BEIO

Boletín de Estadística e Investigación Operativa

Volumen 37
Número 1
Marzo 2021

ISSN: 2387-1725

Revista Oficial de la Sociedad de Estadística e Investigación Operativa

BEIO Revista Oficial de la Sociedad de Estadística e Investigación Operativa

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. 37, No. 1, Marzo 2021

Índice

Boletín de Estadística e Investigación Operativa
Vol. 37, No. 1, Marzo 2021, pp. 1-2

Editorial

Anabel Forte
Depto. de Estadística e IO
Facultad de Ciencias Matemáticas
Universidad de Valencia
✉ anabel.forte@uv.es

Francisco Parreño
Depto. de Matemáticas
Escuela Superior de Ingeniería Informática
Universidad de Castilla-La Mancha
✉ francisco.parreno@uclm.es

Hace pocos meses comenzaba nuestra andadura como editores del Boletín de Estadística e Investigación Operativa (BEIO), una andadura de la que he aquí el primer fruto, el primer número que hacemos como editores. Pero este fruto no es solo merito nuestro si no de quienes nos precedieron en el puesto, M. Teresa Santos y Salvador Naya, a quienes queremos agradecer todo el trabajo realizado durante estos años, así como su valiosa ayuda en el traspaso de las riendas de esta publicación.

Del mismo modo, nos gustaría dar las gracias al comité editorial, Rosa M. Crujeiras, César Guitérrez, Pedro Revilla y M. Carmen Escribano, cuyo trabajo resulta fundamental para asegurar la calidad de los artículos publicados. Por supuesto, agradecer también al equipo técnico, Antonio Elías y María Jesús Gisbert por su eficiente labor sin la cual, esta publicación sería imposible. También agradecer a Paula Saavedra por aceptar formar parte del equipo técnico.

BEIO es la muestra fehaciente del compromiso de la Sociedad de Estadística e Investigación Operativa (SEIO) con la divulgación de un área de investigación que, a pesar de estar presente en todas las ciencias, resulta casi invisible y es, en muchas ocasiones, mal interpretada.

La reciente pandemia ha dejado patente la necesidad, no solo de la metodología estadística y de investigación operativa, sino también de generar cultura estadística en una sociedad que cada día se ve avocada a la comprensión de gráficas y de datos que aparecen por doquier.

Creemos firmemente, que BEIO debe continuar siendo un engranaje de divulgación y de transmisión de conocimiento dentro de la SEIO, trasladando a

toda ella y en un lenguaje común, las investigaciones que se están llevando a cabo en la actualidad. Pero también creemos que es importante que este boletín sea una ventana abierta a la sociedad en general que permita mostrar como esta ciencia contribuye a la comprensión del mundo que nos rodea.

Es por ello por lo que mostramos nuestra intención de hacer la revista más accesible y atractiva, manteniendo las secciones actuales, pero añadiendo otras nuevas que puedan ser de interés a un público más general contribuyendo, en la medida de nuestras posibilidades, a amplificar y arraigar la cultura de la Estadística y la Investigación Operativa en nuestra sociedad.

Entre estas nuevas secciones nos gustaría incorporar resúmenes de las tesis que se han defendido recientemente en el ámbito de la SEIO. Resúmenes que muestren, de forma divulgativa, el trabajo que están realizando las personas más jóvenes de nuestra sociedad. Con esta finalidad hemos incorporado al equipo editorial a Paula Navarro y Mercedes Pelegrín, como vínculo con una nueva generación que viene con mucha fuerza. Gracias a las dos por aceptar llevar esta nueva sección.

Así mismo, creemos importante incorporar una sección que nos ayude a entender cuáles son los retos del futuro desde el punto de vista del tejido empresarial. Cuáles son los retos a los que se están enfrentando o se van a enfrentar las empresas y que necesitan, de un modo u otro, nuestra colaboración. Con este fin hacemos un llamamiento a quienes trabajáis directamente con empresas para que nos contéis cuales son las aplicaciones que estáis llevando a cabo y que animéis a dichas entidades a que nos expliquen cuales son los retos que están por venir.

No queremos concluir esta editorial sin agradecer a quienes habéis publicado en BEIO y a quienes leéis la revista asiduamente. Sin vosotros y vosotras está publicación perdería su razón de ser.

