priors to the one way random effects model developed in Cano *et al* , 2007a, 2007b, and we take advantage of these papers to improve the presentation of some results in section 4 of Cano *et al* , 2007a.

Finally, in section 6 we present some relevant conclusions and outline oncoming research.

2. Introducing the integral priors

Integral priors are a new methodology to deal with Bayesian model selection problems considered as a generalization of hypothesis testing problems, since they allow to compare non-nested models. The integral priors were proposed in Cano *et al.*, 2008, where under mild assumptions it was proved that they are unique up to a multiplicative constant that is canceled out in the computation of the Bayes factor.

Integral priors are introduced in this section and we also state here how they operate. They are related to the expected posterior priors mentioned above as it can be seen from its very same definition. To be concrete, integral priors are defined as the solutions of the system of the two following integral equations

$$\pi_1(\theta_1) = \int \pi_1^N(\theta_1 \mid x) m_2(x) dx \tag{2.1}$$

and

$$\pi_2(\theta_2) = \int \pi_2^N(\theta_2 \mid x) m_1(x) dx, \tag{2.2}$$

where again x is an imaginary minimal training sample and $m_i(x) = \int f_i(x \mid \theta_i)\pi_i(\theta_i)d\theta_i$, $i = 1, 2$. We emphasize that in this system both priors $\pi_i(\theta_i)$, $i = 1, 2$, are the *incognita*.

Several arguments were given in previous papers to derive these equations. They were summarized in Cano *et al.*, 2016, where it is said that "we are dealing with a two steps procedure looking for the greater objectivity. In the first one by considering objective priors for estimation we are letting the data speak by themselves. On the other hand, in the second one, by considering the system of integral equations above, we are letting the objective priors for estimation speak by themselves. In summary, a sensible way to get priors close to the initial default priors, and with predictive distributions $m_1(x)$ and $m_2(x)$ as close as possible, is by means of equations (1) and (2). These equations balance each model with respect to the other one since the prior $\pi_i(\theta_i)$ is derived from the marginal $m_j(x)$; and therefore from $\pi_j(\theta_j)$, $j \neq i$, as an unknown generalized expected posterior prior". After all, when $m_1(x) = m_2(x)$, the integral priors are expected posterior priors.

One good property of the integral priors as stated in Cano *et al.*, 2008, is that in the continuous case, when the Markov chain with transition density

$$Q\left(\theta_1' \mid \theta_1\right) = \int g\left(\theta_1, \theta_1', \theta_2, x, x'\right) dx dx' d\theta_2,$$

where

$$g\left(\theta_1, \theta_1', \theta_2, x, x'\right) = \pi_1^N\left(\theta_1' \mid x\right) f_2\left(x \mid \theta_2\right) \pi_2^N\left(\theta_2 \mid x'\right) f_1\left(x' \mid \theta_1\right),$$

is recurrent then, there exists a unique solution $\{\pi_1(\theta_1), \pi_2(\theta_2)\}$ to the integral equations system up to a multiplicative constant. In this case $\pi_1(\theta_1)$ is the invariant σ-finite measure for $Q(\theta_1' \mid \theta_1)$. In a similar way there exists a parallel Markov chain on the parameter space Θ_2 with the same properties. In addition, if we are unable to explicitly find the unique pair of integral priors, the corresponding Bayes factor can be approximated by simulation. Therefore, we can operate in the theoretical way, finding the invariant measure of the Markov

chain with transition density $Q(\theta_1' \mid \theta_1)$, or in the empirical one, obtaining a realization of this Markov chain and using it to approximate the corresponding Bayes factor. The transition $\theta_1 \to \theta_1'$ consits of the following four steps:

$$\begin{aligned} 1.\ x' &\sim f_1(x' \mid \theta_1) \\ 2.\ \theta_2 &\sim \pi_2^N(\theta_2 \mid x') \\ 3.\ x &\sim f_2(x \mid \theta_2) \\ 4.\ \theta_1' &\sim \pi_1^N(\theta_1' \mid x). \end{aligned}$$

Analogously, beginning in step 3 followed by steps 4, 1 and 2 we obtain the transition density $\theta_2 \to \theta_2'$. That is, we jump from parameters to samples and between models. It is worth to mention that to operate in the empirical way we just need to simulate from the models and the posteriors that is likely to be easy. In Cano *et al.*, 2007a, 2007b, using the theoretical way we obtained a couple of integral priors and its corresponding Bayes factor for the nested case of the one way random effects model.

3. An easy implementation: The case of testing a normal mean with known variance

We consider here the case of testing a normal mean with known variance. Let $\mathbf{x} = (x_1, ..., x_m)$ be a random sample from $N(\theta, \sigma^2)$, where σ^2 is known and we test $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. Of course, the minimal training sample consists of just a single observation x, the default priors are $\pi_1^N(\theta) = \delta_{\theta_0}(\theta)$ and $\pi_2^N(\theta) = c$ and therefore $\pi_1^N(\theta \mid x) = \delta_{\theta_0}(\theta)$ and $\pi_2^N(\theta \mid x) = N(\theta \mid x, \sigma^2)$. Now, equations (2) and (3) provide the integral priors $\pi_1(\theta) = \delta_{\theta_0}(\theta)$ and $\pi_2(\theta) = N(\theta_0, 2\sigma^2)$. On the other hand, considering the four steps above. The Markov chain for the simplest model is constant and equal to θ_0 while the transition $\theta \to \theta'$ of the Markov chain for the complex model is just made of two steps: 1. $x' = \theta_0 + \varepsilon_1$ and 2. $\theta' = x' + \varepsilon_2$, where ε_1 and ε_2 are independent $N(0, \sigma^2)$ random variables. It follows that $\theta' = \theta_0 + \varepsilon$, with $\varepsilon \sim N(0, 2\sigma^2)$ which yields again the $N(\theta_0, 2\sigma^2)$ as the integral prior for the complex model. In summary, in this example we can compute the exact Bayes factor that results to be

$$B_{21}(\bar{\mathbf{x}}) = \frac{1}{\sqrt{2m+1}} \exp\left(\frac{m^2(\bar{\mathbf{x}} - \theta_0)^2}{(2m+1)\sigma^2}\right),$$

or we can easily simulate a Markov chain $\theta_i, i = 1, ..., L$ of length L for the parameter of the complex model and obtain an approximate Bayes factor as

$$B_{21}^L(\bar{\mathbf{x}}) = \frac{\sum_{i=1}^L f(\bar{\mathbf{x}}|\theta_i)/L}{f(\bar{\mathbf{x}}|\theta_0)},$$

where $f(\bar{\mathbf{x}}|\theta)$ is the normal density for the sample mean.

	m						
z (P-value)	1	5	10	20	50	100	1,000
1.645 (0.1)	0.42	0.44	0.49	0.56	0.65	0.72	0.89
1.960 (0.05)	0.35	0.33	0.37	0.42	0.52	0.60	0.82
2.576 (0.01)	0.21	0.13	0.14	0.16	0.22	0.27	0.53
3.291 (0.001)	0.086	0.026	0.024	0.026	0.034	0.045	0.124

Tabla 1: Posterior probabilities of the null hypothesis using the prior proposed in Berger and Sellke, 1987.

	m						
z (P-value)	1	5	10	20	50	100	1,000
1.645 (0.1)	0.41	0.49	0.56	0.63	0.72	0.79	0.92
1.960 (0.05)	0.32	0.37	0.42	0.50	0.60	0.68	0.87
2.576 (0.01)	0.16	0.14	0.16	0.20	0.27	0.34	0.62
3.291 (0.001)	0.045	0.024	0.026	0.031	0.045	0.061	0.166

Tabla 2: Posterior probabilities of the null hypothesis using the integral prior $N(\theta_0, 2\sigma^2)$.

When both models have the same prior probability, the Bayes factor is equal to the ratio of their posterior probabilities and therefore the posterior probability of the null hypothesis is $P(M_1|\bar{\mathbf{x}}) = (1 + B_{21}(\bar{\mathbf{x}}))^{-1}$. The accuracy of B_{21}^L has been illustrated in Cano *et al.*,2016. For it, we took $\theta_0 = 0$ and for several values of m, $m = 1, 5, 10, 20, 30, 50$, and σ, $\sigma = 1, 2, 3$, samples of size m were generated from the $N(\theta, \sigma^2)$ with θ ranging from -1 to 1 and step equal to 0.005. The exact and the approximate posterior probabilities -using $10,000$ iterations of the Markov chain- for the complex model were computed and they were found to be very similar. In all the cases very reasonable results were observed, the lowest probabilities for the complex model were obtained as the sample mean $\bar{\mathbf{x}}$ goes to zero, increasing as $\bar{\mathbf{x}}$ moved away from zero and the sample size m increased. In addition it was observed that increments in σ yield declines in the posterior probability of the complex model.

On the other hand, Berger and Sellke, 1987 carried out a subjective Bayesian analysis for this testing problem, they after a careful near objective discussion chose the $N(\theta_0, \sigma^2)$ prior. For several sample sizes the posterior probabilities of the null hypothesis they obtained are reproduced in Table 1. Likewise, the posterior probabilities of the null hypothesis obtained using the integral priors are in Table 2. The results are in agreement and in both tables the large m phenomenon called Lindley's paradox is observed, see Lindley, 1957.

To finish it is worth to mention a nice property of integral priors that appears in this example and it is likely to be present when comparing nested models. The

integral prior for the complex model concentrates mass in the null, concretely here we have seen how the initial flat prior por the complex model $\pi_2^N(\theta) = c$ yields the integral prior $\pi_2(\theta) = N(\theta_0, 2\sigma^2)$. Moreover, in this case when we iteratively apply the procedure to obtain iterated integral priors, the ones that are obtained for the complex model converge to the null hypothesis. This is straightforward followed iteratively applying the procedure above stated to derive the integral priors from the initial default priors.

4. Some more complex applications

4.1. Testing the location parameter of a Cauchy distribution

The problem of testing the location parameter of a Cauchy distribution has been considered in Cano and Salmerón, 2013, using integral priors with constrained imaginary training samples. Let $\mathcal{C}(x|\theta,\sigma)$ be the Cauchy density with location θ and scale σ:

$$\mathcal{C}(x|\theta,\sigma) = \frac{1}{\pi\sigma\left(1+\left(\frac{x-\theta}{\sigma}\right)^2\right)}.$$

We are interested in testing the hypothesis $\theta = 0$. The application of the integral priors needs the simulation from the two posteriors distribution, the posterior under $M_1 : \theta = 0$, and the posterior under $M_2 : \theta \neq 0$.

Using that the Cauchy density can be written as a mixture of the normal and the gamma:

$$\mathcal{C}(x|\theta,\sigma) = \int_0^{+\infty} N(x|\theta,\sigma^2/\lambda)\mathcal{G}(\lambda|1/2,2)d\lambda,$$

the posterior distribution for the Cauchy parameters $\pi(\theta,\sigma|x)$ given the imaginary minimal training sample $x = (x_1, x_2)$ and the prior $\pi^N(\theta,\sigma) \propto 1/\sigma$, is the marginal of

$$\pi(\theta,\sigma,\lambda_1,\lambda_2|x) \propto \frac{1}{\sigma}N(x_1|\theta,\sigma^2/\lambda_1)N(x_2|\theta,\sigma^2/\lambda_2)\mathcal{G}(\lambda_1|1/2,2)\mathcal{G}(\lambda_2|1/2,2).$$

Therefore the simulation from $(\theta,\sigma) \sim \pi(\theta,\sigma|x)$ can be performed as follows:

1. $\lambda = (\lambda_1,\lambda_2) \sim \pi(\lambda|x)$
2. $\sigma \sim \pi(\sigma|\lambda,x)$
3. $\theta \sim \pi(\theta|\sigma,\lambda,x)$

First, to simulate $\pi(\lambda|x)$, note that

$$\pi(\lambda|x) \propto \mathcal{G}(\lambda_1|1/2,2)\mathcal{G}(\lambda_2|1/2,2)\int \frac{1}{\sigma}N(x_1|\theta,\sigma^2/\lambda_1)N(x_2|\theta,\sigma^2/\lambda_2)d\theta d\sigma$$

$$= \mathcal{G}(\lambda_1|1/2,2)\mathcal{G}(\lambda_2|1/2,2)\frac{1}{2|x_1-x_2|}$$

and therefore the simulation is straightforward. Then, to simulate $\pi(\sigma|\lambda,x)$ and $\pi(\theta|\sigma,\lambda,x)$ note that

$$\pi(\theta,\sigma|\lambda,x) \propto \frac{1}{\sigma}N(x_1|\theta,\sigma^2/\lambda_1)N(x_2|\theta,\sigma^2/\lambda_2)$$

$$\propto \frac{1}{\sigma^3}\exp\left(-\frac{1}{2\sigma^2}\left(\lambda_1(x_1-\theta)^2+\lambda_2(x_2-\theta)^2\right)\right)$$

$$= \frac{1}{\sigma^3}\exp\left(-\frac{1}{2\sigma^2}\left(H_1(\lambda,x)+(\lambda_1+\lambda_2)(\theta-H_2(\lambda,x))^2\right)\right)$$

$$= \frac{1}{\sigma^3}\exp\left(-\frac{H_1(\lambda,x)}{2\sigma^2}\right)\exp\left(-\frac{(\theta-H_2(\lambda,x))^2}{2\sigma^2/(\lambda_1+\lambda_2)}\right),$$

where

$$H_1(\lambda,x) = \lambda_1\lambda_2(x_1-x_2)^2/(\lambda_1+\lambda_2)$$

and

$$H_2(\lambda,x) = (\lambda_1x_1+\lambda_2x_2)/(\lambda_1+\lambda_2).$$

Therefore $\pi(\theta|\sigma,\lambda,x)$ is the normal density with mean $H_2(\lambda,x)$ and variance $\sigma^2/(\lambda_1+\lambda_2)$. Moreover

$$\pi(\sigma|\lambda,x) = \int \pi(\theta,\sigma|\lambda,x)d\theta \propto \frac{1}{\sigma^2}\exp\left(-\frac{H_1(\lambda,x)}{2\sigma^2}\right)$$

and to simulate $\pi(\sigma|\lambda,x)$, we made $v \sim \mathcal{G}(1/2, 2/H_1(\lambda,x))$ and we take $\sigma = 1/\sqrt{v}$.

For model M_1 the posterior distribution is

$$\pi(\sigma|x_1,x_2) \propto \frac{\sigma}{(\sigma^2+x_1^2)(\sigma^2+x_2^2)}.$$

This distribution can be simulated using the probability integral transform solving the equation

$$\frac{\sigma^2+x_2^2}{\sigma^2+x_1^2} = \left(\frac{x_2^2}{x_1^2}\right)^{1-u},$$

where $u \sim U(0,1)$.

For $\theta = 3$, $\theta = 1$ and $\theta = 0$, we have simulated three samples of size 20

dataset	min	max	$b = 30$	$b = 80$	$b = 150$
$(\theta = 3)$	-3.3	42.9	0.0192	0.0197	0.0200
$(\theta = 1)$	-3.7	19.0	0.5346	0.5649	0.5625
$(\theta = 0)$	-8.9	15.8	1.8986	1.9103	1.9663

Tabla 3: Bayes factor B_{12}^A in favor of the simpler model $\theta = 0$, using integral priors for different constraints and for 3 simulated datasets.

from the Cauchy $\mathcal{C}(\theta, 2)$. The associated Markov chain has been simulated for 3 different constraints $|x_i| \leq b$, $i = 1, 2$, with $b = 30$, $b = 80$ y $b = 150$. Table 3 shows the values of the Bayes factors that are obtained and the range of the data.

The Bayes factors in the last row of Table 3 provided evidence in favor of $\theta = 0$, and the p-value associated with a t-test for the simulated data with $\theta = 0$ was 0.418. The p-values that were obtained with the data simulated from $\theta = 3$ and $\theta = 1$ were 0.05715 and 0.007039, and the corresponding Bayes factors provided a similar response. Another dataset was simulated from $\theta = 0$ for which the data range was from -28 to 2.8, the p-value was 0.029, and Bayes factors obtained were 4.0444, 4.0031 y 4.0401, showing more intensively the lack of robustness and consistency of the t-test to departures from normality.

4.2. Hypothesis testing in binomial regression models

The methodology of integral priors has been satisfactorily applied in binomial regression models with a general link function, see Salmerón *et al.*, 2015. Binomial regression models (and specially the logistic regression model) are some of the main techniques used in analytical Epidemiology to estimate the effect of an exposure on an outcome. To test the effect of specific exposure factors we state the problem as a Bayesian model selection one and we solve it using objective Bayes factors with integral priors. We formulate the problem as follows. Suppose that $\{(y_i, x_i);\, i = 1, \ldots, n\}$ are independent observations, where $y_i \sim Ber(p_i)$ is a Bernoulli distributed random variable, $x_i = (x_{i1}, \ldots, x_{ik})$ is a vector of covariates, X is the matrix with rows $x_1, \ldots, x_n$, and $g(p_i) = x_i\beta$, $i = 1, \ldots, n$, where $g(p)$ is the link function, and $\beta = (\beta_1, \ldots, \beta_k)^{\mathrm{T}} \in \Theta \subseteq \mathbb{R}^k$ is the vector of the regression coefficients with $x_{ik} = 1$. For a given value $k_0 \in \{1, \ldots, k-1\}$ we want to test the hypothesis $H_0 : (\beta_1, \ldots, \beta_{k_0}) = (0, \ldots, 0)$ *versus* $H_1 : (\beta_1, \ldots, \beta_{k_0}) \neq (0, \ldots, 0)$. Each hypothesis provides a competing model to explain the sample data. This hypothesis testing is equivalent to the

problem of selecting between the models M_1 and M_2, with

$$\begin{aligned} M_1: \quad & y_i \mid x_i, \theta_1 \sim Ber(p_i),\ g(p_i) = x_i\theta_1 \quad (i = 1, \ldots, n) \\ & \theta_1 = (\theta_{11}, \ldots, \theta_{1k})^{\mathrm{T}} \in \Theta_1 \subseteq \mathbb{R}^k,\ \theta_{1j} = 0 \quad (j = 1, \ldots, k_0), \end{aligned}$$

$$\begin{aligned} M_2: \quad & y_i \mid x_i, \theta_2 \sim Ber(p_i),\ g(p_i) = x_i\theta_2 \quad (i = 1, \ldots, n) \\ & \theta_2 = (\theta_{21}, \ldots, \theta_{2k})^{\mathrm{T}} \in \Theta_2 \subseteq \mathbb{R}^k. \end{aligned}$$

To compute the posterior probability of each model one needs the specification of the prior distributions. In the literature diffuse, vague, or flat priors and objective ones like the Jeffreys prior, 1961, or the reference prior (Bernardo, 1979; Berger and Bernardo, 1989), are the methods commonly used to estimate the parameters of regression models. Since the Jeffreys prior for binomial regression models is usually a proper distribution, see Ibrahim and Laud, 1991, and Chen, Ibrahim, and Kim, 2008, it can be used to compute Bayes factors for testing H_0 *versus* H_1. However, the Jeffreys's prior does not concentrate mass around the null model (see, e.g., Casella and Moreno, 2006, Casella and Moreno, 2009, and references therein), and therefore the Jeffreys prior is not appropriate for Bayesian model selection.

It is important to note that in regression models there exist different training samples, one for each set of rows of the design matrix with the appropriate dimension. To overcome this issue, in linear models, Berger and Pericchi, 2004, have suggested that imaginary training samples can be obtained by first randomly drawing rows from the design matrix and then generating data from the regression model. We have adapted the above procedure to deal with binomial regression models and the simulation of the Markov chains which need the simulation of training samples. For a detailed description of the associated Markov chain we refer the reader to Salmerón *et al.*,2015.

Breast cancer mortality

The following example is used to illustrate the application of integral priors in binomial regression models. We study the relation of receptor level and stage with the 5-year survival indicator, in a cohort of women with breast cancer, see Greenland, 2004.

The logistic link function was used for this example, and we tested the effect of the receptor. The maximum likelihood estimates exhibits an association between receptor level and mortality, with 2.51 as the estimate for the odds ratio and a p-value of 0.02. We have approximated the integral priors $\pi_1(\theta_1)$ and $\pi_2(\theta_2)$ based on the simulated Markov chains. For $T = 1000$, 5000, and 10,000, we have run 50 Markov chains of length T, and we have approximated the posterior probability using importance sampling. Table 4 shows the mean and the standard deviation of the 50 estimates of the posterior probability of model M_2: there is a high probability of a true association between receptor level and mortality.

	$T = 1000$	$T = 5000$	$T = 10000$
Mean	0.710	0.722	0.726
Standard deviation	0.020	0.010	0.008

Tabla 4: Estimates of the posterior probability of model M_2, based on 50 Markov chains of length T and an importance sampling approximation supported by T simulations.

Figure 1 shows the marginal integral priors for model M_2. These marginal priors concentrate mass around zero, although the marginal prior for the coefficient of the receptor level is more concentrated (the null hypothesis is that this coefficient is equal to zero). The first row provides the priors for the coefficient of the receptor level and the intercept, the second row corresponds to the stage.

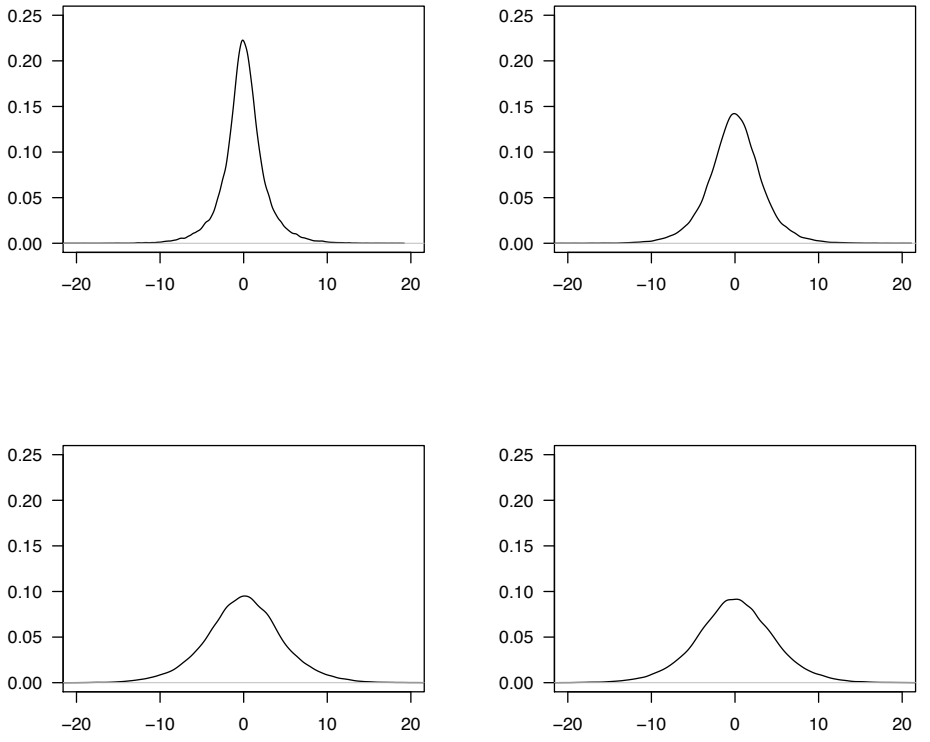


Figura 1: Non-parametric approximations to the integral priors , 2 (top, left: receptor level; top, right: intercept; bottom, left and right: stage) based on 50,000 iterations of the associated Markov chain.

5. Revisiting the application of the integral priors to the one way random effects model

5.1. The problem

We consider the random effects model

$$M : y_{ij} = \mu + a_i + e_{ij}, i = 1, ..., k;\ j = 1, ..., n,$$

where the variables $e_{ij} \sim N(0, \sigma^2)$ and $a_i \sim N(0, \sigma_a^2)$, $i = 1, ..., k;\ j = 1, ..., n$, are independent. We are interested in the selection problem between models with parameters:

$$M_1 : \theta_1 = (\mu_1, \sigma_1, 0) \text{ and } M_2 : \theta_2 = (\mu_2, \sigma_2, \sigma_a).$$

The default priors we use to derive the integral priors in equations (2.1) and (2.2) are the reference priors $\pi_1^N(\theta_1) = c_1/\sigma_1$ and $\pi_2^N(\theta_2) = c_2\sigma_2^{-2}(1 + (\sigma_a/\sigma_2)^2)^{-3/2}$. Note that $\pi_1^N(\theta_1)$ is the reference prior for model M_1 and $\pi_2^N(\theta_2)$ is the reference prior for model M_2 for the ordered group $\{\sigma_a, (\sigma, \mu)\}$ when $n = 1$. We use the prior $\pi_2^N(\theta_2)$ to keep this section within a methodological level. Under these assumptions the sample densities for the two models are:

$$f_1(\mathbf{y} \mid \theta_1) = \prod_{i=1}^{k} N_n(\mathbf{y}_i \mid \mu_1 \mathbf{1}_n, \sigma_1^2 \mathbf{I}_n)$$

and

$$f_2(\mathbf{y} \mid \theta_2) = \prod_{i=1}^{k} N_n(\mathbf{y}_i \mid \mu_2 \mathbf{1}_n, \sigma_2^2 \mathbf{I}_n + \sigma_a^2 \mathbf{J}_n),$$

where $\mathbf{y}_i = (y_{i1}, ..., y_{in})'$, $\mathbf{y} = (\mathbf{y}_1, ..., \mathbf{y}_k)'$, $\mathbf{1}_n = (1, ..., 1)'$, $\mathbf{I}_n$ is the identity matrix of dimension n and $\mathbf{J}_n$ the square matrix of dimension n with all the entries equal to one.

Let S be the total sum of squares, $\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \overline{y})^2$, decomposed as $S = S_1 + S_2$, where $S_1 = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \overline{y}_i)^2$ and $S_2 = \sum_{i=1}^{k} n(\overline{y}_i - \overline{y})^2$. The Bayes factor $B_{21}^N(\mathbf{y})$ is obtained in Cano *et al.*, 2007a, as

$$B_{21}^N(\mathbf{y}) = \frac{c_2}{c_1} \int_0^\infty (1 + nu^2)^{-\frac{k-1}{2}} \left(1 - \frac{nu^2}{1 + nu^2} \frac{S_2}{S}\right)^{-\frac{nk-1}{2}} (1 + u^2)^{-\frac{3}{2}} du, \quad (5.1)$$

which unfortunately depends on the arbitrary ratio c_2/c_1 and to avoid this indeterminacy we will use integral priors instead of the original default priors. The priors $\{\pi_1^N, \pi_2^N\}$ are integral priors when $c_1 = c_2$. However, to ensure the uniqueness of the integral priors recurrence of their associated Markov chains is needed and as both chains are of the same type (see Cano *et al.*, 2008) we have explored the Markov chain associated with the simpler model.