Nos despedimos animando a quienes estáis pensando enviar vuestros artículos, vuestras propuestas y vuestras ideas de mejora a que nos las hagáis llegar porque entre todos y todas conseguiremos que esta ventana al mundo sea más amplia y entre por ella más luz.

Anabel Forte y Francisco Parreño
Editores de BEIO

Boletín de Estadística e Investigación Operativa
Vol. 37, No. 1, Marzo 2021, pp. 3-18

Estadística

Phase-Type distributions: computational aspects and applications in electronics

Juan Eloy Ruiz Castro, Christian J. Acal González and Ana María Aguilera del Pino

Departamento de Estadística e Investigación Operativa y IMAG
Universidad de Granada

✉ jeloy@ugr.es, ✉ chracal@ugr.es,
✉ aaguiler@ugr.es

Abstract

Reliability is an area of statistics that analyzes the behaviour of systems subject to failures where probability plays a fundamental role in modeling, solving and optimization problems. It is usual to develop methodologies that allow a detailed study through classic distributions. An important aspect is the estimation of the parameters. A class of non-negative distributions, called phase-type distributions, makes it possible to model complex problems with well-structured results, thanks to its matrix-algebraic form. The computational aspects of the estimation in this field, through statistical programmes or applications such as R, Matlab or EMpht, are revised and applied to a real data set from RRAM memories in order to prove that this approach is better than the classic statistical analysis employed in this area.

Keywords: Reliability, Modeling, Probability, Phase-type distributions, Resistive memories.
AMS Subject classifications: 97K80, 97K60, 62N05, 62P30.

1. Introducción

El análisis de la fiabilidad o supervivencia es un conjunto de técnicas del campo de la Estadística y la Probabilidad en diversas áreas de la ciencia como pueden ser la ingeniería o la medicina. En términos generales, esta rama del conocimiento es la encargada de estudiar el comportamiento de los sistemas, cuyo funcionamiento está condicionado y afectado por ciertas variables que son incontrolables (generalmente por condiciones ambientales) y provocan que estos sistemas estén sometidos a un continuo desgaste. Sirva como ejemplo el estudio

del tiempo de vida de una bombilla (sistema), que puede estar influenciado por el ambiente y el trato que reciba la propia bombilla. Otro aspecto a destacar es que el tiempo de vida (o análogamente la ocurrencia de fallo) es aleatorio entre las distintas unidades experimentales, ya que no todas vivirán el mismo tiempo aunque hayan sido fabricadas y estén operando en igualdad de condiciones. Siguiendo con el ejemplo de la bombilla, es claro que el tiempo de vida de cada una de ellas será diferente. En consecuencia, es aquí donde la Estadística, y especialmente la Probabilidad, juegan un papel fundamental en la modelización de los sistemas, puesto que dichos tiempos de vida o fallo podrán ser ajustados mediante una distribución de probabilidad conocida. Además, otro matiz a tener en cuenta es que, aunque por comodidad se hable de tiempo, hay que poner de manifiesto que un análisis de fiabilidad puede tratar de estudiar otras variables diferentes que no sean específicamente el tiempo, pero que estén altamente relacionadas con él. Por ejemplo, la aplicación que se muestra en el presente trabajo se refiere al caso particular de las memorias resistivas que son actualmente una de las tecnologías de memoria más prometedoras y están llamadas a reemplazar, o al menos complementar en algunos nichos de aplicación, las actuales memorias no volátiles de tipo flash que dominan el mercado ([15]). El functionamiento de estas memorias está basado en la formación y ruptura de un filamento conductor, cuyo proceso depende del voltaje que se suministre. Sin embargo, aunque la variable de interés sea el voltaje y no el tiempo, es claro que estos aparatos han estado funcionando en paralelo un determinado tiempo hasta el mencionado voltaje de fallo. Para más información acerca del caso particular de la modelización del tiempo de fallo de las memorias resistivas véase el trabajo de Long et al. [11].