5.2. Exploring the Markov chain associated with the simpler model

The transition density of the associated Markov chain, $\theta_1 \to \theta_1'$, has been obtained in Cano *et al.*,2007a, as

$$\mu_1' = \mu_1 + \sigma_1\alpha$$

and

$$\sigma_1' = \sigma_1\beta,$$

where

$$\beta = \frac{\sqrt{w}}{4} \mid \varepsilon_3 - \varepsilon_4 \mid \sqrt{z} \mid \xi_1 - \xi_2 \mid,$$

$$\alpha = \overline{\xi} + \frac{\varepsilon_2\sqrt{z}}{2\sqrt{2}} \mid \xi_1 - \xi_2 \mid + \frac{\varepsilon_3 + \varepsilon_4}{2}\sqrt{z} \mid \xi_1 - \xi_2 \mid /2 + \beta\varepsilon_1/\sqrt{2}$$

and $\xi_1, \xi_2, \varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4 \sim N(0,1)$, $u \sim q_1(u) = (1+u^2)^{-3/2}$, $z \sim q_2(z) \propto z^{-3/2}e^{-1/z}$ and $w \sim p(w) \propto w^{-3/2}e^{-1/w}$.

Regarding the autonomous chain (σ_n), $(\log\sigma_n)$ is a recurrent random walk since $E(\log\beta) = 0$ and $E((\log\beta)^2) < +\infty$. Although, for the whole chain (μ_n, σ_n), we have not been able to establish recurrence so far because the second order moments for α and β are not finite. However, under some assumptions the recurrence of the chain could be established and then the integral priors would be unique. This is done in the following proposition.

Proposición 5.1. *Let $\{\pi_1(\theta_1), \pi_2(\theta_2)\}$ be the class of integral priors. Suppose that each integral prior $\pi_1(\theta_1) = \varphi(\sigma_1)$ does not depend on μ_1, then the integral priors are unique up to a multiplicative constant.*

Proof

It follows, see Cano *et al.*, 2008, that $\pi_1(\theta_1)d\theta_1$ is an invariant σ-finite measure for the Markov chain

$$\mu_1' = \mu_1 + \sigma_1\alpha$$

$$\sigma_1' = \sigma_1\beta.$$

If $p_2(\beta)p_1(\alpha|\beta)$ denotes the density function of (α, β) then the transition density $Q(\theta_1'|\theta_1)$ is

$$\frac{1}{\sigma_1^2} p_2\left(\frac{\sigma_1'}{\sigma_1}\right) p_1\left(\frac{\mu_1' - \mu_1}{\sigma_1}\Big|\frac{\sigma_1'}{\sigma_1}\right)$$

and it follows from the invariance property that

$$\varphi(\sigma_1') = \int \frac{1}{\sigma_1} p_2\left(\frac{\sigma_1'}{\sigma_1}\right) p_1\left(r\Big|\frac{\sigma_1'}{\sigma_1}\right) \varphi(\sigma_1) dr d\sigma_1 = \int \frac{1}{\sigma_1} p_2\left(\frac{\sigma_1'}{\sigma_1}\right) \varphi(\sigma_1) d\sigma_1.$$

Therefore $\varphi(\sigma_1)d\sigma_1$ is an invariant σ-finite measure for the recurrent Markov chain $\sigma_1' = \sigma_1\beta$ meaning that $\varphi(\sigma_1)$, and therefore $\pi_1(\theta_1)$, have to be proportional to $1/\sigma_1$, and the proposition is proved.

□

Observación 5.1. *Note that if each integral prior* $\pi_1(\theta_1)$ *can be written as* $\pi_1(\theta_1) = \varphi_1(\mu_1|\sigma_1)\varphi_2(\sigma_1)$ *with*

$$\int \varphi_1(\mu_1|\sigma_1)d\mu_1 = 1, \forall\sigma_1 > 0,$$

then, from the invariance property it follows that

$$\varphi_2(\sigma_1') = \int \frac{1}{\sigma_1} p_2\left(\frac{\sigma_1'}{\sigma_1}\right)\varphi_2(\sigma_1)d\sigma_1$$

and again $\varphi_2(\sigma_1)$*, and therefore* $\pi_1(\theta_1)$*, have to be proportional to* $1/\sigma_1$ *and the integral priors are unique up to a multiplicative constant.*

The performance of the integral priors $\pi_1^N(\theta_1)$ and $\pi_2^N(\theta_2)$ was satisfactorily illustrated with two popular data sets found in Box and Tiao, 1973, pages 246 and 247, respectively. We computed the Bayes factors for the two sets of data, using equation (5.1), with $c_1 = c_2$. The Markov chain associated with the simpler model when we use this methodology provided some insight on how integral priors work out. However, the question of whether or not the integral priors for this problem are unique is still an open problem and we are dealing with it exploring the results that are obtained when the Markov chain for the simpler model is run many times in different conditions. The consistency of this Bayes factor when the number of groups goes to infinity, when the number of observations per group goes to infinity, and when both go to infinity has been recently established in Kang *et al.*, 2015.

6. Conclusions and oncoming research

This paper revises the theory of integral priors. We have explained how integral priors operate in Bayesian model selection and we have illustrated their use with several examples ranging from the simple case of testing a normal mean with known varince to more complex situations.

An automatic tool to compute Bayes factors has been developed as we only have to simulate from the involved models and their posterior distributions once a default prior has been assigned to each model. This will be enough to compute (integral) Bayes factors that will be unique provided that its associated Markov chain is recurrent. This methodology can directly be applied to the comparison of nonnested models contrary to what happens with other methodologies that need to be adapted for it.

Several situations may arise when applying this methodology. If we are able to obtain the unique invariant proper distribution we can straightforward compute the unique integral Bayes factor, this is the case of the problem of testing a normal mean with known variance. Nevertheless if we can just establish null recurrence of the associated Markov chain we even are able to compute the

unique integral Bayes factor.

In some situations for which we were not able to ensure uniqueness of the integral priors they were simulated after imposing a constraint on the imaginary training samples space that implied uniqueness of the integral priors, see Cano and Salmerón, 2013.

In other situations like the one sided testing for the exponential model we were able to state uniqueness of the integral priors but we could not explicitly find them and we solve the problem using approximated Bayes factors; these two problems are solved in Cano and Salmerón, 2013. Future applications of this methodology to multiple comparison and therefore to variable seletion are in progress.

Referencias

[1] Bayarri, M. J., Berger, J. O., Forte, A. and García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *Annals of Statistics*, **40**-3, 1550–1577.

[2] Berger, J. O. and Bernardo, J. M. (1989). Estimating a product of means: Bayesian analysis with reference priors. *J. Am. Statist. Assoc.*, **84**, 200–207.

[3] Berger, J. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, **91**-433, 109–122.

[4] Berger, J. O. and Pericchi, L. R. (2004). Training samples in objective Bayesian model selection. *Ann. Statist.*, **32**, 841–869.

[5] Bernardo, J. M. (1979). Reference posterior distribution for Bayesian inference. *J. R. Statist. Soc. B*, **41**, 113–147.

[6] Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Wesley.

[7] Cano, J. A., Iniesta, M. and Salmerón D. (2016). Explaining the integral priors for Bayesian model selection. Technical Report. Departamento de Estadística e I. O. Universidad de Murcia.

[8] Cano, J. A., Kessler, M. and Moreno E. (2004). On intrinsic priors for nonnested models. *Test*, **13**-2, 445–463.

[9] Cano, J. A., Kessler, M. and Salmerón, D. (2007a). Integral priors for the one way random effects model. *Bayesian Analysis*, **2**-1, 59–68.

[10] Cano, J. A., Kessler, M. and Salmerón, D. (2007b). A synopsis of integral priors for the one way random effects model. *Bayesian Statistics*, **8**, 577–582. Oxford University Press.

[11] Cano, J. A. and Salmerón, D. (2013). Integral Priors and Constrained Imaginary Training Samples for Nested and Non-nested Bayesian Model Comparison. *Bayesian Analysis*, **8**-2, 361–380.

[12] Cano, J. A., Salmerón, D. and Robert, C. P. (2008). Integral equation solutions as prior distributions for Bayesian model selection. *Test*, **17**-3, 493–504.

[13] Casella, G. and Moreno, E. (2006). Objective Bayesian variable selection. *J. Am. Statist. Assoc.*, **101**, 157–167.

[14] Casella, G. and Moreno, E. (2009). Assessing robustness of intrinsic tests of independence in two-way contingency tables. *J. Am. Statist. Assoc.*, **104**, 1261–1271.

[15] Chen, MH., Ibrahim, J. G., and Kim, S. (2008). Properties and implementation of Jeffreyss Prior in binomial regression models. *J. Am. Statist. Assoc.*, **103**, 1659–1664.

[16] Greenland, S. (2004). Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. *Am. J. Epidemiol.*, **160**, 301–305.

[17] Ibrahim, J. G. and Laud, P. W. (1991). On Bayesian analysis of generalized linear models using Jeffreyss prior. *J. Am. Statist. Assoc.*, **86**, 981–986.

[18] Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press, London.

[19] Kang, S., Wang, M. and Lu, T. (2015). On the consistency of the objective Bayes factor for the integral priors in the one-way random effects model. *Statistics and Probability Letters*, **103**, 17–23.

[20] Pérez, J. M. and Berger, J. (2002). Expected posterior priors for model selection. *Biometrika*, **89**-3, 491–512.

[21] Salmerón, D., Cano, J. A. and Robert, C. P. (2015). Objective Bayesian hypothesis testing in binomial regression models with integral prior distributions. *Statistica Sinica*, **25**-3, 1009–1023.

Boletín de Estadística e Investigación Operativa
Vol. 32, No. 2, Julio 2016, pp. 112-129

Investigación Operativa

A brief introduction to two-stage Stochastic Optimization

Julene Escudero
Unidad de Epidemiología Clínica
Biocruces Health Research Institute
✉ julene.escuderoargaluza@osakidetza.eus

María Merino
Dpto. de Matemática Aplicada y Estadística e Investigación Operativa,
Universidad del País Vasco
✉ maria.merino@ehu.eus

Abstract

This article is a brief introduction to two-stage Stochastic Optimization, an area of Mathematical Programming for modeling optimization problems that involve uncertainty. Stochastic modeling consists of optimizing the expected value over a set of possible scenarios, being feasible under (almost) all of them. The so-called here-and-now solution will be compared with the wait-and-see solution and the expected result of using the expected value. Numerical simulations are reported for comparison purposes, where uncertainty is modeled via several discrete probability distributions.

Keywords: Stochastic Optimization, Wait and See, Expected Value, Value of Perfect Information, Value of Stochastic Solution.
AMS Subject classifications: 90C15, 90C05.

1. Introducción

La Programación u Optimización Estocástica es un área de la Investigación Operativa encargada de optimizar modelos que contienen incertidumbre. Los problemas de optimización determinista se formulan bajo parámetros conocidos, pero los problemas reales casi siempre incluyen parámetros inciertos en el momento en que hay que tomar las decisiones. Los problemas estocásticos consideran que los datos desconocidos vienen dados o pueden ser estimados por distribuciones de probabilidad. El objetivo es encontrar una política que sea factible para todos (o casi todos) los posibles casos y optimiza la esperanza sobre

una función de las decisiones y las variables aleatorias. Los modelos se formulan, resuelven analítica o numéricamente y se analizan para proveer de información útil a la persona tomadora de decisiones.

La Optimización Estocástica tiene aplicaciones en un amplio rango de áreas desde las finanzas a la optimización en transporte o energía.

Algunos de los libros fundamentales son Kall et al., 1988, Kall and Wallace, 1994, Prekopa, 1995, Wallace and Ziemba, 2005, Shapiro et al., 2009, Alonso-Ayuso et al., 2009, Birge and Louveaux, 2011, Pflug , 2012, King and Wallace, 2012 y Pflug and Pichler, 2014. Una lista de bibliografía de Programación Estocástica ha sido compilada en van der Vlerk, 2007a y sobre Programación Estocástica Entera en van der Vlerk, 2007b; para publicaciones españolas véanse los libros Alonso-Ayuso et al., 2004 y Ramos et al., 2008, así como una mención a esta disciplina en esta revista en Escudero and López Cerdá, 2012 y Escudero, 2009. La referencia a nivel mundial es la *Stochastic Programming Society* (SPS), Sección Técnica de la *Mathematical Optimization Society* (MOS); a nivel europeo el *European Working Group on Stochastic Programming and Applications* (EWGSP) aprobado en 2012 dentro de la *European Association of OR Societies* (EURO); y a nivel estatal la *Red Temática de Optimización Bajo Incertidumbre* (RETOBI) coordinada por el profesor Andrés Ramos, de la Universidad Pontificia Comillas (Madrid).

Este trabajo se organiza de la siguiente manera: en la Sección 2 se presenta una introducción a la teoría básica de la Optimización Estocástica, también conocida como Optimización bajo Incertidumbre; en la Sección 3 se ilustran los resultados obtenidos en la optimización de un caso práctico en dos etapas para los modelos presentados en base a diferentes distribuciones de probabilidad; finalmente, en la Sección 4 se encuentran las conclusiones.

2. Optimización Estocástica

Un problema lineal determinista consiste en encontrar una solución óptima, que minimice (o maximice) una función objetivo lineal sujeta a un conjunto de restricciones lineales:

$$\begin{array}{ll} min & z = cx \\ \text{sujeto a} & Ax = b \\ & x \geq 0, \end{array} \tag{2.1}$$

donde c es el vector fila de costes de dimensión $1 \times n$, x es el vector columna de variables de decisión de dimensión $n \times 1$, $A \in \mathcal{M}_{m \times n}$ es la matriz de las restricciones y b es el vector columna de términos independientes de las restricciones de tamaño $m \times 1$ (*right hand side*, RHS). En un problema determinista, c, A y b son datos conocidos.

La función objetivo es $z = cx$, donde $\{x | Ax = b, x \geq 0\}$ es el conjunto de soluciones factibles. Una solución óptima factible, x^*, es aquella que cumple con

la desigualdad $cx \geq cx^*$ para cualquier x factible.

Típicamente los problemas lineales tratan de buscar la solución del mínimo coste bajo restricciones de demanda que se deben satisfacer o la solución del máximo beneficio bajo recursos limitados. Dado que maximizar una función objetivo z es lo mismo que minimizar $-z$, sin pérdida de generalidad, este trabajo se desarrollará para problemas de minimización.

2.1. Optimización lineal bajo incertidumbre

Los problemas lineales estocásticos son problemas lineales donde algunos datos no son conocidos con antelación. La incertidumbre se puede representar con variables aleatorias bajo la forma de distribuciones de probabilidad, densidades o medidas de probabilidad. Los problemas que analiza la Optimización Estocástica, consideran que tanto los coeficientes de la función objetivo, como la matriz de restricciones o los términos independientes pueden tener componentes estocásticas.

Una técnica que modeliza y recoge adecuadamente la incertidumbre, es la denominada *Análisis de escenarios*. Esta metodología parte de conocer un conjunto finito de valores de los parámetros estocásticos, representativo del conjunto de todos los posibles valores de los mismos. Habitualmente la toma de decisiones está sujeta a un horizonte temporal, cuyos periodos de tiempo se agrupan en etapas de acuerdo a la incertidumbre.

Definición 2.1. *Una* ***etapa*** *de un horizonte temporal dado, es un conjunto de periodos de tiempo en los que tiene lugar la realización de parámetros inciertos. El conjunto de etapas a lo largo del horizonte temporal se denota mediante* $\mathcal{T}$*.*

Definición 2.2. *Un* ***escenario*** *es una realización de los parámetros inciertos y deterministas a lo largo de las etapas del horizonte temporal. El conjunto de escenarios se representa mediante* Ω*. Dado un árbol de escenarios, el conjunto de nodos se representa mediante* $\mathcal{G}$*.*

La incertidumbre simboliza la condición aleatoria de un problema, la cual se representa en términos del experimento aleatorio. El conjunto de todos los posibles resultados lo representaremos por el conjunto Ω. Los resultados pueden combinarse en subconjuntos de Ω llamados sucesos. Cada suceso elemental $\omega \in \Omega$ determina un escenario $\xi^\omega = (c^\omega, A^\omega, b^\omega)$, es decir, una particular realización de los parámetros aleatorios del problema. La colección de sucesos aleatorios se denota por $\mathcal{F}$, siendo $\mathcal{F}$ una $\sigma - algebra$ de las partes de Ω. A cada evento $A \in \mathcal{F}$ se le asocia el valor $P(A)$, llamado probabilidad, tal que $0 \leq P(A) \leq 1$, $P(\Omega) = 1$ y $P(\cup_{n\geq 1} A_n) = \sum_{n\geq 1} P(A_n)$ si $A_1, \ldots, A_m \in \mathcal{F}$ son incompatibles dos a dos. $(\Omega, \mathcal{F}, P)$ se denomina espacio de probabilidad.

Los problemas de Optimización Estocástica tienen su origen a mediados sel siglo XX derivados de la Optimización Lineal en los estudios de Dantzig, 1955

y Beale, 1955. Consideremos el problema estocástico lineal de dos etapas con recurso fijo:

$$\begin{aligned} \mathcal{Q}_{SP} = \quad & min \;\; cx + E_{\boldsymbol{\xi}}[min \;\; (q^{\omega} y^{\omega})] && && \text{(2.2a)} \\ & s.a. \;\; Ax = b && && \text{(2.2b)} \\ & \qquad T^{\omega} x + W y^{\omega} = h^{\omega} && \forall \omega \in \Omega && \text{(2.2c)} \\ & \qquad x, y^{\omega} \geq 0 && \forall \omega \in \Omega, && \text{(2.2d)} \end{aligned}$$

donde c es el vector fila de los coeficientes de la función objetivo para la variable x de la primera etapa de dimensión $n_1 \times 1$, b es el vector RHS para las restricciones de la primera etapa de dimensión $m_1 \times 1$, A es la matriz conocida para las restricciones de la primera etapa de dimensión $m_1 \times n_1$, h^{ω} es el vector RHS para las restricciones de la segunda etapa de dimensión $m_2 \times 1$, q^{ω} es el vector columna de los coeficientes de la función objetivo para la variable y^{ω} de dimensión $n_2 \times 1$ y por último, T^{ω} es la matriz tecnológica de dimensión $m_2 \times n_1$ y W la matriz de recurso fijo para los distintos escenarios ω de dimensión $m_2 \times n_2$. Los componentes estocásticos del problema vienen dados por el vector $\xi^{\omega} = (q^{\omega}, T^{\omega}, h^{\omega})$, $\omega \in \Omega$.

El conocido como *Problema Determinista Equivalente* (DEP) es

$$\begin{aligned} min \;\; z = \quad & cx + \mathcal{Q}(x) && && \text{(2.3a)} \\ & s.a. \;\; Ax = b && && \text{(2.3b)} \\ & \qquad x \geq 0, && && \text{(2.3c)} \end{aligned}$$

donde $Q(x) = E_{\boldsymbol{\xi}}[Q(x, \xi^{\omega})]$ y $Q(x, \xi^{\omega}) = min_y \{q^{\omega} y \mid Wy = h^{\omega} - T^{\omega} x, \; y \geq 0\}$.

Sea la variable aleatoria $\boldsymbol{\xi}$, la cual consideraremos discreta con un número finito de valores ξ^{ω} con probabilidad $P(\boldsymbol{\xi} = \xi^{\omega}) = p^{\omega}$ tal que $\sum_{\omega \in \Omega} p^{\omega} = 1$. La esperanza matemática se define como $E[\boldsymbol{\xi}] = \sum_{\omega \in \Omega} p^{\omega} \xi^{\omega}$ y la varianza como $Var[\boldsymbol{\xi}] = E[\boldsymbol{\xi} - E[\boldsymbol{\xi}]]^2$. Por simplificar la notación, se denotará ξ^{ω} como ω. La incertidumbre del problema se puede representar mediante un árbol de escenarios, cuyos niveles pueden estar relacionados con los periodos del horizonte temporal, ver Figura 1. El camino que une el nodo raíz con una hoja representa un escenario, una posible realización de la incertidumbre. En cada nodo del árbol existe una variable a decidir, es decir, una decisión que debe ser tomada. En cada etapa hay tantos nodos como realizaciones de los parámetros inciertos y en cada etapa se dispone de la información necesaria para la toma de decisiones. En la Figura 1 se ilustra un árbol de 3 etapas, $\mathcal{T} = \{1, 2, 3\}$ y 6 escenarios, $\Omega = \{1, \ldots, 6\}$, formando un árbol de 10 nodos, $\mathcal{G} = \{1, \ldots, 10\}$.

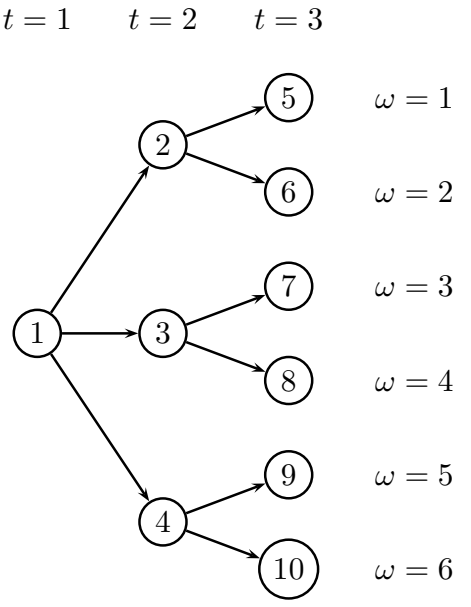


Figura 1: Ejemplo de árbol de escenarios

La representación compacta del problema lineal estocástico de dos etapas de recurso fijo viene dado en (2.4).

$$\begin{array}{rlll} \mathcal{Q}_{SP} = & min & cx + \sum_{\omega\in\Omega} p^\omega q^\omega y^\omega & \\ & s.a. & Ax = b & \\ & & T^\omega x + W y^\omega = h^\omega & \forall \omega \in \Omega \\ & & x, y^\omega \geq 0 & \forall \omega \in \Omega, \end{array} \tag{2.4}$$

donde la función objetivo minimiza el valor esperado bajo el conjunto de escenarios y las restricciones se satisfacen bajo todos y cada uno de los escenarios. En efecto, $\mathcal{Q}_{SP} = E[Z_{SP}] = \sum_{\omega\in\Omega} p^\omega Z^\omega_{SP}$, donde $Z^\omega_{SP} = cx_{SP} + \sum_{\omega\in\Omega} q^\omega y^\omega_{SP}$ siendo x_{SP} y y_{SP} solución de (2.4).

Observación 2.1. *En este trabajo se denota por $Z_{(\cdot)} : \Omega \to I\!R$ la variable aleatoria que representa la función objetivo bajo cada escenario del modelo de optimización, donde $Z_{(\cdot)}(\omega) \equiv Z^\omega_{(\cdot)}$, siendo $Z^\omega_{(\cdot)} = cx_{(\cdot)} + q^\omega y^\omega_{(\cdot)}$, $\omega \in \Omega$, con función de distribución $F_{Z_{(\cdot)}}(z) = P(Z_{(\cdot)} \leq z)$, $\forall z \in I\!R$. $Z_{(\cdot)}$ típicamente puede representar los costes en problemas de minimización o las ganancias en problemas de maximización. En este trabajo, como hemos mencionado anteriormente, $Z_{(\cdot)}$ representa los costes en el modelo $(\cdot)$.*

2.2. Solución WS y valor de la información perfecta (EVPI)

El valor esperado de la solución *espera y observa* (*Wait and See, WS*) se define como el promedio de los valores de las funciones objetivo de cada problema en

cada situación de incertidumbre, véase Madansky, 1960. Es decir, consideramos que podemos *esperar* hasta que la incertidumbre se despeje y calcular el promedio sobre los resultados. Tras resolver $|\Omega|$ problemas (2.5), uno bajo cada escenario $\omega \in \Omega$:

$$\begin{aligned} Z^{\omega}_{WS} = \quad & min \quad cx + q^{\omega} y^{\omega} \\ & s.a. \quad Ax = b \\ & \qquad T^{\omega} x + W y^{\omega} \leq h^{\omega} \\ & \qquad x, y^{\omega} \geq 0, \end{aligned} \tag{2.5}$$

el resultado del WS se obtiene como el promedio de las funciones objetivo resultantes bajo cada escenario, donde $Z_{WS} = (Z^{\omega}_{WS})_{\omega \in \Omega}$:

$$\mathcal{Q}_{WS} = E_{\boldsymbol{\xi}}[Z_{WS}] = \sum_{\omega \in \Omega} p^{\omega} Z^{\omega}_{WS}. \tag{2.6}$$

También se puede implementar en un único problema equivalente:

$$\begin{aligned} \mathcal{Q}_{WS} = \quad & min \quad \sum_{\omega \in \Omega} p^{\omega} (c x^{\omega} + q^{\omega} y^{\omega}) && && \text{(2.7a)} \\ & s.a. \quad Ax^{\omega} = b && \forall \omega \in \Omega && \text{(2.7b)} \\ & \qquad T^{\omega} x^{\omega} + W y^{\omega} = h^{\omega} && \forall \omega \in \Omega && \text{(2.7c)} \\ & \qquad x^{\omega}, y^{\omega} \geq 0 && \forall \omega \in \Omega. && \text{(2.7d)} \end{aligned}$$

donde la función objetivo (2.7a) es la suma ponderada de las $|\Omega|$ funciones objetivo en los respectivos escenarios, y la variable x^{ω} puede tomar distinto valor a lo largo del conjunto de escenarios.

Definición 2.3. *El valor esperado de información perfecta (Expected Value of Perfect Information, EVPI), para problemas de minimización, se define como la diferencia entre el valor $\mathcal{Q}_{SP}$ y $\mathcal{Q}_{WS}$:*

$$EVPI = \mathcal{Q}_{SP} - \mathcal{Q}_{WS} \tag{2.8}$$

Este valor mide la cantidad máxima que la persona tomadora de decisiones estaría dispuesta a pagar a cambio de la información completa y exacta sobre el futuro, véase Raiffa and Schlaifer, 1961.

2.3. Solución EEV y Valor de la Solución Estocástica (VSS)

Una tentación a la hora de resolver un problema con parámetros inciertos es reemplazar todas las variables aleatorias por su valor esperado. Se denomina solución del problema del valor esperado (*Expected Value*, EV) al valor $\mathcal{Q}_{EV}$ que

se obtiene del siguiente problema determinístico:

$$\begin{array}{rlll} \mathcal{Q}_{EV} = & min & cx + E_{\boldsymbol{\xi}}[\mathbf{q}]y & \\ & s.a. & Ax = b & \\ & & E_{\xi}[\mathbf{T}]x + Wy = E_{\boldsymbol{\xi}}[\mathbf{h}] & \\ & & x, y \geq 0 & \end{array} \tag{2.9}$$

El resultado esperado de utilizar la solución que proporciona el valor esperado (*Expected result of using the EV solution*, EEV), $\mathcal{Q}_{EEV}$, es el resultado de implementar las decisiones de primera etapa proporcionadas por el modelo EV (2.9), x_{EV}, en el modelo estocástico (2.4). Es decir,

$$\begin{array}{rlll} Q_{EEV} = cx_{EV} + min & \sum_{\omega\in\Omega} p^{\omega} q^{\omega} y^{\omega} & & \\ s.a. & W y^{\omega} = h^{\omega} - T^{\omega} x_{EV}, & \forall \omega \in \Omega & \\ & y^{\omega} \geq 0, & \forall \omega \in \Omega & \end{array} \tag{2.10}$$

Este problema se puede descomponer en $|\Omega|$ problemas independientes:

$$\begin{array}{rlll} Z^{\omega}_{EEV} = & cx_{EV} + min & q^{\omega} y^{\omega} & \\ & s.a. & W y^{\omega} \leq h^{\omega} - T^{\omega} x_{EV} & \\ & & y^{\omega} \geq 0 & \end{array} \tag{2.11}$$

Así, el valor Q_{EEV} es equivalente al promedio de los valores de las funciones objetivo de cada uno de los problemas (2.11) en los que se ha fijado la variable de la primera etapa x_{EV}.

$$\mathcal{Q}_{EEV} = E_{\boldsymbol{\xi}}[Z_{EEV}] = \sum_{\omega\in\Omega} p^{\omega} Z^{\omega}_{EEV} \tag{2.12}$$

Definición 2.4. *El valor de la solución estocástica (Value of Stochastic Solution, VSS) en problemas de minimización se define como:*

$$VSS = \mathcal{Q}_{EEV} - \mathcal{Q}_{SP} \tag{2.13}$$

Este valor mide el costo por ignorar la incertidumbre.

Para un estudio de los problemas EEV y valor de la solución estocástica VSS en el contexto de optimización estocástica multietapa, véase Escudero et al., 2007.

2.4. Desigualdades básicas

Las siguientes relaciones entre los valores definidos en la sección anterior fueron establecidas por Madansky, 1960.

Proposición 2.1. *Para modelos lineales de minimización se tienen las siguientes desigualdades:*

$$Q_{WS} \leq Q_{SP} \leq Q_{EEV} \tag{2.14}$$

En los modelos de maximización se darían las desigualdades contrarias.

Demostración: Dado que la solución óptima del problema (2.4) es solución factible del problema (2.7), se obtiene la primera desigualdad. Teniendo en cuenta que la solución óptima de (2.10) es una solución factible de (2.4), se obtiene la segunda desigualdad. ■

De esta proposición se concluye que el valor del problema estocástico SP (2.4) nunca podría superar el valor bajo información perfecta, modelo WS (2.7), y que nunca será peor que el resultado esperado de utilizar la solución promedio, modelo EEV (2.10).

Proposición 2.2. *En problemas estocásticos de minimización con coeficientes fijos en la función objetivo y matriz de recurso W fija, se tiene:*

$$Q_{EV} \leq Q_{WS} \tag{2.15}$$

En los modelos de maximización se daría la desigualdad contraria.

Demostración: Se basa en la conocida desigualdad de Jensen, véase Jensen, 1906, que establece que para cualquier función convexa $f(\boldsymbol{\xi})$ de $\boldsymbol{\xi}$: $Ef(\boldsymbol{\xi}) \geq f(E\boldsymbol{\xi})$. Para aplicar este resultado, necesitamos demostrar que $f(\xi) = min\, z(x, \xi)$ es una función convexa de $\xi = (h, T)$. En primer lugar, obsérvese que $z(x, \xi) = cx + \mathcal{Q}(x, h, T) + \delta(x|Ax = b, x \geq 0)$, donde $\delta(x|X)$ es la función indicador del punto x en el conjunto X, es simultáneamente convexa en x, h y T. Sean ξ_1 y ξ_2 tales que $z(x_1, \xi_1) = f(\xi_1)$ y $z(x_2, \xi_2) = f(\xi_2)$, entonces la convexidad se sigue de la siguiente manera: $\lambda f(\xi_1) + (1 - \lambda) f(\xi_2) = \lambda z(x_1, \xi_1) + (1 - \lambda) z(x_2, \xi_2) \geq z(\lambda(x_1, \xi_1) + (1 - \lambda)(x_2, \xi_2)) \geq min_x z(\lambda(x_1, \xi_1) + (1 - \lambda)(x_2, \xi_2)) =$
$f(\lambda\xi_1 + (1 - \lambda)\xi_2)$. ■

De esta proposición se observa que debemos desconfiar del resultado que promete el valor esperado, puesto que podría ser mejor que el que ofrece la solución bajo información perfecta. Además, las soluciones del modelo EV (2.9) podrían no ser implementables bajo alguno de los escenarios, por lo que el modelo EEV (2.10) sería infactible, y por tanto el valor de la solución estocástica inestimable.

Proposición 2.3. *Para cualquier problema estocástico:*

$$0 \leq EVPI \tag{2.16}$$

$$0 \leq VSS \tag{2.17}$$

Demostración: Es inmediata a partir de la Proposición 2.1. ■

Proposición 2.4. *Para problemas estocásticos con matriz de recurso fija y coeficientes en la función objetivo fijos:*

$$EVPI \leq \mathcal{Q}_{EEV} - \mathcal{Q}_{EV} \tag{2.18}$$

$$VSS \leq \mathcal{Q}_{EEV} - \mathcal{Q}_{EV} \tag{2.19}$$

Demostración: Es inmediata a partir de la Proposición 2.2. ■

3. Caso de estudio

En esta sección se exponen las simulaciones del que hemos llamado *problema del granjero*. Contiene la experiencia computacional del problema con 100 escenarios y para cinco distribuciones de rendimientos distintos, de manera que se puedan ilustrar los conceptos previamente expuestos y se puedan analizar las respuestas que nos ofrecen los modelos básicos de Optimización Estocástica presentados en la Sección 2.

La experiencia computacional se ha realizado bajo el sistema operativo Windows 8.1 de 64 bits con un procesador AMD 10, 1.90GHz, 4.00 GB (3.43GB utilizables) RAM y 4 cores. Todos los códigos están implementados en C++ bajo Visual Studio 2013 usando COIN-OR, véase COIN-OR, 2012, Pérez and Garín, 2010.Los resultados y gráficos estadísticos que se muestran en este capítulo se han obtenido mediante R, véase R Project, 2016.

3.1. El problema del granjero (PG)

El problema seleccionado está inspirado en un ejemplo del capítulo 1 del libro Birge and Louveaux, 2011.Un granjero quiere cultivar $\mathcal{I}$ productos en b acres de tierra y durante el invierno debe decidir cuánta tierra destinar a cada tipo de cosecha. El granjero sabe cuantas toneladas de cada producto son necesarias para la comida del ganado. Estas cantidades pueden obtenerse de la granja o comprarse en un almacén y cualquier exceso de producción será vendido. Además, la Comisión Europea impone una cuota a la venta de ciertos productos. Conocido el rendimiento medio base en las tierras y los costes de plantación, se quiere optimizar la distribución de los cultivos que minimice los gastos del granjero.

El problema está compuesto por los siguientes conjuntos, parámetros y variables:

Conjuntos:

$\mathcal{I}_G$, conjunto de materias primas para el alimento del ganado.

$\mathcal{I}_R$, conjunto del resto de materias cultivables.

$\mathcal{I}$, conjunto total de materias, $\mathcal{I} = \mathcal{I}_G \cup \mathcal{I}_R$.

$\mathcal{I}_2$, conjunto de índices para la venta de materias, $\mathcal{I}_2 := \{1, \ldots, I_G + 2I_R\}$.

Ω, conjunto de escenarios representando la incertidumbre.

Parámetros determinísticos:

c, vector de costes de plantación, $c = (c_i)_{i \in I}$.

q^V, vector de precios de venta, $q^V = (q_i^V)_{i \in I_2}$.

q^C, vector de precios de compra, $q^C = (q_i^C)_{i \in I_G}$.

b, acres de tierra disponibles.

r, vector de rendimientos base para los productos cultivados, $r = (r_i)_{i \in I}$.

h, vector de toneladas de producto para el ganado, $h = (h_i)_{i \in I_G}$.

$\bar{w}$, vector de la cuota que impone la Comisión Europea, $\bar{w} = (\bar{w}_i)_{i \in I_R}$.

Parámetros estocásticos:

ξ, vector aleatorio correspondiente a la tasa de variación del rendimiento base, $\xi = (\xi^\omega)_{\omega \in \Omega}$, donde $\xi^\omega = (\xi_i^\omega)_{i \in I}$.

Variable de primera etapa:

x, vector de acres de tierra a cultivar, $x = (x_i)_{i \in \mathcal{I}}$.

Variables de segunda etapa:

y, vector de las toneladas de compra de productos para el ganado, $y = (y_i^\omega)_{i \in I_G, \omega \in \Omega}$.

w, vector de las toneladas de venta de todos los productos, $w = (w_i^\omega)_{i \in I_2, \omega \in \Omega}$.

El modelo lineal estocástico de dos etapas de riesgo neutro para el PG viene dado en (3.1):

$$Q_{SP} = \quad min \quad \sum_{i \in \mathcal{I}} c_i x_i + \sum_{\omega \in \Omega} p^\omega \left(\sum_{i \in \mathcal{I}_G} q_i^C y_i^\omega - \sum_{i \in \mathcal{I}_2} q_i^V w_i^\omega \right) \tag{3.1a}$$

$$s.a. \quad \sum_{i \in I} x_i \leq b \tag{3.1b}$$

$$(r_i \cdot \xi_i^\omega) \cdot x_i + y_i^\omega - w_i^\omega \geq h_i, \ \forall i \in \mathcal{I}_G, \ \forall \omega \in \Omega \tag{3.1c}$$

$$(r_i \cdot \xi_i^\omega) \cdot x_{I_G+i} - w_{I_G+i}^\omega - w_{I+i}^\omega \geq 0, \ \forall i \in \mathcal{I}_R, \ \forall \omega \in \Omega \tag{3.1d}$$

$$x_i \geq 0, \ \forall i \in \mathcal{I} \tag{3.1e}$$

$$y_i^\omega \geq 0, \ \forall i \in \mathcal{I}_G, \ \forall \omega \in \Omega \tag{3.1f}$$

$$w_i^\omega \geq 0, \ \forall i \in \mathcal{I}_2, \ \forall \omega \in \Omega \tag{3.1g}$$

$$w_{I_R+i}^\omega \leq \bar{w}_i, \ \forall i \in \mathcal{I}_R, \ \forall \omega \in \Omega \tag{3.1h}$$

donde (3.1a) es la función objetivo, que representa la esperanza matemática de la variable costes Z, es decir, $E[Z] = \sum_{\omega \in \Omega} p^\omega Z^\omega$, donde para cada escenario

$\omega \in \Omega$ el coste Z^ω se define como:

$$Z^\omega = \sum_{i \in \mathcal{I}} c_i x_i + \sum_{i \in \mathcal{I}_G} q_i^C y_i^\omega - \sum_{i \in \mathcal{I}_2} q_i^V w_i^\omega,$$

(3.1b) es la restricción sobre las plantación de acres, (3.1c) es la restricción de las materias primas para el alimento del ganado, (3.1d) es la restricción para el resto de materias primas, (3.1e), (3.1f) y (3.1g) reflejan el carácter no negativo de las variables y (3.1h) es la restricción sobre la cuota de producción.

3.2. Distribuciones de probabilidad y datos deterministas

El problema estocástico toma la expectativa respecto de la distribución de probabilidad, la cual se da por conocida. Sin embargo, en aplicaciones prácticas no se conoce la distribución y se tiene que estimar desde un conjunto de datos o utilizando juicios subjetivos. En este problema la única fuente de incertidumbre es la tasa de variación del rendimiento de las plantaciones, ξ_i^ω, $i \in \mathcal{I}$, $\omega \in \Omega$. Se ha considerado la misma distribución para todos los productos $i \in \mathcal{I}$, con rango común $[0.5, 1.5]$ de acuerdo a cinco distribuciones de probabilidad. Se han realizado cinco muestras aleatorias simples de tamaño $|\Omega| = 100$, de las siguientes características, cuyo histograma se muestra en la Figura 2.

Tasa 1: Tasa uniforme que proviene de una distribución uniforme en el intervalo $[0.5, 1.5]$, $\mathcal{U}[0.5, 1.5]$.

Tasa 2: Tasa con forma de campana, proviene de una distribución normal de media 1 y desviación estándar 0.2, $\mathcal{N}(1, 0.2)$.

Tasa 3: Tasa con cola a la derecha, proviene de una distribución Beta de parámetros 0.3 y 1, $Beta(0.3, 1)$, desplazada 0.5 unidades a la derecha, dado que la distribución Beta toma los valores en el intervalo $[0, 1]$.

Tasa 4: Tasa con cola a la derecha, proviene de una distribución Beta de parámetros 5 y 1, $Beta(5, 1)$, desplazada 0.5 unidades a la derecha.

Tasa 5: Tasa con cola a la derecha, proviene de una distribución Beta de parámetros 0.3 y 0.3, $Beta(0.3, 0.3)$, desplazada 0.5 unidades a la derecha.

Respecto a los parámetros deterministas, consideramos que el granjero quiere cultivar $\mathcal{I} = \{$trigo, maíz, remolacha$\}$ productos en $b = 500$ acres de tierra. Sea $\mathcal{I}_G = \{$trigo y maíz$\}$ el conjunto de materias primas para el alimento del ganado e $\mathcal{I}_R = \{$remolacha$\}$ el resto de materias, $|\mathcal{I}| = 3$, $|\mathcal{I}_G| = 2$, $|\mathcal{I}_R| = 1$ e $|\mathcal{I}_2| = 4$. Sea $r = (r_i)_{i \in I} = (2.5, 3, 20)$ el vector de rendimientos base para los productos cultivados. Sea $h = (200, 240)$ las toneladas de producto necesarios para alimento del ganado. Sea $c = (150, 230, 260)$ el coste de plantación de los productos $\mathcal{I}$. Sean $q^V = (170, 150, 36, 10)$, $q^C = (238, 210)$ los precios de venta y compra de

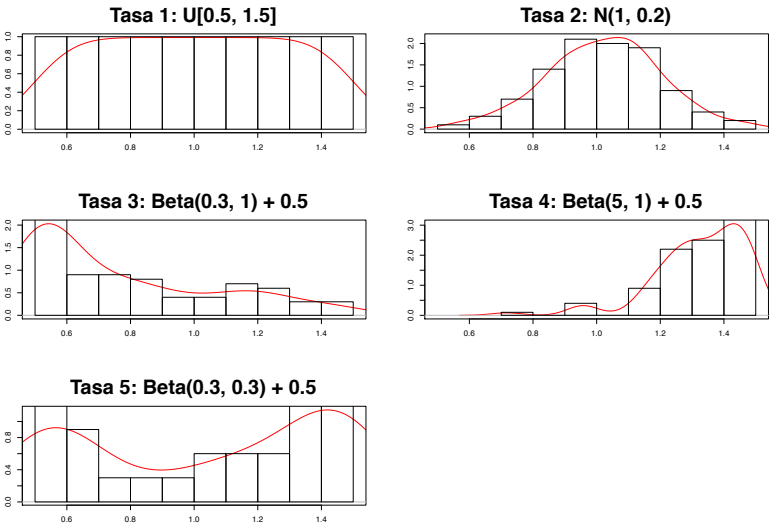


Figura 2: Histograma para las 5 tasas

los productos en $\mathcal{I}_2$ e $\mathcal{I}_G$, respectivamente. Sea $\bar{w}_i = 6000$ la cuota que impone la Comisión Europea para las materias $i \in I_R$.

3.3. Resultados, análisis y discusión

Las cinco tasas de rendimientos nos ofrecen resultados muy diversos en las modelizaciones previamente descritas. En la Tabla 1 está la comparación de los resultados del modelo del promedio EV (2.9), modelo espera-y-observa WS (2.7), modelo estocástico SP (2.3) y el resultado esperado de utilizar la solución que proporciona el valor esperado, modelo (EEV) (2.10). Además se indican las medidas EVPI y VSS.

Tabla 1: Comparación según tasas de EV, WS, SP, EEV, EVPI y VSS

Tasa	$\mathcal{Q}_{EV}$	$\mathcal{Q}_{WS}$	$\mathcal{Q}_{SP}$	$\mathcal{Q}_{EEV}$	EVPI	VSS
Tasa 1: $-$	-118600	-110468	-101648	-96573	8820	5075
Tasa 2: $\cap$	-123973	-120973	-113729	-111980	7244	1750
Tasa 3: $\subset$	-45520	-38148	-32792	-18894	5356	13898
Tasa 4: $\supset$	-185720	-196411	-191003	-190033	5408	970
Tasa 5: $\cup$	-129644	-117264	-110266	-99870	6998	10396

Como se ha mencionado, el EVPI mide la cantidad máxima que la persona tomadora de decisiones, en este caso, el granjero estaría dispuesto a pagar a cambio de la información completa y exacta sobre el futuro, es decir, el rendimiento. El

mayor valor resulta con la tasa 1 uniforme, mientras que con las tasas asimétricas 3 y 4, resulta el menor valor. $EVPI^1 > EVPI^2 > EVPI^5 > EVPI^4 > EVPI^3$. El valor VSS mide el costo que supone ignorar la incertidumbre, el mayor costo por ignorarla es el que ofrece la tasa 3 con cola a la derecha, seguido de la tasa 5 en forma de U (más desfavorables); por el contrario, la tasa 4 con cola a la izquierda (más favorable), es la que menos costo tendría en el caso de ignorar la incertidumbre. $VSS^3 > VSS^5 > VSS^1 > VSS^2 > VSS^4$.

En la Figura 3 se muestra el histograma de Z_{SP} para la Tasa 2 con distribución normal, las curvas de densidad suavizadas para Z_{WS}, Z_{SP} y Z_{EEV}, así como los valores esperados (función objetivo). Se observa que $\mathcal{Q}_{EV} < \mathcal{Q}_{WS} < \mathcal{Q}_{SP} < \mathcal{Q}_{EEV}$, de acuerdo a las Proposiciones 2.1 y 2.2. Y por tanto, se satisfacen las Proposiciones 2.3 y 2.4 en desigualdad estricta. Lo que resalta la conveniencia de utilizar modelos de optimización estocástica frente a la tentación de simplificar el problema sustituyendo los parámetros estocásticos por sus promedios.

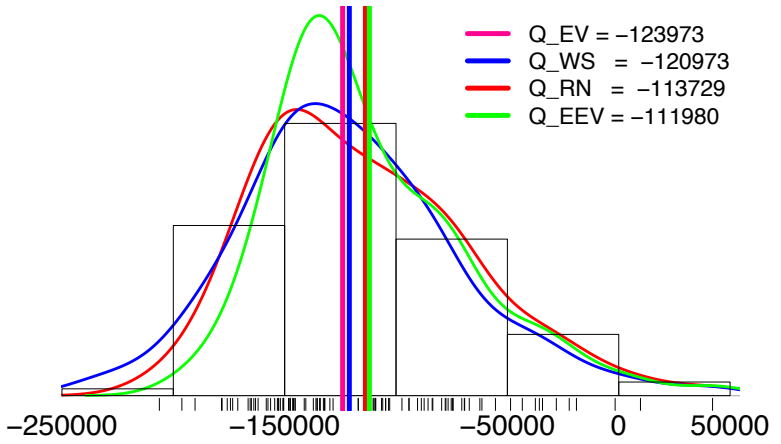


Figura 3: Histograma de Z_{SP}

En la Figura 4 se han recogido los histogramas y valores esperados de Z_{SP}, Z_{WS} y Z_{EEV} para las cinco tasas de variación del rendimiento. Como era de esperar, se observa que la forma de las distribuciones de costos son inversas a las distribuciones de tasas, lo que principalmente se aprecia en las tasas asimétricas 3 y 4. En efecto, a mayor tasa de variación del rendimiento, menor costo esperado y viceversa.

Por último, en la Tabla 2 se muestra la comparativa del vector de soluciones de primera etapa, es decir, la distribución de los productos a cultivar en los acres de tierra disponibles, de acuerdo a todos los modelos presentados: modelo SP (2.4), WS (2.7) (se muestran decisiones en el mejor y peor escenario) y EEV (2.10).

Nótese que las decisiones a implementar son muy diferentes dependiendo del

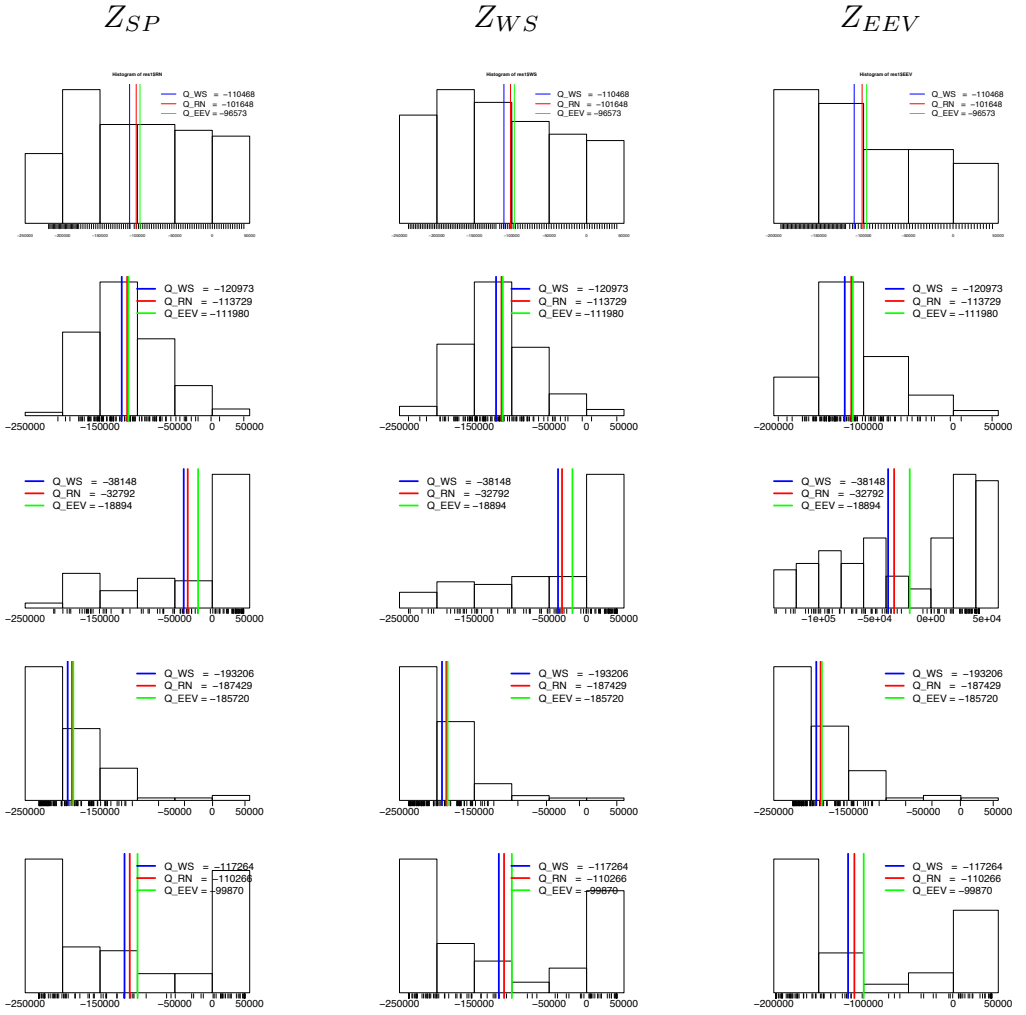


Figura 4: Z_{SP}, Z_{WS} y Z_{EEV} bajo las cinco distribuciones

Tabla 2: Comparativa del vector de decisiones según tasas

Problema	Tasa 1: $-$	Tasa 2: $\cap$	Tasa 3: $\subset$	Tasa 4: $\supset$	Tasa 5: $\cup$
SP	(174, 86, 240)	(150, 82,268)	(157, 88, 255)	(222, 66, 212)	(213, 76, 211)
WS (max)	$(160, 0, 340)$	$(160, 0, 340)$	$(160, 0, 340)$	$(160, 0, 340)$	$(160, 0, 340)$
WS (min)	$(247, 53, 200)$	$(247, 53, 200)$	$(247, 53, 200)$	$(247, 53, 200)$	$(247, 53, 200)$
EEV	(120, 80, 300)	(128, 78, 294)	(106, 0, 394)	(221, 61, 280)	(136, 77, 287)

objetivo que se plantee y de la distribución de los parámetros aleatorios. Las soluciones bajo el mejor y peor escenario no varían según tasas, ya que las cinco distribuciones comparten la misma tasa mínima (-0.5) y tasa máxima (0.5). En el modelo estocástico, con respecto al del valor medio, en general se propone cultivar más del primer producto y menos del segundo. En cuanto al modelo

estocástico según distribuciones, las tasas 2 y 3 son similares, con la menor decisión de cultivar el producto 1, después la tasa uniforme y por último bajo las tasas 4 y 5, similares entre sí, se propone mayor cultivo de dicho producto; de manera inversa ocurre con el tercer producto y no se observan diferencias tan importantes entre tasas para el segundo. En cuanto al EEV el primer producto va aumentando y el tercero decreciendo en las tasas en el siguiente orden: 3, 1, 2, 4 y 5.

4. Conclusiones

En este trabajo se han introducido los conceptos básicos de Optimización Estocástica en dos etapas, comparando con la solución bajo información perfecta y la solución basada en el escenario promedio. Se ha comprobado el cumplimiento de las Proposiciones 2.1 a 2.4 en desigualdad estricta para todos los casos. Por lo que, podemos confirmar lo inadecuado de fiarse de la solución que ofrece el problema del valor esperado, EV, ya que prevé un costo medio mejor que el que ofrece la solución bajo información perfecta, WS. Los resultados obtenidos para los problemas SP, WS y EEV siguen la distribución de costos inversa a la distribución de la tasa de variación del rendimiento. Los resultados obtenidos para el valor de la solución estocástica ponen de relieve el interés de considerar la solución estocástica, pese a su mayor complejidad, frente a otro tipo de modelizaciones más simplistas, basadas en sustituir los parámetros inciertos por sus promedios.