La primera distribución empleada en un análisis de fiabilidad fue la distribución Exponencial en el trabajo de Epstein y Sobel en 1953 ([8]), que ha sido la distribución de referencia durante décadas para modelizar el tiempo de fallo en multitud de áreas, debido en gran medida a sus buenas propiedades, sencillez y aplicabilidad. Con el paso de los años, sin embargo, esta distribución quedó obsoleta, ya que modela el comportamiento de las unidades que fallan a una tasa constante, independientemente del tiempo acumulado, y esto no siempre ocurre en el mundo real. Este hecho provocó que se empezaran a usar otros tipos de distribuciones algo más complejas. Normalmente, las distribuciones más empleadas en este campo en la actualidad son la distribución Erlang, la distribución Weibull, la distribución Log-Normal o la distribución Gamma, entre otras. En la mayoría de los estudios de fiabilidad no se suele profundizar mucho más allá y se entiende que estas distribuciones, que son bien conocidas y ampliamente aplicadas en ramas de la ciencia, ingeniería o medicina, son suficientes para lidiar con los problemas reales que puedan surgir. No obstante, y especialmente en sistemas tecnológicos que están en continuo desarrollo por

el auge e importancia que está sufriendo este sector, los sistemas sometidos a estudio en el día a día suelen ser mucho más complejos, con estructuras internas más sofisticadas, que provocan que estas distribuciones no se ajusten bien a los datos, lo que impide una correcta interpretación de la realidad a través de los resultados. Ante esta situación, el analista se encuentra obligado a utilizar un enfoque diferente que solucione el problema planteado y mejore el ajuste.

Bajo el contexto descrito anteriormente, en los últimos años y gracias al desarrollo computacional, se han introducido las distribuciones Tipo Fase, que aún siendo un tipo de distribuciones muy poco conocidas, presentan una gran flexibilidad y unas excelentes propiedades que hacen que sean un candidato a tener en cuenta en estudios de fiabilidad. Por ejemplo, en [1] se demuestra mediante un estudio de simulación que las distribuciones Tipo Fase funcionan mejor que otras distribuciones clásicas consideradas en el sentido de que el ajuste que proporcionan es más preciso, poniendo de manifiesto que las distribuciones Tipo Fase deben ser tenidas en cuenta cuando se analicen datos experimentales relacionados con el fallo de un sistema. Asimismo, en [2] y en [15] se aplican estas distribuciones al caso específico de las memorias resistivas.

Las distribuciones Tipo Fase fueron introducidas por Neuts en 1975 (ver [12] y [13] para un estudio detallado). Esta clase de distribuciones no negativas generalizan una amplia colección de distribuciones conocidas como pueden ser la distribución Exponencial, la distribución Erlang o la distribución Coxiana, entre otras. Entre sus principales ventajas destacan que permiten modelizar problemas no exentos de dificultad, proporcionando resultados con una estructura clara y sencilla de interpretar gracias a su forma algebraico matricial. No obstante, uno de los resultados más interesantes de las distribuciones y razón fundamental por la que son aplicadas en múltiples ocasiones (no solamente en estudios de fiabilidad) fue el que obtuvo Asmussen en [4]. Este resultado detalla que cualquier distribución de probabilidad no negativa puede ser aproximada tanto como se desee mediante una distribución Tipo Fase. De esta manera, conocida la distribución ideal para modelizar el tiempo de vida del sistema, se puede extraer información adicional de cómo funciona el sistema, cuáles son sus principales características o cuánto se espera que viva, entre otros resultados.

El ajuste por máxima verosimilitud de los parámetros involucrados en una distribución Tipo Fase se consigue a través del algoritmo iterativo EM, el cual alterna dos pasos: esperanza y maximización. Este método fue desarrollado por Asmussen en [3] y asumido por Buchholz en [6]. En la red están disponibles una serie de paquetes implementados en Matlab y en R, e incluso existen aplicaciones para ordenador, que pueden ser utilizados para determinar la estructura de las distribuciones Tipo Fase y para la estimación de sus parámetros. Sin embargo, el coste computacional de estas funciones implementadas en los paquetes estadísticos nombrados puede ser alto si el número de fases óptimo es elevado.

Entonces, en situaciones donde es conocida la estructura de las distribuciones Tipo Fase, puede ser recomendable implementar un código que rebaje el costo computacional mencionado. En este trabajo se ha desarrollado un algoritmo en R que puede ser utilizado cuando los datos experimentales siguen una distribución Tipo Fase con estructura Erlang, como es el caso del voltaje de fallo en memorias resistivas.

Este trabajo queda estructurado de la siguiente manera. La Sección 2 está centrada en definir las distribuciones Tipo Fase y en describir sus características más importantes. También se muestra la expresión que toman algunas de las funciones más comunes en un estudio de fiabilidad en el caso de que las distribuciones Tipo Fase sean consideradas. En la Sección 3 se consideran algunos de los paquetes encargados de estimar la estructura y parámetros de una distribución Tipo Fase y se describe el código desarrollado por los autores del trabajo. La aplicación a los datos reales se puede ver en la Sección 4. Finalmente, se incluye una sección en la que se muestran las conclusiones obtenidas en el presente trabajo.