Referencias

[1] Alonso-Ayuso, A., Cerdá, E., Escudero, L., and Sala, R. (2004). *Optimización Bajo Incertidumbre*. Tirant lo Blanch.

[2] Alonso-Ayuso, A., Escudero, L., and Pizarro, C. (2009). *Introduction to Stochastic Programming*. Ediciones Dykinson.

[3] Beale, E. M. (1955). On minimizing a convex function subject to linear inequalities. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 173–184.

[4] Birge, J. R. and Louveaux, F. (2011). *Introduction to Stochastic Programming*. Springer Science & Business Media.

[5] Dantzig, G. B. (1955). Linear programming under uncertainty. *Manag. Sci.*, **1**:197–206.

[6] Êopt. Grupo de investigación en estadística y optimización `http://www.et.bs.ehu.es/~eopt/es`.

[7] Escudero, L. F. (2009). Algunas reflexiones personales sobre la I-O. *BEIO*, **25**:158–171.

[8] Escudero, L. F., Garín, M. A., Merino, M., and Pérez, G. (2007). The value of the stochastic solution in multistage problems. *TOP*, **15**:48–64.

[9] Escudero, L. F. and López Cerdá, M. A. (2012). SEIO and the history of OR in Spain. *BEIO*, **28**:24–56.

[10] EURO. European Association of OR Societies
`https://www.euro-online.org`.

[11] EWGSP. European Working Group on Stochastic Programming and Applications
`http://www.mii.lt/ewgsp/`.

[12] COIN-OR (2012). COIN-OR: Computational infrastructure for operations research. `http://www.coin-or.org/`.

[13] GOE. Grupo de Optimización Estocástica
`http://www.ehu.eus/eu/web/goe/home`.

[14] Jensen, J. L. W. V. (1906). Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Math.*, **30**:175–193.

[15] Kall, P., Ruszczynski, A., and Frauendorfer, K. (1988). Approximation techniques in stochastic programming. *Numerical techniques for stochastic optimization*, pages 33–64.

[16] Kall, P. and Wallace, S. (1994). *Stochastic Programming.* John Wiley.

[17] King, A. J. and Wallace, S. W. (2012). *Modeling with Stochastic Programming.* Springer Series in Operations Research and Financial Engineering.

[18] Madansky, A. (1960). Inequalities for stochastic linear programming problems. *Manag. Sci.*, **6**:197–204.

[19] MOS. Mathematical optimization society
`http://www.mathopt.org/`.

[20] Pérez, G. and Garín, M. A. (2010). On downloading and using COIN-OR for solving linear/integer optimization problems. BILTOKI **2010-05**, UPV/EHU.

[21] Pflug, G. and Pichler, A. (2014). *Multistage Stochastic Optimization.* Springer.

[22] Pflug, G. C. (2012). *Optimization of Stochastic Models: the Interface Between Simulation and Optimization*, volume 373. Springer Science & Business Media.

[23] Prekopa, A. (1995). *Stochastic Programming*. Kluwer Academic Publishers, Dordrecht.

[24] R Project (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria `https://www.r-project.org/`.

[25] Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Boston: Clinton Press, Inc.

[26] Ramos, A., Alonso-Ayuso, A., and Pérez, G. (2008). *Optimización Bajo Incertidumbre*. Universidad Pontificia Comillas.

[27] RETOBI. Red temática de optimización bajo incertidumbre `http://www.iit.upcomillas.es/~retobi/`.

[28] Shapiro, A., Dentcheva, D., and Ruszczyǹski, A. (2009). *Lectures on Stochastic Programming: Modeling and Theory.* MPS-SIAM Book Series on Optimization 9.

[29] SPS. Stochastic programming society `http://stoprog.org`.

[30] van der Vlerk, M. H. (1996-2007a). Stochastic integer programming bibliography. World Wide Web, `http://www.eco.rug.nl/mally/biblio/sip.html`.

[31] van der Vlerk, M. H. (1996-2007b). Stochastic programming bibliography. World Wide Web, `http://www.eco.rug.nl/mally/spbib.html`.

[32] Wallace, S. W. and Ziemba, W. T. e. (2005). *Applications of Stochastic Programming.* MPS-SIAM Book Series on Optimization 5.

Acerca de las autoras

Julene Escudero Argaluza es licenciada en Matemáticas por la Universidad del País Vasco y postgraduada en el máster interuniversitario en Modelización e Investigación Matemática, Estadística y Computación. En la actualidad, trabaja en la Unidad de Epidemiología Clínica del Hospital Universitario Cruces, Biocruces Health Research Institute.

María Merino Maestre es profesora agregada (contratada doctora) en el Departamento de Matemática Aplicada y Estadística e Investigación Operativa de

la Universidad del País Vasco. Sus lineas de investigación comprenden la optimización estocástica, programación lineal, entera y mixta 0-1, gestión del riesgo, aplicaciones financieras, industriales, logísticas y sociales. Pertenece al Grupo de Investigación en Estadística y Optimización (Êopt) coordinado por M. Araceli Garín Martín y el Grupo de Optimización Estocástica (GOE) coordinado por Gloria I. Pérez Sainz de Rozas.

Boletín de Estadística e Investigación Operativa
Vol. 32, No. 2, Julio 2016, pp. 130-147

Estadística Oficial

Iris: International automatic coding system of causes of death. Its use in the Spanish mortality statistics

Jesús Carillo Prieto and Mª Rosario González García
Subdirección General de Estadísticas Sociales Sectoriales
Instituto Nacional de Estadística
✉jesus.carrillo.prieto@ine.es, ✉rosario.gonzalez.garcia@ine.es

Abstract

Statistics on causes of death are the main source of information for epidemiological research or social and health policy decisions. Mortality statistics consider the underlying cause of death. The selection of the underlying cause is based on the guidelines and rules described in the International Classification of Diseases (ICD). Although highly qualified coders perform the selection of the underlying cause, discrepancies in the interpretation of the ICD reduce the homogeneity of mortality statistics at international level. The interest in improving the mortality quality data has prompted the researchers to develop automatic systems for the coding and selection of the underlying cause of death. Iris, the promising automatic coding software used by an increasing number of countries, is the result of many years of effort and international cooperation.

Keywords: Causes of death statistics, underlying cause of death, International Classification of Diseases, automatic systems, coding
AMS Subject classifications: 62P10, 62P25

1. Introducción

La Estadística de Defunciones según la causa de muerte es una de las mayores fuentes de información para la investigación epidemiológica y para la toma de decisiones en políticas sanitarias y sociales. La gran demanda de información en esta materia, tanto a nivel nacional como internacional, obliga a los productores de esta estadística a velar por la calidad y la comparabilidad de los datos.

La codificación manual de la causa de muerte está afectada por los mismos problemas que la codificación manual en general: requiere tiempo, necesita numerosos recursos humanos y económicos y es muy sensible a los errores sistemáticos de los codificadores. No obstante, es necesario señalar que la codificación manual

de la causa de muerte tiene además sus propios problemas específicos, ya que debe basarse en las directrices descritas por la Organización Mundial de la Salud (OMS) en la Clasificación Internacional de Enfermedades (CIE), véase OMS, 2016a.

Esta clasificación consta de tres volúmenes:

- Volumen 1: Lista tabular (1157 páginas). Presenta una lista de enfermedades clasificadas por grandes grupos (22 capítulos) con códigos asignados a cada una de ellas. Presenta más de 12.000 entradas diferentes. Contiene además, el Informe de la Conferencia Internacional de la Décima Revisión, una clasificación histológica y de comportamiento de los tumores, definiciones y Listas tabulares.

- Volumen 2: Manual de Instrucciones (244 páginas). Contiene las reglas de codificación.

- Volumen 3: Índice alfabético de la Clasificación de Enfermedades (758 páginas). Contiene las enfermedades clasificadas por orden alfabético para su mejor manejo, así como tablas de medicamentos y productos químicos.

Esta clasificación se caracteriza por su complejidad y sus numerosas excepciones. Por ello, la experiencia y los conocimientos médicos del codificador son fundamentales en esta tarea. La formación de un buen codificador puede llevar entre uno y dos años, lo que hace aconsejable disponer de un equipo de profesionales estable.

A partir de las condiciones informadas en el certificado médico de defunción por el médico certificador (causa inmediata, causa intermedia, causa inicial o fundamental y otros procesos) el codificador, aplicando las reglas del volumen 2 de la CIE, debe seleccionar la causa básica de defunción, es decir, la enfermedad o lesión que inició la cadena de acontecimientos patológicos que condujeron directamente a la muerte, o las circunstancias del accidente o violencia que produjo la lesión fatal. Las tabulaciones y los análisis estadísticos están basados en la causa básica.

La calidad de la estadística de defunciones según la causa de muerte está ligada fundamentalmente al nivel de detalle de información proporcionado por el médico en el certificado de defunción y a la variabilidad en la interpretación de la CIE durante el proceso de selección manual de la causa básica de defunción (Harteloh et al., 2010).

La información aportada por el médico certificador es de vital importancia en este proceso, para ello no solamente es necesario conocer la patología padecida por el fallecido, sino tener también la formación necesaria para hacer la correcta certificación de la muerte (Consejería de Sanidad de la Región de Murcia, 2016).

Por su parte, la búsqueda de una solución, que elimine o minimice las discrepancias de interpretación de la CIE, es lo que ha llevado a varios países a lo

largo de las últimas décadas a desarrollar sistemas automáticos de codificación y selección de la causa básica.

2. Antecedentes

A finales de los años 60 el US National Centre for Health Statistics desarrolló el primer codificador automático para causas de muerte: Mortality Medical Data System (MMDS), véase US National Centre for Health Statistics, 2016.

Este codificador consta de cuatro partes funcionales:

- SUPERMICAR, trasforma el literal de enfermedades informadas en el certificado en códigos ERN (Entity Reference Numbers). Cada ERN engloba los términos sinónimos de una enfermedad (por ejemplo, tendrá la misma ERN el cáncer de laringe, el tumor maligno de laringe, cáncer laríngeo, neoplasia laríngea, etc.). Sin embargo, las distintas formas de presentarse una enfermedad, constituyen entidades nosológicas diferentes, tales como la insuficiencia cardiaca, insuficiencia cardiaca aguda, insuficiencia cardiaca congestiva. Solamente admite entradas de texto en inglés.

- MICAR, tiene como entrada los ERN, aplica las tablas internas que interrelacionan los diferentes códigos de las enfermedades descritas (reglas de codificación múltiple), en el último paso transforma el código ERN resultante del proceso en códigos CIE. Por ejemplo, se describe en un mismo certificado un cáncer de pulmón y un cáncer de mama cada uno con su ERN asignada por el módulo anterior. Micar aplica las reglas y establece que si existen dos cánceres descritos en el mismo certificado y uno de ellos pertenece a un sitio al que frecuentemente van las metástasis (caso del pulmón), éste tomará el código de tumor secundario, considerando al otro cáncer como primario. Los códigos salientes de este módulo serían un cáncer de pulmón secundario y un cáncer de mama primario.

- ACME, tiene como entrada los códigos CIE que proporciona el módulo Micar, aplica las tablas de decisión de secuencias lógicas basadas en las reglas del volumen 2 de la CIE y da como salida el código de la causa básica de defunción. Siguiendo el ejemplo anterior, por la regla 2 de selección se seleccionaría el tumor secundario de pulmón por ser el primero informado, posteriormente se aplica la regla de modificación C. Asociación que establece que con un tumor secundario en presencia de uno primario se debe seleccionar el primario, cáncer de mama. Así la causa básica de defunción en este certificado sería el cáncer de mama.

- TRANSAX, compila las causas múltiples de defunción.

El problema que surge a nivel internacional reside en que estos códigos ERN no siempre se pueden ligar a expresiones no inglesas. Esto explica que el codificador

sea adoptado en su totalidad por países de habla inglesa como Estados Unidos, Reino Unido y Australia, mientras que otros países como Suecia, Brasil, Francia o Dinamarca implanten únicamente ACME en el proceso de sus estadísticas.

Es inevitable pensar que utilizando únicamente ACME, podríamos obtener la causa básica de defunción, sin embargo, se trataría de una codificación incompleta y en muchos casos incorrecta, ya que no dispondríamos del módulo MICAR que interrelaciona entre sí los códigos informados en cada certificado.

Por otra parte, la opción de trabajar exclusivamente con ACME obliga a introducir directamente en el codificador los códigos CIE correspondientes a las condiciones informadas en el certificado médico de defunción o bien a desarrollar un sistema que lo realice de forma automática. La clave podría residir en la disponibilidad de un certificado electrónico. Los elevados costes de esta herramienta y diversas dificultades legales e institucionales relacionadas con la firma electrónica del médico certificador, explican que su uso no esté muy extendido, al menos entre los países de la Unión Europea, con excepciones, como puede ser el caso de Dinamarca y la reciente implantación del certificado electrónico en Portugal (Assembleia Da República Português, 2012).

Algunos países, ante la necesidad de un sistema de codificación automático partiendo de un texto desarrollaron, en los primeros años de la década de los 90, codificadores automáticos a nivel nacional. Suecia, uno de los países pioneros en esta materia, desarrolló Mikado. Aunque se trató de un reto interesante, el problema al que se enfrentó fue la gran inversión económica que supuso un proyecto de esta envergadura, ya que se trataba de un sistema que continuamente debía ser actualizado.

La conclusión a la que llegaron otros países como Francia y Alemania, que también se embarcaron en proyectos similares, iba en la misma línea: el desarrollo de sistemas automáticos a nivel nacional, además de no ser viable económicamente, no resolvía el problema de comparabilidad a nivel internacional de las estadísticas de defunciones según la causa de muerte (Eurostat, 1998a).

3. Iris, sus orígenes

No fue hasta 1997 cuando Lars Age Johansson (National Board of Health and Welfare, Suecia) y Gerard Pavillon (Centre d'épidémiologie sur les causes médicales de décès, Francia) empiezan a trabajar sobre un proyecto común de codificación automática. Pocos años después, Eurostat promueve iniciativas para la mejora y coordinación de las estadísticas de defunciones según la causa de muerte. Según reflejan varios informes de la Oficina de Estadística Europea (Pavillon and Johansson, 2001, Eurostat, 1998b), entre las distintas recomendaciones se encontraba el desarrollo de sistemas de codificación automático. Aunque esta idea tardó en ser aceptada por muchos países, actualmente nadie duda de que es la única solución para conseguir la comparabilidad a nivel internacional, además

de un ahorro de recursos, sin olvidarnos de las mejoras en los indicadores de puntualidad y oportunidad de los datos.

En 2001, en el marco de un proyecto de Eurostat, Lars y Pavillon describieron un primer borrador de lo que podría ser un sistema de codificación automático común para todos los países de la Unión Europea. Posteriormente, se pasó al desarrollo de un software que cumpliera con las especificaciones descritas en el informe: Iris. En 2004, Alemania y Hungría se unieron al proyecto, lo que permitió integrar en el grupo de trabajo profesionales con diferentes cualificaciones. Tras la anexión de Italia en 2010, el Core Group Iris quedó constituido por cinco países.

Las subvenciones de Eurostat y la decisión de Alemania de ceder un programador a tiempo completo, permitieron grandes avances en el desarrollo del software, de manera que en 2011 se pudo ofrecer a Eurostat y a los Estados Miembros una versión suficientemente buena de Iris.

Como consecuencia de la crisis y de los recortes a los que tuvo hacer frente Eurostat, Iris fue calificado como una prioridad negativa. Esto significaba que aunque la Oficina de Estadística Europea apoyaba y consideraba fundamental la implantación de Iris en todos los países miembros, no podría destinar más fondos a este proyecto, lamentando y siendo conscientes de las consecuencias que tendría en términos de calidad y comparabilidad de las estadísticas de defunciones si no se conseguía continuar con el desarrollo y la implantación de este codificador automático internacional.

Iris necesitaba un respaldo institucional que garantizase su continuidad. El Institute of Medical Documentation and Information (DIMDI, Alemania) decide acoger dentro de su institución al Instituto Iris, dándole así una estructura legal (Iris Institute, 2016).

Actualmente Iris ha atravesado la frontera europea y son muchos los países de otros continentes que han implantado o están trabajando en la implantación de Iris en sus estadísticas de defunciones. La filosofía de sus fundadores se basa en que el software debe ser gratuito para permitir el acceso a todos los países del mundo interesados en su implantación, por ello, Iris se financia con las aportaciones anuales que de forma voluntaria y en la cuantía que es posible proporcionan algunos países.

4. El software Iris

El software Iris toma como base el sistema americano MMDS. El primer objetivo fue modificar el módulo MICAR de manera que pudiera ser utilizado de forma universal. Para ello, crea un motor interno que transforma el código CIE de un diccionario en cualquier idioma en código ERN que es interpretado por MICAR y tras aplicar las tablas de interrelación lo vuelve a transformar en código CIE para su entrada en ACME, dando este último el código de causa básica. Por

tanto, el reto inicial consistía en conseguir que todos los aspectos relacionados con el idioma se pudieran almacenar en tablas independientes, de forma que no interfiriesen con las tablas de decisión de la causa básica de defunción y que fuesen fácilmente modificables y adaptables a las necesidades de cada usuario. Conseguido este reto, los productores de estadísticas a nivel nacional tendrían que asumir la tarea de la construcción del diccionario de literales diagnósticos en su propio idioma asociados a su código CIE.

Iris comienza trabajando con ACME, pero su finalidad es construir sus propias tablas para la selección de la causa básica. Detrás de esta decisión hay diversas razones, pero fundamentalmente está basada en la falta de documentación sobre el funcionamiento de ACME y en la no inclusión por parte de sus gestores de las actualizaciones de la CIE publicadas anualmente por la OMS. El Reglamento (UE) 328/2011 de la Comisión de 5 de abril de 2011 sobre la estadística de defunciones según la causa de muerte de obligado cumplimiento para los países de la Unión Europea contempla la consideración de estas actualizaciones (European Union, 2011).

Después de este preámbulo general, abordaremos a continuación algunos aspectos a nivel técnico. Tras una sencilla instalación del software Iris, el codificador automático presenta una intuitiva interfaz que facilita el trabajo a los usuarios. Esta interfaz, que por defecto está en inglés, puede ser traducida al idioma deseado a través de una tabla específica incluida en la base de datos. Sobra decir que es responsabilidad, en nuestro caso del Instituto Nacional de Estadística (INE), la traducción al castellano para la implantación en España. Esta traducción se ha puesto a disposición del Instituto Iris para que pueda ser distribuida a otros países de habla hispana que estén interesados en este proyecto.

El diseño del certificado médico de defunción con el que trabaja Iris es el recomendado por la OMS (Figura 1) y coincide con el certificado vigente en España (Figura 2), véase Arimany et al., 2009.

Iris puede ser utilizado en dos modos: modo de entrada de código y modo de entrada de texto. En el modo de entrada de código, el usuario debe introducir los códigos CIE correspondientes a las enfermedades informadas en los certificados médicos de defunción. Este modo es sencillo y práctico para aquellos países, cítese el caso de la Republica Checa, en los que el médico certificador no informa mediante texto sino utilizando los códigos CIE, o en aquellos otros que deseando utilizar Iris no disponen de diccionario de literales diagnósticos. Una vez instalado el software, Iris está preparado para ser utilizado en modo de entrada de código.

En la práctica internacional habitual, el médico certificador cumplimenta el certificado de defunción especificando con texto los distintos diagnósticos y causas de defunción en la Parte I del certificado (causa inmediata, intermedia, inicial o fundamental) y en la Parte II (otros procesos), en este caso Iris tendrá que ser utilizado en modo de entrada de texto. Para ello, se precisa la construcción de un diccionario en el idioma de trabajo que asocie a cada término un código CIE.

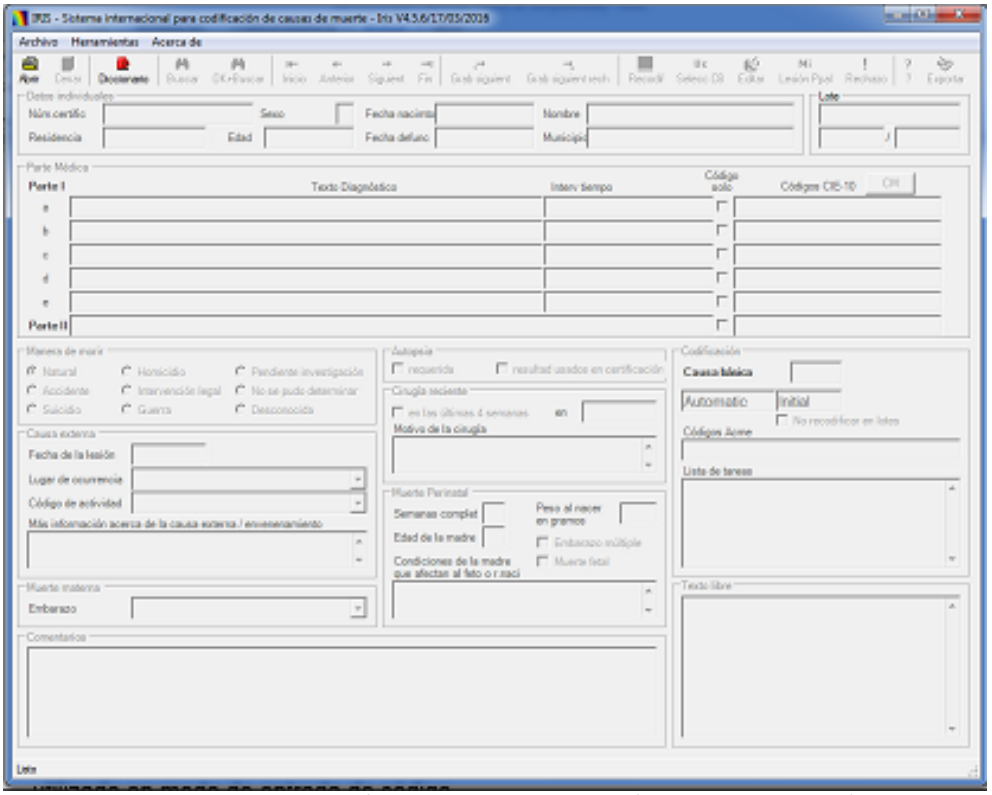


Figura 1: Diseño del certificado médico de defunción en el software Iris.

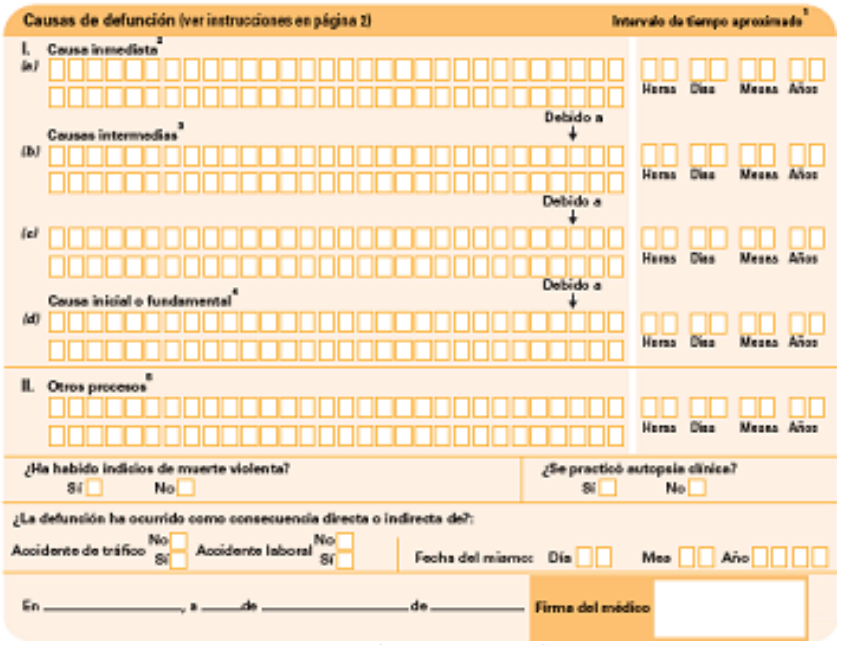


Figura 2: Parte médica del certificado de defunción vigente en España.

Para poder entender cómo se procesa la información en el codificador automático Iris, centrémonos en un par de ejemplos, sin olvidarnos que el objetivo es la selección de la causa básica de defunción atendiendo a las complejas reglas de la CIE:

Ejemplo 1:

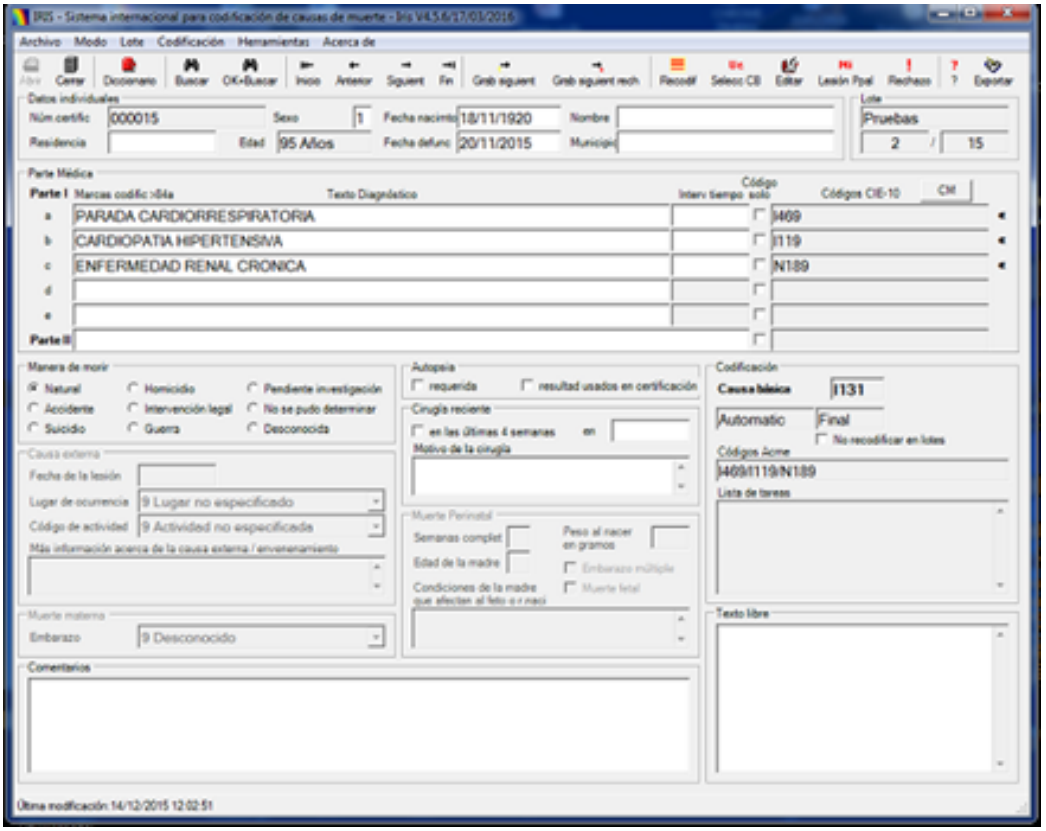


Figura 3: Ejemplo 1.