2. Distribuciones Tipo Fase

Las distribuciones Tipo Fase, introducidas por primera vez por Neuts, son consideradas en diversas ramas de la ciencia e ingeniería gracias a que aceptan representaciones algorítmico-matriciales que son muy útiles de cara al análisis teórico y cálculo numérico. Asimismo, las distribuciones Tipo Fase constituyen una clase versátil que permite que cualquier distribución de probabilidad no negativa pueda ser aproximada tanto como se desee mediante una distribución Tipo Fase.

Una variable aleatoria no negativa X sigue una distribución Tipo Fase si su función de distribución viene dada por la siguiente expresión:

$$F(t) = 1 - \alpha e^{\mathbf{T}t}\mathbf{e}, \quad t \geq 0, \tag{2.1}$$

donde $\mathbf{e}$ es un vector columna de dimensión apropiada cuyos elementos son 1, α es un vector sub-estocástico de orden m, es decir, un vector cuyos elementos son no negativos y $\alpha\mathbf{e} \leq 1$, donde m es un entero positivo, $\mathbf{T}$ es un sub-generador de orden m, es decir, $\mathbf{T}$ es una matriz $m \times m$ tal que todos los elementos de la diagonal principal son negativos, todos los elementos fuera de la diagonal son no negativos o cero, todas las filas suman valores no positivos o cero y es invertible.

De la definición anterior se pueden extraer otras funciones empleadas en el ámbito de la fiabilidad como son la función de densidad, $f(t)$, la función de fiabilidad, $R(t)$, la función razón de fallo, $h(t)$ y la función de fallo acumulada, $H(t)$. Las expresiones que adoptan estas funciones en las distribuciones Tipo Fase figuran a continuación:

$$
\begin{aligned}
f(t) &= \alpha e^{\mathbf{T}t}\mathbf{T}^0, \quad t \geq 0, \quad \mathbf{T}^0 = -\mathbf{T}\mathbf{e}, \\
R(t) &= \alpha e^{\mathbf{T}t}\mathbf{e}, \quad t \geq 0, \\
h(t) &= \frac{\alpha e^{\mathbf{T}t}\mathbf{T}^0}{\alpha e^{\mathbf{T}t}\mathbf{e}}, \quad t \geq 0, \\
H(t) &= \int_0^t \frac{\alpha e^{\mathbf{T}t}\mathbf{T}^0}{\alpha e^{\mathbf{T}t}\mathbf{e}}, \quad t \geq 0.
\end{aligned}
\tag{2.2}
$$

Una definición alternativa a la dada anteriormente, puede ser la que dio Neuts por primera vez basándose en la idea de que estas distribuciones son definidas en el campo de las cadenas de Markov. Supongamos que se dispone de una cadena de Markov absorbente en tiempo continuo $\{I(t),\ t \geq 0\}$ con generador infinitesimal $\mathbf{Q}$ que será absorbida por el estado $m+1$ con probabilidad uno. Entonces, una variable aleatoria Tipo Fase X se define como el tiempo de absorción por el estado $m+1$ de la cadena de Markov en tiempo continuo, dado que la distribución inicial de la cadena de Markov es $(\alpha, 1-\alpha\mathbf{e})$. Ambas definiciones son equivalentes pero esta última es bastante útil en el modelado estocástico, ya que relaciona una variable aleatoria Tipo Fase con las cadenas de Markov en tiempo continuo. A modo de recordatorio, se aclara que un generador infinitesimal es una matriz finita/infinita para la cual todos los elementos fuera de la diagonal son no negativos, todos los elementos de la diagonal son negativos o cero y todas las filas suman cero (matriz conservativa).

En resumen, una distribución Tipo Fase se representa considerando solo los estados transitorios de la cadena de Markov asociada a través del par $(\alpha, \mathbf{T})$, siendo $\alpha = (\alpha_1, \ldots, \alpha_m)$ y $\mathbf{T} = (q_{ij})_{i,j=1,\ldots,m}$ donde q_{ij} representa la intensidad de transición del estado i al estado j y α_i la probabilidad de estar inicialmente en el estado i.

Por otro lado, un importante resultado relacionado con estas distribuciones lo formuló Asmussen en [4] en forma de teorema en el que demostró que el conjunto de distribuciones Tipo Fase es denso en el conjunto de distribuciones de probabilidad no negativas. Este teorema implica una las razones principales por la que se usan ampliamente en el modelado estocástico: las distribuciones Tipo Fase pueden aproximar cualquier distribución de probabilidad no negativa tanto como se desee. Otra razón importante es la interpretación probabilística asociada a las representaciones de las Tipo Fase y sus buenas propiedades pudiendo expresar los resultados de forma algorítmico matricial. Además, esta clase de distribuciones es cerrada bajo una serie de operaciones tales como el mínimo, máximo, suma, etc. y poseen la propiedad de falta de memoria parcial.

En la introducción del presente trabajo se ha comentado que las distribuciones Tipo Fase engloban a una serie de distribuciones conocidas como la distri-

bución Exponencial, distribución Erlang o la distribución Coxiana, entre otras. Estas distribuciones vienen precedidas porque las distribuciones Tipo Fase tienen representaciones matriciales que no son únicas, originando las distribuciones mencionadas. A modo de ejemplo, si se obtiene que el vector de probabilidades iniciales α y la matriz de transición $\mathbf{T}$ toman las siguientes expresiones:

$$\alpha = (1, 0, \ldots, 0, 0); \quad \mathbf{T} = \begin{pmatrix} -\lambda & \lambda & & \\ & -\lambda & \ddots & \\ & & \ddots & \lambda \\ & & & -\lambda \end{pmatrix}_{mxm}, \tag{2.3}$$

entonces, esta estructura es bien conocido que corresponde a una distribución Erlang con estructura Tipo Fase. En [10] se recopila una colección de las expresiones que adoptan otras distribuciones mediante estructura Tipo Fase.

Finalmente, en cuanto a la estimación de las distribuciones Tipo Fase en estudios de fiabilidad, se suele utilizar un método gráfico para estimar los parámetros de las distribuciones de probabilidad clásicas cuando las estimaciones de sus parámetros presentan serias dificultades de cálculo como puede suceder, por ejemplo, con la distribución Weibull. Esta técnica paramétrica, que está basada en el principio de mínimos cuadrados, es utilizada con bastante frecuencia debido a su sencillez y porque permite una primera idea gráfica del ajuste. En [1] se detalla metodológicamente como funciona dicha técnica. Sin embargo, las distribuciones Tipo Fase no se pueden linealizar y por lo tanto, este método queda descartado para la estimación de los parámetros involucrados en ellas. Hay que tener en todo momento claro que en el ajuste de las distribuciones Tipo Fase se trabaja con un problema de optimización difícil dado que la representación de una distribución Tipo Fase es, en general, altamente redundante. Es por ello por lo que normalmente se recurre a un algoritmo iterativo, denominado algoritmo EM, para estimar los parámetros de una distribución Tipo Fase mediante máxima verosimilitud. Este algoritmo fue desarrollado por Asmussen en [3] y asumido por Buchholz en [6] y se basa en dos pasos principalmente: esperanza y maximización, de ahí su nombre. Debido a la dificultad de la notación y puesto que no es el objetivo del presente trabajo, se invita al lector a redirigirse a las referencias mencionadas si se estuviera interesado en la formulación del mismo.

3. Aspectos computacionales

Matlab ([18]) y R ([17]) quizás sean los dos programas informático-estadísticos más empleados en los últimos tiempos en el ámbito de la investigación en el campo de la Estadística. Incluso, se podría concluir que en los últimos años R ha sido, probablemente, el programa rey en la cima de la colina, aunque también es cierto que últimamente le ha salido un duro competidor en el mercado como es

Python ([16]). Centrándonos en Matlab y especialmente en R, estos programas contienen paquetes en los que vienen implementadas funciones que trabajan con las distribuciones Tipo Fase.

En Matlab está disponible el paquete `butools.ph` (desarrollado por BuTools Team ([7])) el cual requiere que se inicie el script `BuToolsInit` para poder cargar dicho paquete. Este paquete contiene una serie de funciones básicas (momentos de una distribución tipo fase, la función de distribución, generación de números aleatorios de una Tipo Fase, etc.) y funciones algo más complejas referidas a herramientas de caracterización inversa y métodos de transformación de representaciones (distribuciones Tipo Fase de orden 2 y orden 3 en forma canónica y transformación de una representación exponencial en una representación acíclica de Tipo Fase, entre otros resultados).