En este ejemplo, el código final de causa básica no coincide con ninguna de las causas múltiples, esto es debido a que internamente Iris hace una asociación de enfermedades entre la cardiopatía hipertensiva y la enfermedad renal crónica para dar como codificación final una enfermedad más informativa que cada una de ellas por separado, la enfermedad cardiorrenal hipertensiva con insuficiencia renal (I13.1).

Ejemplo 2:

Este ejemplo es un poco más complejo desde el punto de vista del proceso interno de Iris, en este caso se produce un cambio en la codificación múltiple con la valvulopatía aórtica (I35.9), y es que en presencia de una valvulopatía mitral que es reumática, asume que la lesión aórtica también lo es, cambiando por tanto el código inicial del diccionario I359 por I069, presente en la codificación

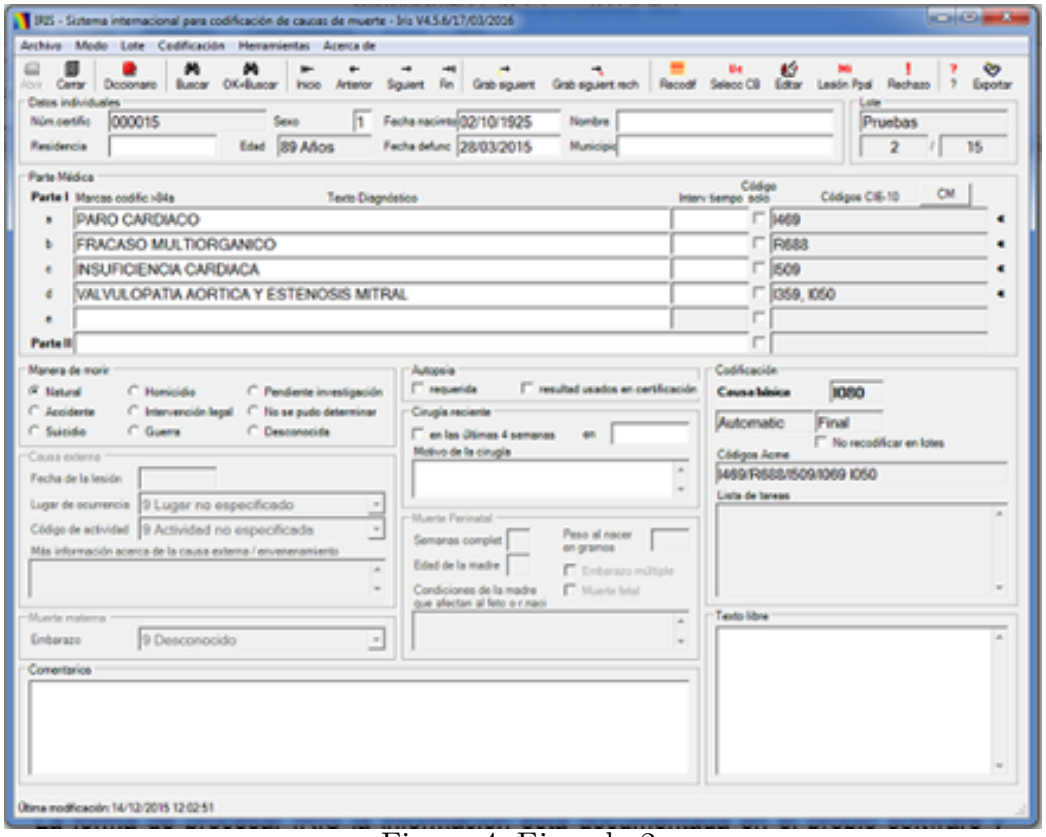


Figura 4: Ejemplo 2.

múltiple del recuadro de Códigos Acme. En este caso se ve la actuación que tiene el módulo MICAR, que interrelaciona las enfermedades descritas tomando decisiones, incluso cambiando el código inicial que aporta el diccionario. Por otra parte, la causa básica ofrece más información que cada una de las causas múltiples, puesto que refleja que existe lesión de las válvulas aórtica y mitral en un solo código (I08.0).

El codificador automático está diseñado para poder procesar lotes de registros, que el usuario habrá elaborado según sus criterios, bien sea por mes de defunción, provincia... Podremos especificar a Iris si queremos que se procese la totalidad del lote o parte de éste. Una vez procesado el lote, Iris nos ofrece un resumen sobre el estado de codificación del lote: número de certificados codificados correctamente, número de certificados rechazados y el motivo del rechazo:

Son varios los motivos por los que un registro puede ser rechazado. Si el motivo del rechazo es por Código (114 en la Figura 5), indica que el literal diagnóstico es un término no registrado en el diccionario; la solución es sencilla: es suficiente con incorporarlo. Si por el contrario el rechazo se debe a Enr, Micar o ACME, el caso debe ser comunicado al Instituto Iris para que lo consideren en su labor de creación de las tablas de decisión, tablas que como se ha mencionado

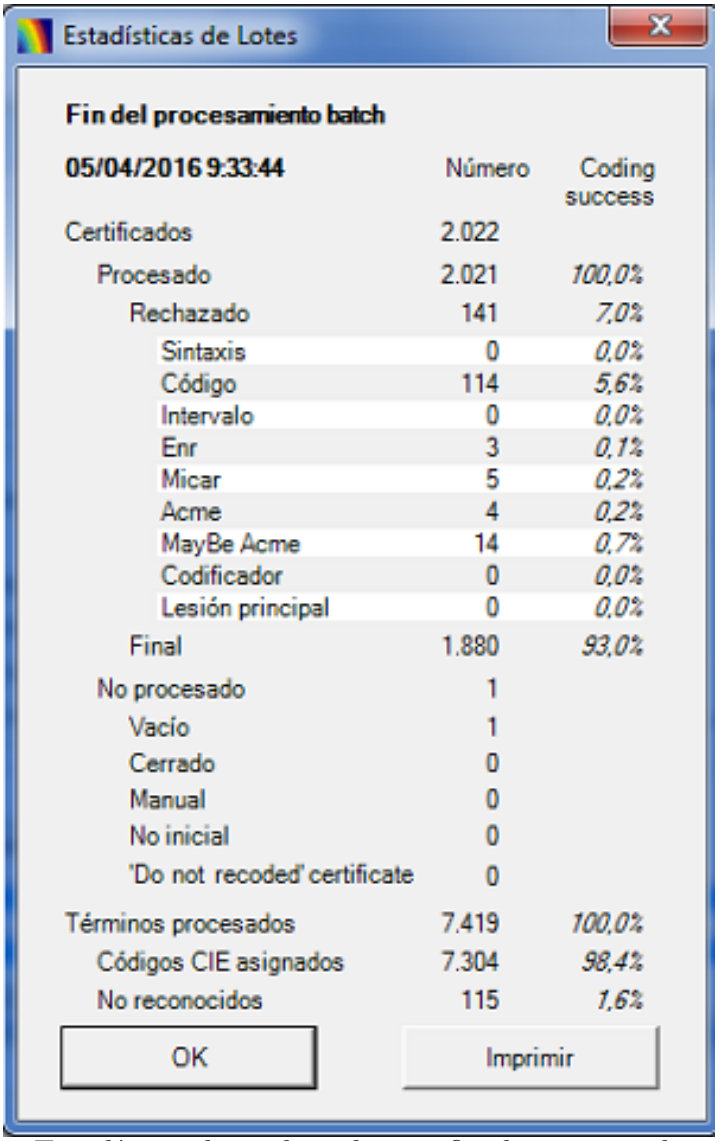


Figura 5: Estadística de un lote de certificados procesados por Iris.

anteriormente se están diseñando para que en algún momento pueda sustituir a ACME. Los casos de MayBe Acme (14 en la Figura 5) nos indican que tras pasar todo el proceso, si existe alguna variable que genera una duda lo marca como tal para solución posterior. La codificación de los casos rechazados se podrá hacer de forma interactiva individualmente. Estos casos, generalmente son complicados y tendrán que ser resueltos por codificadores con elevada experiencia y cualificación de forma manual.

La forma de procesar Iris la información está documentada en el propio software y puede ser consultada para cada uno de los registros, esto permite entender cómo el codificador automático ha determinado una causa básica concreta:

La inclusión anual en el sistema de las actualizaciones de la CIE publicadas

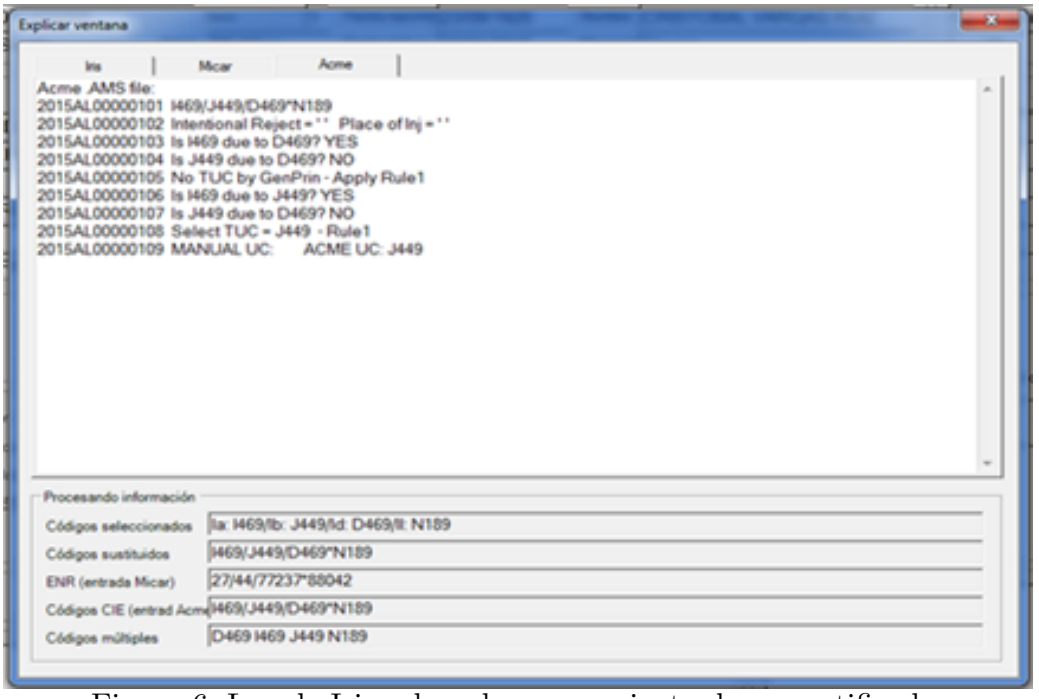


Figura 6: Log de Iris sobre el procesamiento de un certificado.

por la OMS conlleva la incorporación, modificación o eliminación de códigos, y consecuentemente la modificación de las tablas de decisión que conducen a la determinación de la causa básica.

5. Implantación de Iris en España

El INE es el responsable de la Estadística de defunciones según la causa de muerte en España (INE, 2016) y cuenta para ello con la colaboración de las oficinas de estadística y con registros de mortalidad —equipos de codificación— de las comunidades autónomas. Al igual que el resto de los países de la Unión Europea (UE), España lleva a cabo el proyecto bajo el paraguas del Reglamento (UE) 328/2011 de la Comisión de 5 de abril de 2011 sobre la estadística de defunciones (Eurostat, 2016).

El INE sigue el desarrollo del codificador automático Iris prácticamente desde sus inicios. Las pruebas realizadas en 2006 por el equipo de la estadística de defunciones según la causa de muerte con la colaboración de algunas comunidades autónomas confirmaron la compatibilidad de la codificación automática Iris con el proceso de codificación manual realizado en España, que llegó a un 84,2 % de coincidencia a 4 dígitos y a un 89 % de coincidencia a 3 dígitos, garantizando de esta manera la continuidad de las series ante el posible cambio metodológico.

La decisión de implantar el codificador automático Iris conlleva la ardua tarea de la elaboración del diccionario de términos médicos en castellano asociados a su código CIE correspondiente. Las causas de muerte son por lo general in-

formadas por los médicos en el certificado de defunción con un lenguaje rico y variable. La eficacia del diccionario reside en la contemplación de cualquier expresión que el médico certificador pueda incluir en el certificado médico de defunción. Iris cuenta con reglas de estandarización destinadas a normalizar el texto de las causas de muerte, permitiendo de esta manera reducir el tamaño del diccionario. Estas reglas son utilizadas por el creador del diccionario para abordar los sinónimos, las abreviaturas... que puedan ser incluidos en el certificado médico de defunción.

Por ejemplo, la siguiente regla de estandarización:

\b(ARTERIOPATIA|VASCULOPATIA|ENFERMEDAD|PROCESO)\s?
ARTERIO(E)?SCLERO(S[AO]|SIS|TIC[OA])\s?(GENERAL(IZAD[OA])?)?\b

permite transformar 56 entradas diferentes en una sola expresión: ARTERIOSCLEROSIS. La regla de estandarización indica que dichas entradas deben iniciar por la palabra "arteriopatía" o "vasculopatía" o "enfermedad" o "proceso" seguido de "arteriosclerótica" o cualquier variante de esta palabra y llevar o no el término generalizado o generalizada. De ahí que arteriopatía arterioesclerótica, enfermedad arterioesclerótica general y proceso arterioescleroso generalizado se encuentren entre las 56 posibles entradas citadas anteriormente.

Iris ha sido incorporado en el proceso de la estadística de defunciones según la causa de muerte con las defunciones de 2014, para ello ha contado con un diccionario de 102.359 términos y 401 reglas de estandarización. El diccionario actual es el resultado de varios años de trabajo. Para elaborarlo se han considerado distintas fuentes de información: literales teóricos del volumen 1 de la CIE-10, literales procedentes del capítulo 5 del Manual de Causas de Defunción (MCD) —se trata de un listado de términos diagnósticos que no figuran en la CIE con códigos asignados—, diccionarios de codificadores automáticos utilizados en la Comunidad de Madrid y la Comunidad Valenciana, literales rechazados por Iris en las distintas fases de pruebas y literales del diccionario MMDS de México. La consideración de distintas fuentes para la elaboración del diccionario implica la labor de búsqueda y eliminación de duplicados y la construcción de reglas de estandarización que reduzcan el tamaño del mismo. La creación de un diccionario que ofrezca resultados satisfactorios pero que a su vez no sea excesivamente grande facilitará las actualizaciones que anualmente deberán llevarse a cabo.

El diccionario Iris español proporciona unos resultados de codificación y selección de causa básica en torno al 90 % y aunque estos resultados se pueden calificar de óptimos, es importante señalar, que se trata de un diccionario vivo que debe enriquecerse y mejorarse de forma continua. Siempre existirán términos nuevos que será necesario incluir, términos que la experiencia indique que no se utilizan y por consiguiente será aconsejable eliminar y reglas de estandarización que se podrán perfeccionar.

De forma prácticamente simultánea a la creación del diccionario Iris, y ante la ausencia de un certificado médico de defunción electrónico, se ha tenido que abordar el problema de trascripción de términos médicos en formato papel a formato electrónico para poder disponer de las variables de texto que constituyen el fichero de entrada a Iris. Para ello se ha optado por un sistema de lectura óptica de caracteres (OCR), lo que ha obligado a una modificación en 2009 del certificado médico de defunción.

Este sistema consta de dos motores de reconocimiento independientes: R1 que reconoce los caracteres aislados y R2 que reconoce toda la información de la línea y permite mejorar el literal en los casos que el médico no centra la letra en su cuadrícula —foco— correspondiente. Cada uno de ellos se enfrenta al diccionario OCR con más de 164.000 entradas, de estructura diferente al usado por Iris pero con las mismas enfermedades, y aplicando complejas fórmulas de análisis lingüístico y probabilístico generan el literal con mayor índice de confianza.

El diccionario OCR ofrece un literal de entrada que es el que usa el sistema de OCR para comparar su lectura y un literal de correspondencia de salida, que es el que nos presenta como resultado. Todos los literales de salida de OCR están en el diccionario que usa Iris con su correspondiente código.

El mecanismo de OCR también incorpora de forma automática en los casos más frecuentes el signo de puntuación "coma" cuando existe en la misma línea del certificado médico de defunción más de una enfermedad, signo que Iris reconoce para asignar códigos independientes.

El sistema obtiene unos resultados satisfactorios que superan el 80 % de aciertos, sin embargo, queda pendiente entre un 18-20 % en los que el índice de confianza es pequeño y por tanto el literal obtenido del diccionario puede ser diferente al que realmente aparece en el certificado médico de defunción por la simple coincidencia de la posición de algunas letras en el literal leído.

Por este motivo es necesario realizar una revisión exhaustiva de literales antes de procesar un lote por Iris. Para simplificar esta tarea, el INE proporciona a las comunidades autónomas responsables de la codificación una herramienta que permite cotejar la imagen del certificado con el resultado del OCR:

La revisión de literales se introduce en el procesamiento de datos como una tarea adicional anteriormente inexistente, sin embargo, es importante señalar que los recursos necesarios son considerablemente inferiores a los empleados en una codificación manual.

Son muchos los investigadores y las autoridades sanitarias que demandan, además de la causa básica de defunción, las causas múltiples. Es decir, el código correspondiente a cada una de las condiciones informadas por el médico en el certificado. La implantación de Iris podrá satisfacer esta demanda de información. La causa múltiple permitirá realizar estudios más específicos en epidemiología y salud pública.

De todos es conocido que se ha producido un importante incremento de

Figura 7: Aplicación del INE de trabajo e intercambio de información con las CCAA.

enfermedades crónicas en las que suelen confluir varias patologías que si bien por sí solas no producen la muerte, sí pueden interactuar precipitándola. Este es el caso, por ejemplo, de la diabetes, una de las enfermedades que están en el punto de mira de las instituciones sanitarias. Cuando la diabetes aparece informada en el certificado médico de defunción no significa que en todos los casos vaya a ser seleccionada como causa básica, ya que dependerá del resto de enfermedades que también hayan sido mencionadas. Por ejemplo, si la diabetes aparece conjuntamente con un cáncer de pulmón, dependiendo del orden en que se describan ambas enfermedades, las reglas de selección de causa básica pueden penalizar la diabetes desplazándola a favor del cáncer de pulmón. De este modo, se pierde la oportunidad de conocer la verdadera dimensión de la diabetes en la mortalidad, de identificar las asociaciones más frecuentes con otras patologías y, como consecuencia, de adoptar medidas de actuación más eficaces. Por tanto, el análisis multicausal, basado en todas las enfermedades que han sido informadas en el certificado, supondrá un valor añadido esencial a la estadística de causas de muerte.

En relación con la tabulación y explotación de datos sobre causas múltiples de muerte pueden referirse a los siguientes objetivos (López-Zazo, 2004):

a) Determinar la densidad de información de los certificados médicos de defunción, es decir, número de causas descritas en cada certificado, que aumentarán cuanto mayor sea la edad del fallecido.

b) Conocer la "topografía" de distribución de dicha información dentro del

certificado médico de defunción.

c) Investigar la frecuencia de aparición para cada código CIE de causas básicas y causas mencionadas que no son seleccionadas como básicas (en un sentido descriptivo y no relacional), así la arteriosclerosis es penalizada por las reglas de selección y frecuentemente pasa a ser un factor contribuyente de otros procesos que se seleccionan como causa básica.

d) Relacionar las causas mencionadas en asociación para cada causa básica de defunción (para el conjunto de las defunciones o para cualquier subconjunto de las mismas), por ejemplo ver qué otras patologías acompañan al cáncer de hígado.

e) Identificar combinaciones de causas que presentan una cierta frecuencia de aparición, en relación o no con determinada selecciones de causa básica. Por ejemplo la coexistencia de enfermedad cardiaca, enfermedad renal e hipertensión.

6. Iris, una cooperación internacional

Si bien Iris está ligado a los nombres de sus dos fundadores, Lars Age Johansson y Gérard Pavillon, y al del resto del equipo del Core Group, es importante señalar que el éxito del codificador automático es también consecuencia de la implicación de todos los países que, en mayor o menor medida, forman parte del proyecto. La creación de una herramienta que avale la comparabilidad de los datos de las estadísticas de defunciones según la causa de muerte a nivel global, no podría entenderse sin una coordinada cooperación internacional.

De forma periódica se dispone de nuevas versiones del software. Dada la casuística que nos podemos encontrar en un certificado médico de defunción, es conveniente que estas nuevas versiones sean testadas por numerosos usuarios, con el fin de garantizar que las mejoras introducidas responden correctamente a los criterios de codificación, así como para comprobar la incorporación de las actualizaciones de la CIE que anualmente publica la OMS. Cabe destacar el papel activo que desempeña el equipo de causas de muerte del INE en esta labor. Además de posibles errores o diferencias de codificación a nivel nacional, se detectan discrepancias a la hora de interpretar las reglas de codificación establecidas por la CIE. A tal fin, el Core Group Iris actúa como interlocutor ante el Mortality Reference Group (MRG), comité de la Familia Internacional de Clasificaciones de la OMS (FIC-OMS) encargado de identificar y resolver los problemas relacionados con la interpretación de la CIE en su aplicación a las estadísticas de mortalidad (OMS, 2016b). Los dictámenes del MRG pueden implicar modificaciones en las tablas de decisión de Iris y homogeneizar criterios internacionales de codificación cuya práctica no había sido detectada hasta este momento.

Anteriormente hemos explicado como Iris se diferencia de los codificadores automáticos previamente diseñados por su independencia con el idioma, en el sentido de que todos aquellos aspectos relacionados con éste se almacenan en tablas separadas que no interfieren en el funcionamiento del software. Este aspecto, que se podría considerar como el único punto de divergencia entre los diferentes países, comporta también una considerable cooperación internacional. Por citar algún ejemplo, el diccionario de Reino Unido junto con el diccionario de Francia han servido de base para la creación del diccionario canadiense, Portugal muestra interés por el diccionario utilizado en Brasil y el diccionario español, siempre ha sido uno de los más esperados por el Core Group Iris, ya que facilitará la implantación del codificador en todos los países de habla hispana que así lo deseen.

Incluso en el hipotético caso de conseguir una versión de Iris que resuelva el cien por cien de los casos a nivel mundial, la comparabilidad de los resultados puede encontrar una mínima fisura en la creación de los diccionarios. Es labor de los expertos en la materia, asignar a cada expresión médica que pueda ser informada en el certificado médico de defunción un código CIE. Puede ocurrir que en expresiones más complejas que no aparezcan suficientemente especificadas en la CIE existan distintos criterios a nivel internacional a la hora de asignar estos códigos. Siendo conscientes de este hecho, se ha optado por poner a disposición de los productores de estadísticas que trabajen con Iris todos los diccionarios elaborados hasta el momento, de manera que se puedan consultar cómo otros países codifican aquellas expresiones que generalmente plantean mayores dudas.

La satisfacción ante un proyecto que garantiza una mejora en la calidad de los datos y la comparabilidad va inevitablemente unida a una de las mayores preocupaciones que siempre acompañan a un cambio metodológico y, que no es otro, que las implicaciones que pueda tener sobre la ruptura de series. Como es fácil imaginar, son múltiples las comparaciones entre la codificación manual y la automática que se han realizado a tal fin. Resultados que han sido compartidos y discutidos a nivel internacional.

Todos los años, durante el mes de septiembre, tiene lugar la reunión anual de usuarios de Iris en Colonia. Esta ocasión brinda la oportunidad a todos los asistentes de compartir los avances y la experiencia en la implantación del codificador automático en sus respectivos países, además de poder realizar sugerencias para la mejora del software. Se trata de una reunión fructífera y enriquecedora que es testigo desde 2009 del crecimiento exponencial del número de países participantes.

7. El futuro de Iris: MUSE

Las futuras versiones de Iris contarán con la incorporación de MUSE (Multicausal and Unicausal Selection Engine), que sustituirá al MMDS. Esta nueva

herramienta utiliza las tablas de decisión internacionales para la selección de causa básica basadas en reglas precodificadas de conformidad con las directrices del volumen 2 de la CIE-10. MUSE (versión 1.0) es el resultado de la colaboración entre la Oficina Federal de Estadística de Alemania y el Instituto Alemán de Documentación Médica e Información (DIMDI).

El módulo lleva incorporado validaciones de causa múltiple y causa básica, códigos de sustitución de la causa múltiple y selección de la causa múltiple. MUSE se implantó en Alemania con las defunciones de 2012. El Core Group Iris evaluó MUSE para la integración en el software internacional Iris. Las pruebas realizadas confirman que en el 96 % de los casos MUSE y MMDS seleccionan la misma causa básica.

La versión de Iris en la que se incorpora MUSE se implementará en España con las defunciones de 2017.

Referencias

[1] Arimany Manso, J., Barbería Marcalain, E. and Rodríguez Sendín, J. J. (2009). El nuevo certificado médico de defunción. *Revista Española de Medicina Legal*, **35**, 36.

[2] Assembleia Da República Português. Lei nº15/2012 de 3 de abril Institui o Sistema de Informação dos Certificados de óbito.

[3] Consejería de Sanidad de la Región de Murcia. (2016). En: www.murciasalud.es/certifica/ (consultado el 30/04/2016).

[4] . European Union. (2011). Reglamento (UE) Nº328/2011 de la Comisión de 5 de abril de 2011 por el que se aplica el Reglamento (CE) Nº1338/2008 del Parlamento Europeo y del Consejo sobre estadísticas comunitarias de salud pública y de salud y seguridad en el trabajo, por lo que se refiere a las estadísticas sobre las causas de la muerte. En: http://eur-lex.europa.eu/.

[5] Eurostat. (1998a). Coding of Causes of death in European Community. Project 96/S 99-57617/EN - Lot 11. Final report (June 1998) by Pavillon, G., Coleman, M., Johansson, L. A., Jougla, E. and Kardaun, J.

[6] Eurostat. (1998b). Comparability and Quality Improvement of European Causes of Death Statistics. EDC DGV/F3 SOC 98 20108. Final report.

[7] Eurostat. (2016). Causes of death statistics. En: http://ec.europa.eu/eurostat/web/health/causes-death (consultado el 30/04/2016).

[8] Harteloh, P., De Bruin, K. and Kardaun, J. (2010). The reability of causes of death coding in The Netherlands. *Eur. J. Epidemiol.*, **25**, 531–538.

[9] INE. (2016). Estadística de defunciones según la causa de muerte. En: www.ine.es (consultado el 30/04/2016).

[10] Iris Institute. (2016). En: www.dimdi.de/static/en/klassi/irisinstitute/index.htm (consultado el 30/04/2016).

[11] López Zazo, R. (2004). Estadística de mortalidad según causas múltiples de la Comunidad de Madrid. En: www.madrid.org/iestadis/fijas/estructu/demograficas/mnp/descarga/-mor04_1.pdf (consultado el 30/04/2016).

[12] Organización Mundial de la Salud. (2016a). CIE-10 (Décima Revisión de la Clasificación Estadística Internacional de Enfermedades y Problemas relacionados con la Salud). En: www.who.int/classifications/icd/en/ (consultado el 30/04/2016).

[13] Organización Mundial de la Salud. (2016b). The WHO-FIC Mortality Reference Group. En: www.who.int/classifications/committees/mrg/en/ (consultado el 30/04/2016).

[14] Pavillon, G. and Johansson, L. A. (2001). Production of methods and tools for improving causes of death statistics at codification level. *Eurostat working papers, Population and social conditions*, OS/E3/01/COD/11.

[15] US National Centre for Health Statistics. (2016). Mortality Medical Data System. En: www.cdc.gov/nchs/nvss/mmds.htm (consultado el 30/04/2016).

Acerca de los autores

Jesús Carrillo Prieto es licenciado en Medicina y Cirugía por la Universidad de Granada. Pertenece a la Escala Técnica de Gestión de Organismos Autónomos. Ha desarrollado su carrera profesional en el área de Estadísticas Sociales Sectoriales del INE, en concreto, en las estadísticas relacionadas con la salud. Actualmente es Jefe de Servicio.

Mª Rosario González García es licenciada en Ciencias Matemáticas por la Universidad de Salamanca. Pertenece al Cuerpo de Estadísticos Superiores del Estado y al Cuerpo de Profesores de Enseñanza Secundaria. Ha desarrollado su carrera profesional como profesora de educación secundaria hasta su incorporación en 2008 en la Subdirección General de Estadísticas Sociales del INE. Actualmente es Jefa de área de Estadísticas Sanitarias.

Boletín de Estadística e Investigación Operativa
Vol. 32, No. 2, Julio 2016, pp. 148-160

Historia y Enseñanza

Open Source Software for Mathematics and Statistics Teaching

Francisco Rabadán-Pérez and Carolina Cosculluela-Martínez

Departamento de Economía Aplicada I
Universidad Rey Juan Carlos

✉ pacorabadan@pacorabadan.com, ✉ carolina.cosculluela@urjc.es

Raquel Ibar-Alonso

Departamento de Matemática Aplicada y Estadística.
Universidad San Pablo CEU

✉ ribar@ceu.es

Abstract

Open Source software (OSS) and GNU (GNU's Not Unix) or free software, is practically unfamiliar for the Teaching Community. This paper provides a certain number of free tools of special utility for teaching mathematics and statistics. The knowledge of these programs by the academic community will simplify teaching at the same time that it will offer long-term application of the techniques that have been learned by the students. The increasing expansion of Open Source Software in the organizations, and especially in university environments and in the enterprises, will motivate the students to use and to specialize in them. For this purpose, a general criteria to comply with the suitable program for teaching is proposed: 1) accordingly to the potential of implantation with hardware, 2) with respect to the task of teaching; and, 3) regarding its practical use by the student. As conclusion, all software potentially valid selects those tools more suitable for the teaching of mathematics, with special emphasis in the statistics and tools associated with the R project.

Keywords: Free Software, Teaching, Mathematics, Statistics
AMS Subject classifications: 97U04

1. Introducción

Existe un desconocimiento generalizado en cuanto al posible uso de los programas gratuitos por parte de la comunidad docente. Su enseñanza en las aulas cada día va a ser más habitual así como su aplicación en el mundo empresarial.

En este artículo se pretende ofrecer una guía de los programas más adecuados para la enseñanza de las materias de Matemáticas y de Estadística. Con ello, pretende dar respuesta a las preguntas cuyo desconocimiento disuade a profesores y científicos al uso de este tipo de programas.

Se responden a las siguientes preguntas:

1. Primero, ¿es conveniente el uso de programas gratuitos para la enseñanza de las Matemáticas y la Estadística? ¿Qué se les pide para que sean útiles?

2. Segundo, ¿cuál es el software más potente y difundido para el análisis de datos en el mundo OSS y GNU?

3. Y, tercero, ¿qué programas están disponibles?, ¿para qué sirven? Y ¿cuáles son los mas adecuados para la docencia?

Lo que hace a este artículo ser de interés para la comunidad docente es que ofrece una alternativa al software privativo con programas gratuitos disponibles y adecuados para cada una de los análisis que se quieren realizar.

El resto del artículo se organiza de la siguiente manera: en la sección 2, se establecen los requerimientos que tiene que tener el software para ser de utilidad a la comunidad docente; en la sección 3 se presenta el entorno que reúne buena parte del software matemático y estadístico y se presentan los programas con algunas de sus aplicaciones a materias concretas; en la sección 4 se concluye.

2. ¿Qué debemos exigir al software que utilicemos en educación?

En este apartado se sintetizan los requisitos que sería deseable que tuvieran los programas orientados a la enseñanza y que luego se analizarán.

Así, las características que debe tener un programa se podrían clasificar en:

1. Potencial de implantación respecto al hardware: multiplataforma, clonación y virtualización.

2. Potencial de implantación respecto a la labor docente: nivel de orientación a la enseñanza, nivel de dificultad de la materia adecuado al nivel de educación, capacidad de reflejar el proceso de cálculo en relación a los resultados obtenidos y una curva de aprendizaje asequible.

3. Potencial de uso práctico por el alumno: facilidad de uso e interfaz gráfica, libertad de uso del software en el entorno doméstico, vigencia a largo plazo de los conocimientos adquiridos, expectativas de difusión del software en el futuro y que la presentación de resultados y dudas puedan compartirse en foros de debate en Internet.

2.1. Potencial de implantación respecto al hardware

Uno de los requisitos más importante para poder llevar el software a las aulas es que sea capaz de ejecutarse en los sistemas operativos más habituales, actualmente Windows, Mac OS, Linux, Android e iOS. Esto no lo cumplen muchos de los programas por los que se paga una licencia, aunque son cada vez son más las compañías que tienden a programar en Java, GCC, Python, Qt, lo que posibilita la multiplataforma. El software GNU y OSS tiene la ventaja de estar disponible para la practica totalidad de sistemas operativos precisamente por estar programado con este tipo de lenguajes.

Linux, a diferencia de otros sistemas, presenta la ventaja de tener una altísima compatibilidad con el hardware. Por ejemplo, podríamos realizar la instalación en un equipo con procesador AMD y trasladarla a un equipo Intel, con *chipsets* y periféricos absolutamente distintos, sin que la instalación se resienta. En los ordenadores habituales, basados en procesadores Intel, la única restricción es que si elegimos la distribución AMD64, no podremos instalarla en equipos con procesadores a 32bits.

No debemos olvidar iOS y Android, que si bien presentan la limitación de estar diseñados para equipos con una baja capacidad de cálculo, sí suelen ser equipos habituales para la mayoría de los alumnos. Ambos sistemas han influido de forma determinante en el desarrollo de las NTICs (Nuevas Tecnología de Información y Comunicación) que se han ido incorporando en los últimos años cambiando la forma tradicional de enseñar y aprender, tanto por el soporte cada vez más digitalizado como por la creciente incorporación de la enseñanza online (Ruiz *et al.*, 2014). Un ejemplo de aplicación para la enseñanza/aprendizaje de la Estadística mediante tablets es la aplicación APPES.

Uno de los proyectos de código abierto más interesantes es VirtualBox (Virtualbox.org, 2016) de la compañía Oracle, ya que a través de la virtualización de sistemas y un ordenador con potencia media, es posible ejecutar simultáneamente más de un sistema operativo. Esto nos permitiría impartir clase en una máquina virtual, como por ejemplo Mathbuntu (Brin, 2016) que nuestros alumnos podrán usar desde su propio ordenador, o en un servidor remoto al que podrán acceder desde un cliente VNC (*Virtual Networking Computing*) o RDP (*Remote Desktop Protocol*).

Linux ofrece la ventaja de la clonación. Podemos montar la máquina virtual según las necesidades docentes y luego montarla como máquina virtual e incluso clonarla en los equipos de nuestros alumnos. Con VirtualBox podemos tomar

snap-shots (instantáneas), lo que nos permite poder volver al instante en el que hemos sacado la *snap-shots*. Además, podemos sacar una copia de seguridad del entorno de trabajo completo copiando la máquina virtual que se reduce a un conjunto muy pequeño de archivos que, fundamentalmente, son los discos duros y las características del hardware virtual.

Por tanto, el OSS cumple los tres requisitos que exigíamos en el primer apartado de la clasificación: multiplataforma, clonación y virtualización.

2.2. Potencial de implantación respecto a la labor docente

La incorporación de nuevas tecnologías a la metodología docente hace que nos cuestionemos la forma de impartir clase.

Los educadores de metodología cuantitativa son cada vez más dependientes del software especializado para transmitir sus conocimientos lo que plantea ciertos problemas de carácter metodológico: cómo pasar del razonamiento matemático explicado de forma tradicional a la salida de ordenador, y cómo explicar que la salida de ordenador tiene un razonamiento matemático sin el cuál no es posible interpretar correctamente los resultados. Cuando el software utilizado para impartir clase está sujeto a una licencia privativa, normalmente, el estudiante no puede seguir usándolo fuera del centro educativo lo que limita su capacidad de aprendizaje (Culebro *et al.*, 2006).

Hay multitud de trabajos científicos acerca de los beneficios de utilizar software en los cursos en que se enseñan conceptos matemáticos (Ávila *et al.*, 2007), aunque respecto a la comunicación parece que los alumnos siguen prefiriendo la tiza a las transparencias y la pizarra digital (Cosculluela-Martínez *et al.*, 2015). Para realizar los cálculos, sin embargo, no podemos prescindir del software ya que a medida que las asignaturas van ganando en complejidad sería materialmente imposible abordar la resolución de problemas a la manera tradicional.

En los últimos años venimos observando como la sociedad demanda una educación eminentemente práctica para que los conocimientos adquiridos faciliten la incorporación del alumnado al mercado de trabajo. El diseño de las aplicaciones de análisis de datos parece condicionarnos a hacer énfasis en la comprensión de la Matemáticas y en la capacidad de interpretar los resultados, más que en el método de cálculo y la búsqueda de la exactitud numérica. Sin embargo, el principal objetivo en la enseñanza de la metodología cuantitativa ha de ser comprender el procedimiento. Teniendo en cuenta que el tiempo destinado a impartir la asignatura es limitado, el tiempo destinado a enseñar el manejo del software no puede ir en detrimento de la enseñanza de la metodología.

El diseño de software de análisis de datos tiene dos objetivos. El primero respecto al investigador especializado: dotarle de la últimas técnicas para mejorar la calidad de los análisis. El segundo, respecto al usuario medio, que es el que constituye una mayor demanda potencial, conseguir una interfaz gráfica que permita una alta capacidad para interactuar e interpretar los resultados. Algunas

de las aplicaciones comerciales más conocidas han logrado su popularidad gracias a esta última característica.

Son grandes los esfuerzos de las comunidades que respaldan los proyectos GNU y OSS para facilitar el acceso a los que se acercan por primera vez a este tipo de software. Este acercamiento se fundamenta en dos pilares:

1. En la abundante documentación oficial de los propios proyectos y los cada día más habituales *howto's* que son textos orientados a resolver situaciones concretas desarrollados por la comunidad internauta relacionada con el proyecto (desarrolladores y usuarios).

2. 2. En el desarrollo de entornos gráficos que buscan evitar al máximo la tediosa y delicada labor de introducir comandos en el terminal y que habitualmente desemboca en la programación.

   Por tanto, el OSS nace con vocación de ser comprendido y usado por el mayor número de usuarios posibles, lo que cumple el requisito de estar orientado a la enseñanza. Sin embargo, el educador debe ser capaz de discernir cúal es la herramienta que precisa en consonancia con la dificultad de la materia que imparte, con la capacidad de reflejar el proceso de cálculo en relación a los resultados obtenidos y conforme a una curva de aprendizaje adecuada al tiempo del que se dispone para impartir clase.

2.3. Potencial de uso práctico por el alumno

El alumno cuando se enfrenta a un problema matemático en un entorno informático lo hace a través de una pantalla, lo que supone: primero, saber manejar el sistema operativo; segundo, saber manejar el programa; y, tercero, entender qué significan los resultados que obtiene. El profesor tiene que saber en muchas ocasiones explicar los tres procesos de conocimiento, y hacer hincapié, precisamente en el último, pues el alumno puede caer en el error de pensar que si sabe obtener las salidas ha resuelto el problema que se le plantea. En asignaturas de análisis de datos esta situación es habitual.

El Software Libre (GNU) y de Código Abierto (OSS) tiene la gran ventaja de poder usarse dónde, cuándo y durante el tiempo que sea necesario a coste cero, con la garantía de que lo aprendido podrá seguir utilizándose en el futuro. Esto lo convierte en un vehículo que promueve el libre pensamiento, la innovación y se hace indispensable para el desarrollo de naciones como la India (Kalyani *et al.*, 2011). El OSS (Open Source Software) parece estar extendiéndose también en la enseñanza superior de las universidades americanas (Williams van Rooij, 2011). La normativa española, actualmente en vigor, establece como principio para la educación la garantía de acceso a las TIC (Tecnologías de la Información y Comunicación) para propiciar la equidad, la Justicia Social y evitar las brechas que pueden afectar a la integración y la cohesión social (García-Valcarcel *et al.*, 2014).

La desventaja del OSS, sobre todo a mayor dificultad de la materia, es que buena parte de los programas son lenguajes de programación vía línea de comandos, lo que dificulta el primer acercamiento al software. Sin embargo, hay proyectos sólidos que están desarrollando GUI's (*Graphical User Interface*) para facilitar los cálculos más habituales, y permitir romper la puerta fría que supone acercarse a software como R.

Otra gran ventaja del OSS es la comunidad científica global que está detrás, y cuyas innovaciones suelen estar disponibles de forma gratuita y prácticamente inmediata. Esto garantiza a efectos prácticos, en el largo plazo, que los usuarios individuales podrán usar software libre y Open Source para la mayor parte de sus investigaciones.

Por lo tanto, de los requisitos que se exigían en relación a la labor docente, éste cumple: libertad de uso del software en el entorno doméstico, vigencia a largo plazo de los conocimientos adquiridos, expectativas de difusión del software en el futuro y que la presentación de resultados y dudas puedan compartirse en foros de debate en Internet (Ridgway *et al.*, 2006). Queda a cargo del docente seleccionar la interfaz precisa para alcanzar la facilidad de uso .

Por lo tanto, una vez visto que este software cumple los requisitos fundamentales para ser usado en el ámbito docente, se presentan algunos de los programas más interesantes para la enseñanza de la Matemáticas y de la Estadística.

3. Software GNU y Open Source para la docencia de la Matemática y la Estadística

Mathbuntu recopila el software más extendido y utilizado por la Comunidad Científica, además de un conjunto de programas y libros de texto para facilitar el aprendizaje. Por esta razón se describe a continuación.

3.1. Mathbuntu: prácticamente todo el OSS y GNU para la investigación científica.

En torno al proyecto Ubuntu, basado en Debian, han surgido multitud de proyectos como Kubuntu, Lubuntu, Xubuntu, Linux Mint, y entre ellos, respecto de nuestros objetivos, resalta Mathbuntu diseñado por el Dr. Len Brin, profesor de Matemáticas de la Southern Connecticut State University, que se distribuye en dos formatos: distribución completa y script.

El script es un programa de procesamiento por lotes que descarga los paquetes necesarios, el último código fuente disponible y procede a la compilación e instalación de todo el software matemático y estadístico incluido en Mathbuntu. También incluye un conjunto de libros de texto de código abierto.

La distribución completa permite ahorrar tiempo en la instalación y es mas amigable (fácil de instalar) para los usuarios no habituados a Linux, pero presenta el inconveniente de que si queremos tener actualizado el software deberemos

volver a ejecutar el script. El tiempo de instalación vía script puede durar 3 horas o más debido al proceso de descarga y compilación. Pero si el objetivo es la replicación y la redistribución entre nuestros alumnos, esto no supone un problema demasiado grave, ya que tras la primera instalación podemos clonar la máquina real (Brin, 2016)o la máquina virtual.

El software más relevante incluido en Mathbuntu es Sage (SageMath Mathematical Software System, 2016), software matemático; Máxima (Sourceforge.net, 2016), sistema de computación algebraica; R (R-project.org, 2016), computación estadística; en computación numérica GNU Octave (Gnu.org, 2016) y Scilab (Enterprises Scilab, 2016); Geogebra (Geogebra.org, 2016), algebra y geometría interactiva; LaTex (Latex-project.org, 2016), sistema de preparación de documentos; Lurch (Lurchmath.org, 2016), procesador de textos que chequea la sintaxis Matemáticas; Netlogo (Wilensky et al., 2016), modelización de sistemas dependientes del tiempo. Estos programas existen de forma nativa o portada en Linux, FreeBSD, Mac OS y Android, a excepción de Netlogo y Geogebra respecto a FreeBSD y Forks.

A continuación abordaremos el problema de cómo mejorar la capacidad de interactuar con algunos de estos programas y hacerlos más accesibles para el alumno y el proceso de enseñanza.

3.2. Facilitando la capacidad de interacción con el usuario

El OSS y GNU más potente suele estar basado en lenguajes de computación e instrucciones vía terminal lo que hace que la curva de aprendizaje sea difícil al principio para el estudiante. Los GUIs (*Graphical User Interface*) y los IDEs (*Integrated Development Environment*), solucionan parcialmente este problema. En la mayor parte de los casos no se puede prescindir del terminal, pero muchas tareas se pueden ejecutar desde la interfaz visual.

Respecto a LaTex, tenemos dos editores libres multiplataforma: Lyx y Texmaker, ninguno de ellos existe para Android, donde contamos con VerbTex. Aunque no son multiplataforma, para Mac OS tenemos MacTex y en Windows WinEDT.

Para R hay tres IDE‘s fundamentales: R commander,disponible desde el propio R como un modulo; RKWard (Rkward.kde.org, 2016) dependiente del proyecto KDE, pero disponible para Linux, OSX y Windows; y Rstudio (RStudio, 2016) también multiplataforma. El interface mas recomendable por las opciones en menú es RkWard.

Respecto a Maxima podemos encontrar wxMaxima, Maxima for Android, Imaxima, Imath, Kayali y Symaxx2. Los más intuitivos son wxMaxima y el par formado por iMaxiima e Imath.

Scilab, GNU Octave Geogebra y Netlogo tiene su propio Frontend.

A continuación se exponen las propuestas que, desde nuestro modo de ver, son las más indicadas para la impartición de la docencia.

4. Tres propuestas GNU para impartir clase de Estadística

Los programas que se detallan a continuación resaltan respecto al resto de software libre por la alta capacidad de interacción con el usuario que se materializa en una interfaz gráfica que reduce drásticamente la curva de aprendizaje. Estos son: Gnumeric, PSPP y RKward.

4.1. Gnumeric

Es la hoja de cálculo del proyecto Gnome (The Gnome Project., 2016), el entorno gráfico del proyecto GNU. ésta puede ser una plataforma ideal para cursos de Estadística e Introducción a la Econometría. Incluye un menú denominado estadísticas que entre otras cosas nos permite: cálculo de los estadísticos descriptivos más habituales, cálculo del intervalo de confianza para la media, métodos de muestreo aleatorio y sistemático, contraste de hipótesis paramétricos (media y mediana) y no paramétricos (bondad del ajuste), ANOVA, tablas de contingencia, regresión lineal múltiple y análisis de componentes principales. Además, con el módulo de series temporales realiza técnicas de suavizado exponencial y desplazamiento promediado.

4.2. PSPP

PSPP (Gnu.org, 2016) aspira a ser una alternativa gratuita al software de IBM SPSS. Trabaja con el mismo formato de archivo y en ocasiones es capaz de importar ficheros de otras plataformas que el propio SPSS no importa. Cuenta con un menú para la recodificación de variables al igual que SPSS. El interfaz de PSPP sigue la misma estructura que SPSS, sin embargo, es mucho menos completo. Realiza prácticamente las mismas funciones que Gnumeric, a excepción del suavizado exponencial y el desplazamiento promediado, pero además incorpora técnicas de análisis multivariante como análisis factorial y conglomerados de k medias (no incluye actualmente el clúster jerárquico). Es especialmente interesante por la variedad de contrastes no paramétricos que incorpora.

4.3. RKWard

El menú de RKWard es muy completo y está claramente orientado a análisis estadísticos avanzados. Cabe destacar que ofrece técnicas para detección de *outliers*, análisis de potencia del contraste, y contrastes para análisis de series temporales: Box-Pierce, Ljung-Box, KPSS de estacionariedad, y Phillis-Perron Es especialmente útil para la enseñanza de distribuciones de probabilidad. En el menú *distributions*, además de distintos tipos de contraste de normalidad, muestra distribuciones continuas y discretas que se pueden representar gráficamente y ser modificadas por el profesor en clase para mostrar, por ejemplo, el fenómeno de la convergencia y el teorema central del límite. RkTeaching (Sánchez Alberca, 2016) es un módulo en desarrollo por la Universidad San Pablo CEU,

que se integra en el menú de RkWard y está específicamente orientado a la tarea pedagógica (Sánchez Alberca, 2015).

Para adentrarnos aún más en el propio R, con independencia de RKWard, debemos tener en cuenta también:

- statsTeachR (G Reich and S Foulkes, 2016) es un repositorio de acceso abierto con lecciones modulares para enseñar Estadística usando R. Según indica el propio sitio, hecho por profesores para profesores.

- Quick-R (Kabacoff, 2016) es una página web donde disponemos de una amplia variedad de *howtos* muy esquemáticos sobre como abordar técnicas estadísticas de muy distintos niveles.

- Podemos encontrar video-tutoriales, cursos y recursos en GoUmh (Gentleman et al., 2016) de la Universidad de Granada.

- Para cursos específicos de R podemos consultar R Comunidad Hispano (R-es.org, 2016).

5. Conclusiones

Dando respuesta a las preguntas planteadas, y en relación a la primera, sí resulta conveniente el uso del software GNU y Open Source por varias razones: 1ž porque hay expectativas de una profunda difusión y de la mejora de la calidad de la interfaz de este software en el futuro; y, 2ž porque beneficia la justicia social permitiendo que cualquier usuario acceda a las NTIC's. Este software cumple los requisitos necesarios para ser usado en la docencia.

Respondiendo a la segunda de las preguntas planteadas, el mejor modo para acceder a una plataforma operativa completa basada en GNU y OSS para la investigación y la enseñanza de la Matemáticas parece ser Mathbuntu, aunque se pueda disponer de las aplicaciones de forma independiente.

En relación a los programas disponibles, se destacan tres propuestas en el ámbito de la Estadística debido a la alta capacidad de interacción con el usuario: Gnumeric, PSPP, y RKWard.

Se añaden enlaces para disponer de información completa y actualizada de los proyectos y diversos sitios web con howto's para realizar actividades docentes más específicas.

El software citado en este artículo permite a profesores, alumnos e investigadores realizar cálculos estadísticos sin el coste de una licencia privativa, profundizar en el conocimiento matemático y habituarse a un software que cada día será más común en todos los ámbitos.

Referencias

[1] Ávila, M. C., Chourio, E. D., Carniel, L. C. y Álvarez-Vargas, Z. (2007). El software matemático como herramienta para el desarrollo de habilidades del pensamiento y mejoramiento del aprendizaje de las Matemáticas. *Actualidades Investigativas en Educación* , **7(2)**, 1–34.

[2] Batanero, C. y Díaz, C. (Eds.) (2011). *Estadística con Proyectos*. Departamento de Didáctica de la Matemática. Facultad de Ciencias de la Educación. Universidad de Granada (España), 14–16.

[3] Brin, L. (2016). Mathbuntu | Instant access to free mathematical software. [online] Mathbuntu.org. En: http://mathbuntu.org [Consultado el 4 de Abril de 2016].

[4] Cosculluela-Martínez C. e Ibar, R. (2015) ¿Demanda el alumno la enseñanza online? *VI Jornadas en Innovación y TIC Educativas*. JITICE 2015. Madrid (España).

[5] Culebro Juárez, M., Gómez Herrera, W. y Torres Sánchez, S. (2006). Software libre vs software propietario: Ventajas y desventajas. En: http://www.rebelion.org/docs/32693.pdf.[Consultado el 22 de Marzo de 2016].

[6] G-Reich, N. y S-Foulkes, A. (2016). statsTeachR. [online] Statsteachr.org. En: http://statsteachr.org [Consultado el 4 de Abril de 2016].

[7] García-Valcárcel, A., Basilotta, V. y López, C. (2014). ICT in collaborative learning in the classrooms of primary and secondary Education *Comunicar* **21 (42)**, 359.

[8] Geogebra.org. (2016). GeoGebra. [online] En: http://www.geogebra.org/ [Consultado el 4 de Abril de 2016].

[9] Gentleman, R. y Ihaka, R. (2016). Videotutoriales de R y R Commander. [online] Sites.google.com. En: https://sites.google.com/a/goumh.umh.es/[Consultado el 4 de Abril de 2016].

[10] Kabacoff, R. (2016). Quick-R: Home Page. [online] Statmethods.net. En: http://www.statmethods.net/index.html [Consultado el 4 de Abril de 2016].

[11] The Gnome Project. (2016). The Gnumeric Spreadsheet: Free, Fast, Accurate. [online] Gnumeric. En: http://www.gnumeric.org [Consultado el 4 de Abril de 2016].

[12] Gnu.org. (2016). GNU Octave. [online] En: https://www.gnu.org/software/octave/ [Consultado el 4 de Abril de 2016].

[13] Gnu.org. (2016). PSPP - GNU Project - Free Software Foundation. [online] En: https://www.gnu.org/software/pspp/ [Consultado el 4 de Abril de 2016].

[14] Latex-project.org. (2016). LaTeX - A document preparation system. [online] En: https://www.LaTex-project.org/ [Consultado el 4 de Abril de 2016].

[15] Lurchmath.org. (2016). Lurch | The word processor that can check your math. [online] En: http://lurchmath.org/ [Consultado el 4 de Abril de 2016].

[16] Kalyani, P. y Kotwani, G. (2011). Open source software (OSS): Realistic implementation of OSS in school education. *Trends in Information Management* **7 (2)**, 208–17.

[17] Maxima.sourceforge.net. (2016). Maxima, a Computer Algebra System. [online] En: http://maxima.sourceforge.net/ [Consultado el 4 de Abril de 2016].

[18] R-es.org. (2016). R Comunidad Hispano. [online] En: http://r-es.org/category/formacion/ [Consultado el 4 de Abril de 2016].

[19] R-project.org. (2016). R: The R Project for Statistical Computing. [online] En: https://www.r-project.org/ [Consultado el 4 de Abril de 2016].

[20] Ridgway, J., Nicholson, J. y McCusker, S. (2007). Teaching statistics–despite its applications. *Teaching Statistics: An International Journal for Teachers* **29 (2)**, 44–48.

[21] Rkward.kde.org. (2016). Welcome to RKWard. [online] En: https://rkward.kde.org/ [Consultado el 4 de Abril de 2016].

[22] RStudio. (2016). Home. [online] En: https://www.rstudio.com/ [Consultado el 4 de Abril de 2016].

[23] Ruiz-Castro, J. E., Aguilera, A. M., Escabias, M. y Raya-Miranda, R. (2014) Can we learn statistics through a Tablet? Yes, we can APPES. *BEIO*. **30 (2)**, 181–198.

[24] SageMath Mathematical Software System. (2016). SageMath Mathematical Software System - Sage. [online] En: http://sagemath.org [Consultado el 4 de Abril de 2016].

[25] Sánchez-Alberca, A. (2015). Bringing R to non-expert users with the package RKTeaching. *BEIO*. **31 (2)**, 170–188.

[26] Sánchez-Alberca, A. (2016). Un paquete de R para la enseñanza de Estadística. [online] Aprende con Alf. En: http://aprendeconalf.es/rkteaching [Consultado el 4 de Abril de 2016].

[27] Scilab Enterprises. (2016). Home - Scilab. [online] Scilab.org. En: http://www.scilab.org/ [Consultado el 4 de Abril de 2016].

[28] Team, D. (2016). Clonezilla - About. [online] Clonezilla.org. En: http://clonezilla.org/ [Consultado el 4 de Abril de 2016].

[29] Virtualbox.org. (2016). Oracle VM VirtualBox. [online] En: https://www.virtualbox.org/ [Consultado el 4 de Abril de 2016].

[30] Wilensky, U. y Stroup, W. (2016). NetLogo. [online] Center for Connected Learning and Computer-Based Modeling, Northwestern University. Evanston, IL. En: http://ccl.northwestern.edu/netlogo/ [Consultado el 4 de Abril de 2016].

[31] Williams van Rooij, S. (2011). Higher education sub-cultures and open source adoption. *Computers & Education* **57**, 1171–1183.

Acerca de los autores

Francisco Rabadán Pérez es doctor con la distinción *cum laude* en el programa de Análisis Económico y Economía Aplicada (2015) por la Universidad San Pablo CEU y licenciado en Administración y Dirección de Empresas. Es director del área de transporte en la empresa RAC, S.L.. Es profesor asociado en el Departamento de Economía Aplicada I de la Universidad Rey Juan Carlos en la que imparte diversas asignaturas de Estadística. Es el organizador de las Jornadas Internacionales sobre Paradigma Económico Emergente que se han desarrollado en el campus de Aranjuez de la Universidad Rey Juan Carlos. Investiga en el diseño de nuevas metodologías cuantitativas aplicadas al *Big Data* y a las *Smart Cities*.

Carolina Cosculluela-Martínez fue Premio Extraordinario de Doctorado y Premio Funcas a la mejor Tesis Doctoral. Defendió la Tesis en el Departamento de Economía Aplicada y Estadística de la UNED y obtuvo una ayuda de la Fundación Ramón Areces para continuar una de las líneas de investigación propuesta en la misma. Profesora en el Departamento de Economía Aplicada I de la URJC, con una estancia de investigación en *Regional Economic Applied Laboratory* (U. de Illinois, Chicago). Cuenta con más de 15 publicaciones, 2 de ellas JCR, y más de 10 participaciones como investigadora en proyectos de las Consejerías de Empleo e Inmigración. Desarrolla la investigación en materia de *Smart Cities* complementando la desarrollada por la U. de Illinois a la que ha sido invitada recientemente como profesor visitante.

Raquel Ibar-Alonso es doctora en Ciencias Económicas y Empresariales y licenciada en Ciencias Matemáticas. Profesora en el Departamento Interfacultativo de Matemática Aplicada y Estadística de la Universidad San Pablo CEU de Madrid. Miembro del Grupo de Investigación en Convergencia de Medios (INCIRTV) y del proyecto precompetitivo *Smart Cities*: Problemas de accesibilidad a los contenidos digitales en ciudadanos de edad avanzada. Sus líneas de investigación que mantienen un carácter multidisciplinar, se centran en el Análisis Estadístico Multivariante, la *Smart City*, el comportamiento social y en la recogida de información, tanto cualitativa como cuantitativa..

Boletín de Estadística e Investigación Operativa
Vol. 32, No. 2, Julio 2016, pp. 161-170

Opiniones sobre la profesión

Is scientific divulgation mandatory? A little piece of this

Jesús López Fidalgo

Departamento de Matemáticas
Universidad de Castilla–La Mancha
✉ Jesus.LopezFidalgo@uclm.es

Abstract

Whether scientific divulgation is mandatory for a scientist is discussed in this article. This question is particularly addressed to Statistics. The way of doing this is analyzed through the opinion and perspective of the author. Then the book "The hazard does not exist" ("El azar no existe") is presented. Some other books popularizing Statistics are also reviewed. Finally a "false", but illustrative, example of the book is provided.

Keywords: Hazard, Popularizing Statistics
AMS Subject classifications: 62-00, 97K70

1. Divulga que algo queda

El mundo científico se plantea cada vez más la necesidad de divulgar lo que se hace en el campo de la investigación, así como los instrumentos que se utilizan. Pienso que hay un cierto consenso general en que en esto hemos fallado. Hace unos años oí a un investigador básico una frase que me dejó perplejo: "A mi no me importa en absoluto que lo que hago se aplique o no". Es verdad que la frase está sacada de contexto, pero hoy da seguramente casi nadie se atrevería a decir algo así. Al menos algo ha cambiado. Me imagino que divulgar no sea muy rentable para conseguir puntos–ANECA, pero estoy convencido de que estamos obligados a hacerlo. Si no, no podemos quejarnos después de que la financiación para la investigación sea demasiado baja. Nos quejaremos de todas formas... Aunque sea con no poca vergüenza por mi parte, en este artículo me gustaría presentar un librito de divulgación que he escrito recientemente y que he titulado "El azar no existe". Son algunas ideas básicas contadas de modo asequible para cualquiera, siguiendo aquellas palabras atribuidas a Einstein: "No entiendes realmente algo a menos que seas capaz de explicárselo a tu abuela". Aunque hay abuelas con una formación estadística excelente, yendo a la idea de fondo de la frase, eso es

exactamente lo que he tratado de hacer. Ha sido divertido y me lo he tomado como un momento de descanso los domingos por la tarde. Ciertamente el material venía recopilado de diversas charlas y clases impartidas, solo era necesario hilarlo. Antes de dar una visin general del libro, quiero ahondar más en algunas ideas y opiniones acerca de la divulgación, en particular de las matemáticas y más en concreto de la Estadística.

El año internacional de la Estadística fue una ocasión espléndida en todo el mundo para divulgar y dar a conocer en ámbitos muy diversos lo que la estadística y la investigación operativa hacen. En el artículo publicado por BEIO en el volumen (2014) se daba buena cuenta de las actividades que se realizaron durante el año 2013 en España. Se resaltaba también que desde hace años se viene desarrollando un buen número de actividades para promocionar los estudios de Estadística entre los más jóvenes. En el número especial publicado en BEIO (2013) con esta ocasión se incluyeron ocho artículos cuyo objetivo era poner de manifiesto el papel vital de la Estadística y la Investigación Operativa, así como de sus profesionales. Y esto, no solo en el mundo de la educación, la investigación y la empresa, sino en todos los ámbitos de la vida, con los beneficios que esto conlleva para la sociedad en general. Los ocho artículos de ese número especial son idóneos para divulgar el papel de la Estadística más allá del ámbito académico y profesional.

Desde luego hay formas muy diversas de divulgar. Todas aportan algo y por eso cada uno, aprovechando sus "facultades de tiempo libre", puede hacer mucho en este terreno. Nos encontramos así que el que tiene afición a la magia, puede orientar la divulgación a ese terreno. Hay muy buenos ejemplos de ello en el campo más amplio de las matemáticas. Podríamos seguir mencionando a los que tienen dotes para hacer un monólogo divertido, que inyecta divulgación directamente en vena y casi sin enterarse. Aquellos con capacidad de montar espectáculos y atraer la atención de diversos públicos son especialmente valiosos. No soy muy amigo de los "juegos matemáticos" por varios motivos. Por una parte, porque la ciencia, dígase matemáticas o estadística, no es un juego. Otra razón, más importante, es que se puede estar lanzando el mensaje de que las matemáticas, dígase estadística, son un juego que no sirve para otra cosa que para el entretenimiento, eso sí muy sofisticado. Podría incluso pensarse que necesitamos inventarnos estas cosas y mostrarlo así, para que no reduzcan la carga matemática de primaria, secundaria o bachillerato. Otra razón se refiere a esos problemas con idea feliz o incluso con truco. El mensaje que estamos lanzado ahora es que esto es solo para unos pocos privilegiados capaces de tener "ideas felices". En este sentido me quedo con la frase de Picaso "la inspiración siempre me ha pillado trabajando" o con esa otra de George Box "parece que cuanto más trabajo más suerte tengo". No obstante, no discutiré que esta vía sirve, y mucho, especialmente con los más jóvenes. Mucho valor tienen los que encuentran facilidad para entrar en los medios de comunicación. Tú que lees este artículo,

piensa ¿qué se me da bien a mi? y seguramente habrá alguna forma de divulgar aprovechando esa afición o talento particular. ¡Anímate!

2. Necesitamos estadísticos

El auge del llamado Big Data y de los científicos de datos ha puesto de manifiesto la necesidad de un gran número de personas con la formación adecuada para su tratamiento. Suele destacarse que esto no es un problema para una sola área. La interdisciplinariedad es imprescindible para llevar a cabo el proyecto de sacar rentabilidad a las grandes cantidades de información que se nos presentan de un modo más o menos velado. Y en ese equipo interdisciplinar no puede faltar el estadístico. Consecuentemente hay una cierta responsabilidad de formar estadísticos y, por tanto, de atraerlos hacia esta formación. Y en esto hay que convencer a los futuros universitarios, a sus padres, a sus profesores y orientadores y a la sociedad entera, que es de donde salen los artífices de las políticas que determinan el porvenir. La sociedad necesita estadísticos.

En muchos casos, a pesar de la buena orientación y organización, algunas actividades promocionales de la estadística no han tenido la repercusión esperada, por ejemplo en un aumento de alumnos en las titulaciones de Estadística. Con frecuencia los criterios que utiliza un alumno para elegir carrera suelen estar basados en lo que conoce, o cree conocer, y en lo que le gusta, o piensa que le gusta. Por supuesto mira hacia el futuro profesional, no solamente en términos de rentabilidad económica futura sino también del trabajo en sí mismo que realiza un estadístico. Ahí desde luego reside quizá la mayor ignorancia y lo que hace que esta titulación sea poco atractiva, por ser poco conocida. En esto los profesores de secundaria y los orientadores tienen una tarea esencial, pero probablemente necesiten ayuda.

No tengo registrada una lista exhaustiva de las reacciones y respuestas de personas a las que te presentas como estadístico. Desde la consiguiente pregunta de "¿estudiaste la carrera de económicas?" hasta una declaración, sin pudor alguno e incluso con satisfacción, de lo mucho que le costó sacar la estadística en su carrera y que no llegó a entenderla nunca. Otros, especialmente los que se dedican a la investigación o tienen una mente más abierta en el mundo empresarial, manifiestan con humildad que deberían y les gustaría saber más estadística. No es infrecuente que la asignatura de estadística en muchas titulaciones, también en secundaria y bachillerato, sea impartida por expertos en otras materias. No tengo nada en contra de que alguien que domine la estadística la imparta en el ambiente que sea, independientemente de la formación universitaria que haya recibido inicialmente. Es más, conozco casos verdaderamente notables, y de los que he aprendido mucho, de no estadísticos dando muy buenas clases de estadística. Pero con frecuencia falta ese dominio, fundamental en la docencia. Y lo que es casi peor, falta la capacidad de transmitir un cierto entusiasmo por la

materia. Falta también la posibilidad de mostrar con claridad su aplicación más inmediata. De hecho, ahí es donde, bajo mi punto de vista, deberíamos centrar nuestros esfuerzos, en mostrar cómo se aplica en casos reales. Después ya podemos revestirla de un formato divertido o atrayente, pero lo primero es mostrar esa aplicabilidad, que no es tan difícil. Es preferible esto que lo contrario, es decir, buscar algo divertido o curioso y luego revestirlo de apariencia de realidad. Esto no funciona.

3. Pero, he venido a hablar de mi libro

Esta frase ha quedado como una referencia nacional, que todo el mundo utiliza, pero de la que quizá no todo el mundo conoce el origen. Aprovechando el desparpajo mostrado por este conocido escritor, me permito adentrarme en la difícil tarea de presentar mi propio libro "El azar no existe". Ha sido publicado recientemente por la editorial Electolibris, coeditado con la Real Sociedad Matemática Española (RSME). Esta editorial es una spin-off de la Universidad de Murcia en la que ha intervenido un grupo de matemáticos. Su finalidad es editar libros de texto y otras obras matemáticas asegurando una alta calidad.

Es un libro cuyo prólogo tiene forma de prospecto. Esto no supone una focalización del libro única y exclusivamente en las ciencias de la vida. En todo momento se busca atraer la atención y la curiosidad de posibles lectores. Por eso se hace uso de ese formato de prospecto, que todos estamos muy habituados a leer. Por eso la portada imita el formato de una caja de medicinas. Los títulos intentan ser atractivos y provocadores, como por ejemplo "Clones humanos", "Torturar los datos", "¿Qué diferencia el rostro de una mujer del de un hombre?", "Cacicracia: Votar no garantiza la democracia", "Que elija el azar a nuestros gobernantes", "El dato es bello"...

Mediante ejemplos de la vida cotidiana, echando mano del sentido común y nada más, se busca acercar la estadística a no especialistas. Está dirigido a todo tipo de lectores, sin excepción alguna. Se quiere poner en valor que saber algo de estadística ayuda a no dejarse engañar por posibles manipulaciones de diversa procedencia. Por otro lado se muestra como una herramienta muy potente del método científico, que permite conseguir resultados en todos los campos de un modo rápido y eficaz. No es un libro "gordo", en ningún sentido. Es el resultado de plasmar lo que ha venido a la cabeza de un modo natural después de años de experiencia.

Pienso que cumple así las características esenciales de un libro de divulgación. No es un manual para aprender estadística. No es necesario tener ningún conocimiento matemático previo, ni tener afición por las Ciencias. Se van describiendo a lo largo del libro posibles vías de manipulación antes, durante y después del análisis estadístico de los datos. Trata de fomentar el espíritu crítico ante las distintas fuentes de manipulación en la sociedad en la que vivimos. Pretende

mostrar esta ciencia de modo simpático, de forma que pueda leerse en cualquier sitio. La finalidad es mostrar una estadística amable y eliminar el concepto erróneo que muchos tienen de esta ciencia. No tiene fórmulas. Busca despertar el interés general por la estadística y que el lector entienda qué es lo que realmente hace sin entrar en detalles y de una manera muy natural. Muchos profesores universitarios de estadística podrían recomendarlo a sus alumnos o utilizarlo en clase. También los profesores de secundaria de matemáticas podrían utilizarlo a modo de libro que se recomienda leer a los alumnos, como ocurre en las asignaturas de literatura, historia o filosofía.

El libro intenta no caer en la inclusión de curiosidades o "acertijos" que con frecuencia se alejan de la realidad. Las curiosidades y ejemplos tratan de ser muy cotidianos y sencillos. Los estadísticos y matemáticos que lo han leído querrían que diera más detalles técnicos, pero pienso que ahí reside precisamente la eficacia del libro, en no dar detalles que no pueda entender un profano en la materia. Está todo explicado con palabras cotidianas, sin apenas tecnicismos. En algunos momentos pretende ser desafiante y provocador, rompiendo moldes (véase especialmente el capítulo dedicado a la democracia).

La mayoría de los gráficos y fotografías han sido hechos por el autor. Algunos de los personajes son amigos. Fue revisado por un número importante de estadísticos y también por algunas personas muy alejadas del mundo de la estadística o con poca formación.

El origen del título está en unas charlas que he venido impartiendo desde hace años, especialmente desde 2013, y que venía titulando "Sano espíritu crítico a través de la estadística". En una ocasión la coordinadora del ciclo de conferencias en el que se insertaba la charla me prohibió utilizar la palabra estadística, que podría ahuyentar a potenciales asistentes. Así de entrada no me sentó muy bien, porque uno tiene su orgullo de ser estadístico y por eso sin pensarlo muy bien y con un poco de resentimiento le dije que pusiera "El azar no existe". Efectivamente el título era ms atractivo y me llamó una radio para que explicara si por fin había demostrado que el azar no existe.

La portada muestra dos frases, una relativa a la estructura, "Tratamiento para el manipulador patológico" y que hace referencia al formato de caja de medicinas. La otra es provocadora, "Apto solamente para gente que piensa". Es fruto de escribir en una hoja muchas frases que comenzasen por "Apto solamente para ..." hasta que finalmente se impuso esta.

4. Otros libros

Me gustaría recomendar otras lecturas, unas más cercanas a este libro y otras de muy diversa índole. Quizá el que tiene un estilo más semejante es el famoso libro de Darrell Huff de 1954, "Cómo Mentir con Estadísticas". Es un buen libro, muy conocido. La versión original es más sugerente que algunas traducciones.

Muchos de los ejemplos son del mundo americano. A pesar de ser un libro muy antiguo conserva una frescura que lo hace todavía actual. El libro de Tanur y Mosteller de 1989, "La estadística una guía de lo desconocido" es un compendio de 29 casos reales muy interesantes. La traducción se realizó en 1992 bajo los auspicios de la SEIO. Los ejemplos son muy interesantes, aunque lógicamente en aquellos momentos internet apenas estaba desarrollada. Tiene un estilo menos divulgativo y de hecho utiliza conceptos que requieren una cierta formación. El público al que va dirigido es más bien universitario de la rama científica.

El libro "El tigre que no está. Un paseo por la jungla de la estadística" de Blastland y Dilnot (2009) tiene un estilo semejante al propuesto. El libro "Qué es (y qué no es) la estadística" de Sosa Escudero (2014) está más bien dirigido a un público con una cierta formación en el método estadístico. La mayor parte de los ejemplos son del mundo de la economía. Por tanto se podría decir que es un libro orientado a estudiantes y profesionales del mundo de la economía. En su momento tuve ocasión de revisar el libro "Organizando la estadística", en el que Chamoso (2007), mediante un paseo por la ciudad, descubre ejemplos, fundamentalmente de estadística descriptiva, en lo que va viendo. "Estadística para todo(s)" (2014), editado por Etayo Gordejuela y Fernández Fernández recoge las contribuciones más interesantes programadas en el ciclo de talleres divulgativos "Matemáticas en acción" que la Universidad de Cantabria ha venido organizando desde el curso 2004/05. El libro está escrito en un tono divulgativo y va dirigido a personas con curiosidad en temas científicos.

Recientemente se han publicado algunos otros libros de divulgación de la estadística en castellano. Son más frecuentes, sin embargo, los libros de divulgación de las matemáticas, que incluyen algo de estadística.

5. Un falso extracto del libro

En realidad, lo que escribo a continuación no está en el libro, pero es del mismo estilo y probablemente aparecerá en una próxima edición del mismo. Es fruto de un programa de divulgación en la radio en el que cada semana se habla de un teorema matemático. Cuando me pidieron intervenir no tuve duda en elegir el mágico y mítico Teorema Central del Límite (TCL).

La variables Normales, también llamadas de Gauss, aparecen en la naturaleza con muchísima frecuencia: estaturas, pesos, longitudes y otras medidas de seres vivos, fósiles o minerales; la resistencia a la tensión de determinadas piezas de acero o las calificaciones de un examen, por citar unos pocos ejemplos muy dispares. Se trata de datos cuya representación gráfica se ajusta a la bien conocida campana de Gauss. Además muchas variables utilizadas en la estadística proceden de ella de una forma u otra. Todo esto, además de sorprendente, es de gran importancia en la estadística moderna.

Durante mucho tiempo se pensó y se tomó como axioma que con un núme-

ro suficiente de observaciones todas las variables se aproximaban a la Normal. De ahí procede la famosa frase de Lippman: "todos creen en la ley Normal de errores; los experimentadores, porque piensan que es un teorema matemático; los matemáticos, porque creen que es un hecho experimental". Por este motivo recibe el nombre de Normal, lo que no quiere decir que en una variable que no sea de Gauss haya algo de anormalidad.

Aunque parezca increíble, un teorema matemático explica el enigma que rodea a la campana de Gauss. Este teorema garantiza, bajo ciertas condiciones no muy exigentes, que la suma de muchas variables es aproximadamente Normal (Gaussiana). ¿No decimos con frecuencia que determinada cuestión depende de muchas variables? Pues bien, al final eso es lo que ocurre en la naturaleza. Muchas de las magnitudes que observamos o medimos, en realidad son combinación de otras muchas. Por ejemplo, el peso corporal es una combinación de otras variables como la comida ingerida, el ejercicio físico, factores que controlan el metabolismo y un largo etcétera.

Pero, ¿todas las campanas son de Gauss? La respuesta es un rotundo ¡no! ¿Cómo distinguir entonces una campana de Gauss de una que no lo es? Fundamentalmente tiene dos características esenciales. Por una parte es simétrica respecto de la media, de modo que las puntuaciones altas se distribuyen de forma análoga a las bajas. Por otro lado, no hay observaciones extremas, ni demasiado pequeñas, ni demasiado grandes. Un ejemplo un poco radical de campana no Gaussiana es la así llamada, de Cauchy. Números muy extremos no son infrecuentes en este caso. A pesar de ser simétrica en torno a un valor, su media no existe. Este es otro misterio que dejamos para mejor ocasión.

Pero además, alrededor de este teorema gira la Estadística moderna. En particular, nos asegura que si en lugar de trabajar con una variable, lo hacemos con la media aritmética de las observaciones, por ejemplo de unos cuantos experimentos, esa media también se puede considerar parte de la familia Normal, al menos aproximadamente. Para conseguir esto es necesario que se realice un número suficiente de experimentos, que en la práctica no es excesivamente grande. Esto se deriva también de la esencia del teorema. Hay que resaltar que en muchos de los análisis estadísticos lo que interviene habitualmente son medias y por tanto esto es aplicable.

La estadística moderna busca explicar la realidad con modelos matemáticos. Uno de los estadísticos modernos más famosos, George Box, solía decir que "todos los modelos son falsos, pero algunos son útiles". Por eso se buscará un modelo razonable para ajustarlo a la realidad de un determinado fenómeno. Precisamente por no ser un modelo perfecto, existirá una fractura entre él y la realidad que trata de emular. Eso es lo que comúnmente llamamos error de ajuste y que nos gustaría medir de alguna manera para saber lo bien o mal que estamos explicando la realidad. Ahora bien, ese error es la caja donde metemos todo lo que desconocemos o no somos capaces de controlar. Bien empaquetado le

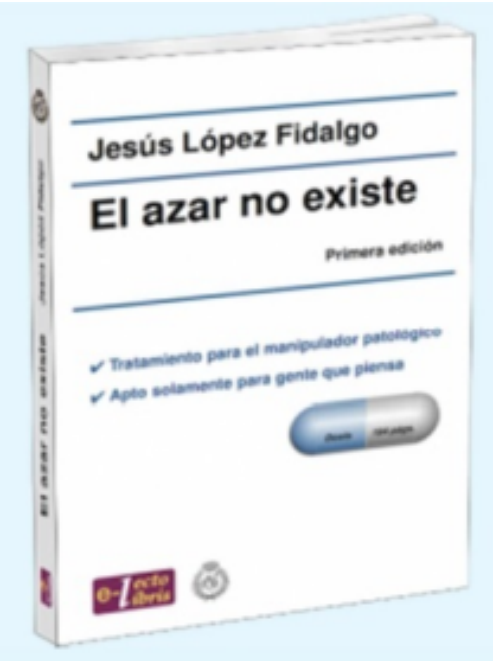
ponemos la etiqueta de azar, y efectivamente se rige por las leyes, bien conocidas, del azar, es decir, por la probabilidad. Si esto fuera tan sencillo como calcular la probabilidad de obtener un seis al lanzar un dado el problema estaría resuelto. Pero precisamente hemos dicho que ahí introducimos lo que desconocemos o no podemos controlar. Parece que hemos llegado entonces a un callejón sin salida después de una apasionante persecución. Y aquí es donde viene en nuestra ayuda el super héroe TCL, que nos permite aproximar esta probabilidad.

Por ejemplo, nos permite calcular la probabilidad de lo raras que son las pruebas que tiene un juez en el supuesto de que el acusado fuera inocente. Esto le llevaría a tomar una decisión, por ejemplo de condenarle si esa probabilidad es pequeña, de modo que las pruebas pueden considerarse claramente determinantes. La analogía del juicio y el juez es válida para cualquier decisión científica. Por ejemplo, si queremos saber si una moneda está trucada, podemos lanzarla 100 veces. Supongamos que obtenemos 63 caras, cuando deberíamos obtener aproximadamente unas 50. La probabilidad de obtener exactamente 63 caras con una moneda no cargada puede calcularse fácilmente y es aproximadamente de 3 entre 1.000. Pero hay que calcular la probabilidad de obtener casos tan raros como éste o más. Así la probabilidad de obtener 63 o más caras sería de 6 entre 1000. Pero también hay casos igual de raros por el otro lado, en concreto de obtener 37 caras o menos. Esa probabilidad es el doble, es decir aproximadamente de 12 entre 1000, que sigue siendo muy pequeña, por lo que sigue siendo muy sospechosa la regularidad de la moneda. Por tanto aquí rechazaríamos la hipótesis de que la moneda es correcta y obraríamos en consecuencia. El problema es que los casos reales a los que nos enfrentamos no son tan sencillos como lanzar una moneda y una probabilidad de este estilo no puede calcularse fácilmente. El teorema central del límite proporciona una herramienta extraordinaria para aproximar este tipo de probabilidades.

Una vez más se comprueba que la matemática no es un invento de un grupo de privilegiados para maltratar al resto de los seres humanos, especialmente en su periodo de formación escolar. Tampoco es un mal menor, que resulta útil en muchos campos. Es realmente algo muy natural, como hemos visto con el Teorema Central del Límite.

Agradecimientos

A todos los que me han dado ideas, han hecho críticas constructivas y me las siguen haciendo. Al Departamento de Matemáticas y al Instituto de Matemática Aplicada a la Ciencia y a la Ingeniería de la Universidad de Castilla-La Mancha que con su contribución económica han hecho posible que este libro viera la luz en un formato atractivo.



Referencias

[1] Aguilera del Pino A. M. (Ed.) (2013). Special Issue on The International Year of Statistics. *BEIO*, **29**,(3).

[2] Blastland M. y Dilnot A. (2009). *El tigre que no está. Un paseo por la jungla de la estadística*. Editorial Turner, Colección Noema.

[3] Box G.E.P. (2013). *An Accidental Statistician. The Life and Memories of George E.P.Box.*. John Wiley & Sons, Inc., Hoboken.

[4] Chamoso, J. (2007). *Organizando la estadística*. Editorial Nivola.

[5] Etayo Gordejuela F. y Fernández Fernández L-A. (Eds.) (2014). *Estadística para todo(s)*. Universidad de Cantabria.

[6] Huff D. (2011). *Cómo Mentir Con Estadísticas*. Crítica.

[7] López–Fidalgo J. (2014). Impact of the International Year of Statistics in Spain, has the effort been worthwhile? *BEIO*, **30**(2), 199-219.

[8] Sosa Escudero W. (2014). *Qué es (y qué no es) la estadística*. Editorial Siglo XXI.

[9] Tanur J.M., Mosteller F. (1992). *La estadística una guía de lo desconocido*. Alianza Editorial.

Acerca de los autores

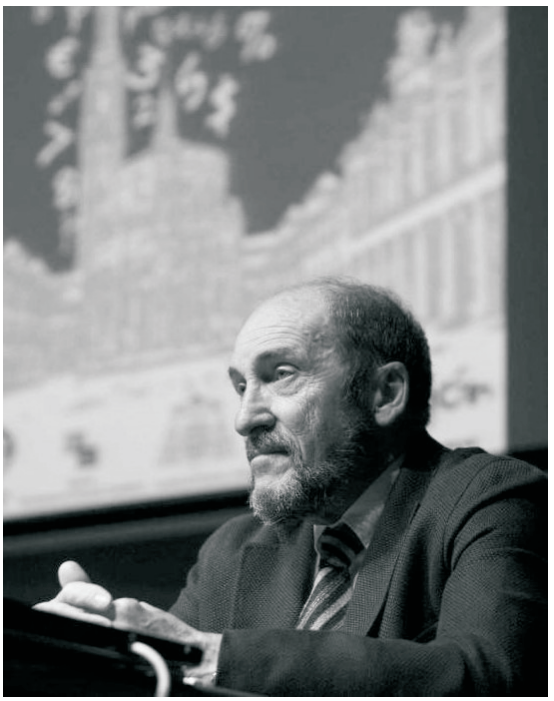
Jesús López Fidalgo es Catedrático de Estadística e Investigación Operativa en la Universidad de Castilla-La Mancha (UCLM). Ha sido Postdoctoral Fellow en la University of Manchester, Institute of Science and Technology (UMIST, 1992), Visiting Scholar en el Department of Biostatistics de la University of California, Los Angeles (UCLA, 1998/99) y Visiting Professor en el Department of Statistics en la University of California, Riverside (UCR, 2005). Ha sido vocal del Consejo Académico y lo es del Ejecutivo de Estadística de la Sociedad de Estadística e Investigación Operativa (SEIO) y editor del Boletín de la SEIO (2005-2008). Es miembro electo de ISI y editor asociado de Test y Sankhya B, entre otras revistas científicas. Ha sido Director de la Escuela Técnica Superior de Ingenieros Industriales y Presidente de la Comisión Electoral de la UCLM de 2008 a 2016. Su línea principal de investigación es el diseño óptimo de experimentos y ha desarrollado múltiples colaboraciones en Estadística aplicada. Ha publicado trabajos en revistas de reconocido prestigio, como son la Journal of the American Statistical Association; Journal of the Royal Statistical Society, series B; Bioinformatics o Technometrics. Desde enero de 2009 hasta diciembre de 2011, ha sido gestor del Programa Nacional de Matemáticas.

Boletín de Estadística e Investigación Operativa
Vol. 32, No. 2, Julio 2016, pp. 171-179

Opiniones sobre la profesión

Pedro Gil (1947-2016). Obituary. A Pedro: Maestro, Mentor, Compañero y Referente

Norberto Corral, María Ángeles Gil and Manuel Montenegro
Departamento de Estadística e I.O. y D.M.
Universidad de Oviedo
✉norbert@uniovi.es, ✉magil@uniovi.es, ✉mmontenegro@uniovi.es

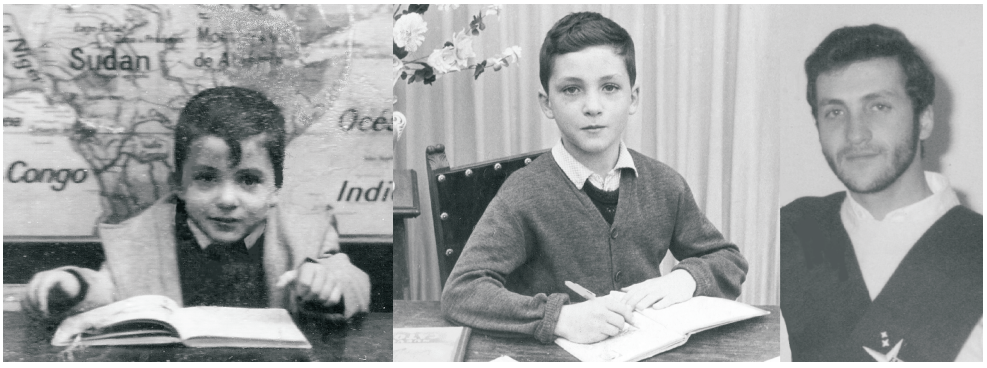


Empezamos a escribir este obituario exactamente un mes tras la despedida de Pedro Gil. Nuestro querido maestro, mentor, compañero y referente. Lo hacemos en representación del que ha sido y será siempre su Departamento, movidos por una necesidad inexorable de compartir nuestra visión de Pedro con tantos y tantos que le habéis conocido.

Somos vástagos científicos de Pedro, como la mayoría de los miembros de su departamento en el que todos nos sentimos orgullosos de ser sus discípulos porque de él hemos aprendido tantas cosas...

1. PEDRO MAESTRO

Pedro nació en Valladolid, en 1947. Siempre tan prudente y pertinente, esperó a nacer el día que el santoral tiene reservado para el nombre que le habían destinado con antelación: el de su abuelo materno. Hijo de maestro con mucha querencia por las Matemáticas y la Música, estaba en su predisposición genética la vocación por las mismas. Y nunca encerró dudas sobre su orientación profesional hacia la primera y su dedicación, de índole lúdica, a la segunda.



Pedro en la etapa escolar infantil (izquierda),
en el Bachiller en el Colegio de los Maristas en Valladolid (medio)
y estudiante universitario en la UCM (derecha)

Tras realizar sus estudios pre-universitarios y el primer año universitario (Selectivo) en Valladolid, en 1965 se desplazó para estudiar Ciencias Exactas en la Universidad Complutense de Madrid, ya que aún faltaban dos años para que en su ciudad natal se pusieran en marcha tales estudios. En cuarto curso, optó por la especialidad de Estadística e Investigación Operativa, y al concluir la licenciatura se incorporó al Departamento correspondiente, dirigido por el Profesor Sixto Ríos (a quien siempre se refirió con admiración, agradecimiento y respeto, como 'Don Sixto' o 'el Jefe', su maestro).

En 1974, Pedro se doctoró en la UCM con la tesis dirigida por el Profesor Ríos titulada "Medidas de incertidumbre e información en problemas de decisión estadística", publicada por la Real Academia de Ciencias Exactas, Físicas y Naturales de Madrid (Gil, 1975). Y dos años más tarde, tras obtener la agregaduría en Investigación Operativa para la Universidad de Oviedo, él inició su andadura como ***maestro***. Y damos fe de que logró completar el camino con creces.

De este modo, con 28 años, Pedro se trasladó a Oviedo con su constituida familia propia (entonces integrada por Pilar y Eva, su mujer e hija mayor), su principal apoyo y motor de toda su vida.

Su única pena era abandonar el nido científico que en torno a su jefe habían creado tantos entrañables compañeros y amigos, de los que afortunadamente nunca se desvinculó. Muchos de ellos habían emprendido el vuelo hacía poco, y otros lo emprenderían en los años posteriores. Nunca olvidó a ninguno de ellos,

y los Congresos de la SEIO supusieron una excusa perfecta para el añorado reencuentro.

1.1. Pedro profesor

Pedro ya había ejercido como profesor ayudante durante seis años en la Complutense. En tan poco tiempo se hizo cargo de diversas materias, lo que le confirió una formación impagable como profesor, que le sirvió no sólo para la docencia que impartiría en los siguientes treinta y cuatro años sino para guiar y asesorar a todos los que hemos formado parte de su departamento. Con muchos de sus estudiantes de esa etapa conservó siempre una relación profesional/personal muy estrecha. Entre ellos, mantuvo una fuerte vinculación con especialistas en Teoría de la Decisión como Pilar García Carrasco (su primera doctoranda) y los hermanos Susi y Sixto Ríos Insua, hijos de su maestro. Y la devoción por la Teoría de la Información Estadística que compartía con Leandro Pardo les hizo llegar a ser grandes amigos; Pedro sentía un orgullo enorme al ver a Leandro convertirse en un investigador de referencia internacional en el campo.

En 1976 la Universidad de Oviedo no contaba entre sus titulaciones con Matemáticas. Pedro, junto con unos pocos neófitos en tareas docentes universitarias bajo su tutela, asumió la enseñanza de Matemáticas y Estadística en las licenciaturas en Química, Biología, Geología y Económicas, haciéndose cargo de impartir varias de ellas. Con los años, se fueron adscribiendo nuevas materias, y tras la incorporación en 1990 de la titulación de Matemáticas (en cuya creación el hacer de Pedro fue determinante), se responsabilizó de la asignatura de Probabilidades y Estadística I y de Teoría de la Información en dicha carrera.

A pesar de ser Pedro el primer matemático que impartía docencia en muchas de las materias en las titulaciones no matemáticas, las reticencias iniciales ante el indudable aumento de rigor y de nivel de exigencia en ellas pronto fueron superadas. Bastó para ello comprobar que ese aumento iba acompañado de unas explicaciones claras, de unas motivaciones bien argumentadas y de una empatía con los alumnos a prueba incluso de los más aversos a las Matemáticas y la Estadística.

Y en cuanto empezó a dar clase de las asignaturas en la Licenciatura de Matemáticas, y descartada esa posible aversión, sus habilidades docentes le hicieron un profesor aún más valorado, si cabe. Pedro sabía mucho y sabía transmitirlo. Y su forma serena de exponer los asuntos de mayor enjundia hacía que, a menudo, los estudiantes creyeran que las lecciones recibidas eran más sencillas de lo que realmente eran. Y ahí estaba él para resolverles las dudas de primera o última hora y, de paso y si se terciaba, escucharles las incertidumbres sobre su presente y futuro profesionales y otros problemas que les aquejaran, aconsejándoles si así lo demandaban.

La puerta de su despacho siempre estaba abierta si él se encontraba dentro. Esa puerta abierta era señal inequívoca de la cercanía de Pedro y constituía el

remedio más eficaz para los recelos que un alumno pudiera tener antes de visitarle. Este sentimiento ha sabido describirlo muy bien Rodríguez-Muñiz, 2016, quien pudo disfrutar de esa oportunidad primero desde la posición de estudiante y más tarde como compañero.

1.2. Pedro investigador

Las numerosas responsabilidades de gestión que Pedro tuvo que asumir desde su llegada a Oviedo, no le permitieron desarrollar la actividad investigadora hasta el punto que le habría gustado, pero era mucha la importancia que Pedro concedió siempre a esa faceta. De hecho, acabamos de verificar cuál era su índice h según la *Web of Science*: 15. Bastante digno para un matemático que, con seguridad, nunca se preocupó de calcularlo y que no dispuso del tiempo de dedicación que habría querido para esa tarea. Pedro trató de inculcarnos que, sin descuidar la docencia, atendiéramos bien a la investigación. Suponía además una forma de seguir apoyando a los más jóvenes, animándoles a que iniciaran sus propias líneas de trabajo, tanto básicas como aplicadas.

Creyó en la publicación de trabajos en buenas revistas antes de que se introdujeran los complementos que reconocían esa política, y no es de extrañar que llegara a tener seis sexenios de investigación, que en mucho tiempo fue una situación poco frecuente en nuestra área.

También apostó sin ambages por la solicitud de proyectos del Plan Nacional de Investigación, dirigiendo un buen número de ellos que en ocasiones llevaron asociados becarios FPI. Y mediante estos proyectos impulsó la participación en congresos nacionales e internacionales y la realización de estancias breves de los miembros de su grupo de investigación.

Los intereses investigadores de Pedro estuvieron siempre relacionados con lo que él mismo denominaba “Las Matemáticas de lo incierto” y que resumió bien en su lección inaugural del Curso 1996-1997 (ver Gil, 1996) y que incluían: las Matemáticas del azar (Probabilidades y Estadística), las Matemáticas de la comunicación (Teoría de la Información) y las Matemáticas de la imprecisión (Lógica Fuzzy). Y esos intereses, con sus matices y concreciones o extensiones, sentaron las bases de gran parte de la investigación que actualmente se desarrolla en su departamento.

En materia de investigación, Pedro no comulgaba con las estructuras piramidales. Por ello, nos fue empujando con suavidad y firmeza a independizarnos. Representaba un acto de confianza en nuestras capacidades, pero en realidad era una prueba más de su generosidad. Y estaba muy a favor del relevo generacional, por lo que renunció en los últimos años de su carrera académica a liderar su grupo de investigación, que ha continuado desarrollando algunas de sus líneas prioritarias junto con muchas otras nuevas.

1.3. Pedro gestor

Desde el momento en que Pedro llegó a Asturias, tuvo que hacerse cargo de diversos puestos de gestión. No le molestaban y sabía de su competencia para ejercerlos, pero en muchos momentos le habría gustado dejarlos a un lado y volcarse más en la docencia y en la investigación. Pero también era consciente de que, para un departamento que acababa de nacer en la Universidad de Oviedo, era conveniente que su responsable principal se involucrara en tareas relativas a los centros en los que se impartía docencia. De este modo, además de Director del Departamento de Matemáticas de Ciencias y Económicas, Pedro fue Secretario de la antigua Facultad de Ciencias (Química, Biología y Geología) y, tras la disgregación de Ciencias en tres Facultades, ViceDecano de la Facultad de Biología, a la que el departamento estaba adscrito y en la que impartía 72 créditos de licenciatura.

A raíz de la implantación de la L.R.U. y de la creación de las áreas de conocimiento y los Departamentos acordes con dicha ley, en la Universidad de Oviedo se constituyó un Departamento de Matemáticas que incluía ocho áreas con alrededor de 200 profesores. Desde su constitución hasta mediados de 1997, Pedro fue Subdirector del mismo.

En 1997, se produjo una partición del Departamento de Matemáticas en cuatro, uno de los cuales integró a las áreas de Estadística e Investigación Operativa y de Didáctica de la Matemática y que sigue vigente en la actualidad. Desde 1997 hasta acogerse a una propuesta de jubilación anticipada en 2010, Pedro fue el Director de ese departamento. Muy a su pesar, no consiguió que nadie le relevara en la dirección mientras él estuvo en activo. Desde sus inicios, el departamento fue un ejemplo absoluto de equilibrio de género, y en la actualidad de los 33 profesores que lo conforman 16 son hombres y 17 mujeres, con un reparto de niveles que parece de diseño. Estamos convencidos de que no fue fruto de una política preconcebida, sino del azar y de la apuesta firme de Pedro por los méritos de cada cual como criterio exclusivo de acceso.

Pedro presidió la SEIO, acompañado por Mariano Valderrama (Vicepresidente de Estadística), Ignacio García Jurado (Vicepresidente de Investigación Operativa) y Susi Ríos (Secretaria General), entre los Congresos SEIO de úbeda y Cádiz, es decir desde Noviembre de 2001 hasta Octubre de 2004. Junto con su mujer, Pilar Fernández de Sanmamed, fueron siempre una presencia segura en todos los Congresos de la Sociedad, incluso tras la jubilación.

Y durante dos años, 2006 y 2007, formó parte del Comité Asesor 1 de la Comisión Nacional Evaluadora de la Actividad Investigadora.

Fue miembro de otras muchas comisiones y comités científicos y ejecutivos, que no vamos a enumerar. A este respecto no podemos pasar por alto su implicación y entrega a una actividad que se tradujo en la materialización de uno de los objetivos que se había marcado al llegar a Asturias: la implantación de la

Licenciatura de Matemáticas. Junto con Javier Valdés, compañero en el Departamento de Matemáticas, y con las colaboraciones puntuales de otros miembros del mismo, elaboraron un Plan de Estudios que se puso en marcha en el curso académico 1990-1991. Aunque esa gestión no llevó asociado un reconocimiento oficial, el papel que Pedro desempeñó en la creación de los estudios de Matemáticas en la Universidad de Oviedo es un elemento indiscutible de su historia.

Pedro hizo mucho por las Matemáticas en Asturias. Y no sólo por las universitarias, ya que siempre fomentó la cooperación entre la enseñanza universitaria y la enseñanza previa. Fue durante muchos años coordinador de COU-PAU, coordinador de la Olimpiada Matemática para alumnos de COU/Bachillerato y, en los últimos años, miembro del jurado en la etapa regional del Concurso Incubadora de Sondeos y Experimentos, para alumnos de ESO y Bachillerato.

2. PEDRO MENTOR

Pedro ha sido mentor de muchas personas y defendía con fervor la necesidad de investigar, repitiendo a menudo que, aunque la tesis doctoral debía ser un trabajo de gran entidad, no tenía que ser nuestra principal contribución investigadora: había que ir a más.

Con estas premisas dirigió veinte tesis doctorales. Su primera doctoranda fue Pilar García Carrasco, y su tesis empezó a supervisarla en la UCM. Estaba dedicada a la comparación de experimentos a través de medidas en el marco de las Teorías de la Información y de la Decisión. Y en la UniOvi dirigió diecinueve tesis. En total, 9 doctorandos y 11 doctorandas.

Los logros de sus discípulos han sido fruto de su magisterio pero, en su afán de restarse méritos, Pedro recurría a citar con frecuencia una sentencia de los Romances del Cid: *"Si non vencí reyes moros, engendré quien los venciera"*. Y no sólo lo aplicaba a sus doctorandos, sino a sus alumnos de Matemáticas, a muchos de los cuales supervisó en sus trabajos académicamente dirigidos. En el homenaje que se le rindió en 2010, al que vamos a referirnos a continuación, Pedro manifestaba que *"No os fijéis sólo en los alumnos excepcionales; hacedlo también en los que no llegan a serlo"*.

Como mentor, Pedro te garantizaba apoyo, trabajo, ánimo, seguimiento y mucha comprensión. Salvo en un aspecto: ensayar concienzudamente las exposiciones. Como él tenía facilidad natural para presentar cualquier tema en un lenguaje asequible y de forma literariamente brillante, no entendía la necesidad de tanto ensayo. Aprendimos a no discutir con él sobre tal cuestión, y simplemente practicábamos 'a escondidas'.

3. PEDRO COMPAÑERO

Como ya hemos mencionado, Pedro llegó a Oviedo con 28 años y su agregaduría (el paso previo a la cátedra, que en esos momentos se obtenía por concurso

de méritos) bajo el brazo. Y se quedó aquí por otros cuarenta. Cuando, tiempo antes de convocarse por la Universidad de Oviedo, se convocó en Santiago de Compostela una cátedra, que habría obtenido sin problemas y le habría acercado mucho a la tierra de su mujer (la Puebla del Caramiñal, donde ahora descansan sus cenizas), sus jóvenes compañeros de la época temieron que sus vidas académicas iban a sufrir un cambio enorme al perder su dirección y protección. Pero al pensar en aquellos jóvenes, en realidad sólo un poco menores que él pero que se sentían tan seguros bajo su guía, decidió permanecer en Asturias.

Y Pedro ha sido feliz en su Departamento, en esta tierra en la que nacieron dos de sus hijos (Juan y Eduardo, también matemáticos), y en la que ha sido tan querido. Pedro ha sido un buen académico, un profesor muy apreciado y un compañero generoso sin afán de protagonismo.

A la izquierda, aún con cara de sorpresa ante el homenaje secreto
que se le brindó en Noviembre de 2010 en el Paraninfo de la Universidad de Oviedo.
A la derecha, Pedro con el Presidente del Principado, el Rector de la Universidad, el Director
del Departamento y el 'Coro Pantaleón'

Pero ninguno de nosotros le ha considerado un compañero más, porque él representó el germen y el corazón de aquel proyecto que se inició en 1976. En este sentido, el homenaje que se le tributó en Noviembre de 2010, y cuya organización desconoció hasta su entrada en el Paraninfo de la Universidad de Oviedo, fue un desfile de cariño, admiración y reconocimiento. Y se constituyó un coro (con ocho cantores, entre ex-alumnos, compañeros y hermanos) expresamente para aderezar las distintas participaciones del homenaje; y, al final, se pidió su colaboración al acordeón en memoria de las veces que había amenizado festividades de la titulación.

A pesar del secreto con el que se llevó la preparación, algunos de los últimos alumnos de Pedro consiguieron enterarse y nos pidieron intervenir con un discurso muy breve. ¡Menos mal que ellos subsanaron nuestro olvido! Pedro lo habría echado mucho en falta.

4. PEDRO REFERENTE

Pedro tuvo siempre mucho ascendente: sobre los amigos, sobre los compañeros, sobre los alumnos. Todos han destacado reiterada y unánimemente que sabía escuchar, que transmitía serenidad, que daba consejos expertos y reflexivos.

Y por ello, y sin buscarlo, fue muy influyente. Hasta en la afición por las Matemáticas. él fue el primer matemático de su familia; y a partir de él siguieron dos hermanos (que además coincidieron hasta en la especialidad), dos hijos, dos hermanos de su mujer, Pilar, dos sobrinos,... Atribuir al azar esta influencia, cuando menos resulta un atrevimiento estadístico.

Uno de sus sobrinos matemáticos, Iago Fernández de Sanmamed, escribió tras la marcha de Pedro la reflexión siguiente, que nos ha autorizado a reproducir porque resume fielmente nuestro sentir:

"¿Qué es un referente? No creo que sea alguien que te enseñe cosas, ni que te ponga las cosas fáciles. Un referente para mí es alguien que enseña un camino, que sólo hablando, opinando y siendo como es, marca la diferencia.

Hoy despido a mi referente, a alguien con un corazón de cristal que sólo siendo así de bueno y transparente fue capaz de llegar a muchísimas personas.

Empecé diciendo que un referente no enseña; me equivoqué, él me enseñó a vivir, a luchar frente a todo, a que siempre hay esperanza y a ser quien soy. A través de ti, a través de ese corazón de cristal, vi matemáticas, vi música y sobre todo vi bondad y una grandísima persona. Te echaré de menos, tío, amigo y referente."

Pedro, ¡cuánto habríamos deseado no tener que escribir estas lineas jamás! Que tu 'corazón de cristal', como dice Iago, no hubiera fallado y hubieras salido airoso como lo hiciste de otros contratiempos en tu salud. Aunque intentemos seguir tu estela, nunca podremos más que aproximarnos, pero ahí estaremos. Tu recuerdo y tu ejemplo estarán siempre con nosotros. De nuevo hemos pedido prestadas unas palabras, en este caso las que te dedicó tu sobrina Sabela Fernández de Sanmamed.

"Como diría Neruda: Te recuerdo como eras en el último otoño. Eras la boina gris y el corazón en calma.

Yo te recuerdo como el compañero perenne, la voz rasgada, las patatitas con pimentón, el cuñadito del alma.

Eres el recuerdo de dos sombras en la huerta, dos sonrisas, dos lectores serenos, un crucigrama y un sudoku con ojos serenos. Eres infinitas matemáticas, pijamas cortos a rayas, manos trabajadas, 'el bueno, el guapo y el listo'. Hoy eres el dolor que rasga el alma".

¡Hasta siempre, Pedro!

Referencias

[1] Gil, P. (1975). Medidas de incertidumbre e información en problemas de decisión estadistica. *Rev. Real Acad. Cienc. Exact. Fis. Natur. Madrid*, **LXIX**, 549–610 (http://www.rac.es:8080/fedora/get/Revistas:REV_20091030_00137/PDF).

[2] Gil, P. (1996). Las matemáticas de lo incierto: lección inaugural del curso académico 1996-1997. *RAE: Revista Asturiana de Economía*, **7**, 203–219 (http://digibuo.uniovi.es/dspace/bitstream/10651/28625/1/-matematicasincierto.pdf).

[3] Rodríguez Muñiz, L.J. (2016). La puerta siempre abierta de Gil. *La Nueva España*, 20/03/2016 (http://www.uniovi.es/-/la-puerta-siempre-abierta-de-gil?redirect).

Acerca de los autores

Norberto Corral es Catedrático de Estadística e Investigación Operativa de la Universidad de Oviedo. Es hijo científico de Pedro y, a día de hoy, Decano de la Facultad de Ciencias (Matemáticas y Física).

María Ángeles Gil es Catedrática de Estadística e Investigación Operativa de la Universidad de Oviedo. Es hija científica de Pedro y hermana de sangre.

Manuel Montenegro es Profesor Titular de Estadística e Investigación Operativa de la Universidad de Oviedo. Es nieto científico de Pedro y su sucesor al frente de la Dirección del Departamento de Estadística e Investigación Operativa y Didáctica de la Matemática.

http://www.seio.es/BEIO