Con respecto a R, existen al menos los tres siguientes paquetes dedicados a las distribuciones Tipo Fase:

- *mapfit* ([14]): implementa métodos de estimación para la distribución Tipo Fase y procesos de llegada Markovianos a partir de datos empíricos (datos puntuales y agrupados) y función de densidad.

- *PhaseType* ([5]): incluye funciones para realizar inferencia bayesiana en datos de tiempo de absorción para distribuciones Tipo Fase, con la idea de incluir también inferencia frecuentista y herramientas de simulación.

- *actuar* ([9]): ha sido publicado recientemente y contiene las famosas funciones de R encargadas de generar números aleatorios, dar la función de densidad o la función de distribución o los cuantiles para una distribución Tipo Fase con vector α y matriz de transición $\mathbf{T}$ (`mapfit` también cuenta con ellas). Es decir, las funciones análogas a las funciones *dnorm(x, mean, sd)*, *pnorm(q, mean, sd)*, *qnorm(p, mean, sd)* y *rnorm(n, mean, sd)* para el caso de la distribución Normal.

Para estimar los parámetros de las distribuciones Tipo Fase (y el valor de la log-verosimilitud), los autores de este trabajo recomiendan utilizar la función *phfit.point* del paquete `mapfit` de R o la aplicación EMpht que está disponible en la red. Ambas herramientas calculan los parámetros de las distribuciones Tipo Fase mediante el algoritmo EM, fijando de antemano el número de fases. Una vez que es fijado el número de fases, el usuario puede indicar si desea obtener una estructura Tipo Fase general para la representación $(\alpha, \mathbf{T})$ o por el contrario está interesado en obtener una estructura más específica como puede ser la correspondiente a una Erlang o Coxiana.

Existen ocasiones en las que al seleccionar una estructura Tipo Fase general, ésta converge a una estructura que le corresponde a una distribución conocida como sucede en [2]. La aplicación EMpht acepta las estructuras Tipo Fase

general, hiper-exponencial, hypo-exponencial, Coxiana y Coxiana general, mientras que `mapfit` solo cuenta con las estructuras Tipo Fase general, canónica e Hyper-erlang. La forma de presentación de los resultados obtenidos es más clara utilizando la aplicación EMpht, pero surge el problema de que si se quisieran realizar análisis complementarios, como por ejemplo realizar un gráfico, esta aplicación no dispone de la posibilidad de poder exportar la estimación obtenida a un fichero de datos que pueda ser utilizado en un programa estadístico posteriormente, teniendo el usuario que copiar a mano las estimaciones conseguidas, proceso realmente laborioso especialmente cuando el número de fases es elevado. Sirva como ilustración que si se seleccionan 3 fases, el número de parámetros estimados sería 12 (3 del vector α y 9 de la matriz $\mathbf{T}$ considerando que la distribución se encuentra inicialmente en un estado transitorio), mientras que para un número de fases fijado en 6 , el número de parámetros estimados sería 42 (6 del vector α y 36 de la matriz $\mathbf{T}$). Por este motivo, los autores recomiendan utilizar la función implementada en `mapfit` si se piensa realizar un estudio más completo y usar la aplicación EMpht simplemente para comprobar la estructura que siguen los datos experimentales.

El problema que presentan estas herramientas computaciones cuando el número de fases es elevado y se dispone de una gran cantidad de datos experimentales es que el coste computacional es demasiado elevado. Si la estructura que sigue dichos datos es desconocida, no habrá más remedio que tener paciencia y trabajar de la manera usual. Sin embargo, en muchas ocasiones es sabido (por estudios previos, por creencias, por la forma que han sido registrados, etc.) que dichos datos pueden ser ajustados mediante una distribución conocida con estructura Tipo Fase. Siguiendo con el ejemplo de las memorias resistivas, después de varios análisis exhaustivos y pormenorizados se ha concluido que el voltaje de fallo de estas memorias puede ser modelado a través de una distribución Erlang con estructura Tipo Fase ([2]; [15]). Esto permite que el usuario se pueda saltar la primera parte de la estimación referida a la estructura que siguen los datos experimentales. En esta situación, donde la estructura de $(\alpha, \mathbf{T})$ es conocida, el usuario puede elaborar un programa que estime los valores de estos parámetros de manera más eficiente que reduzca el coste computacional. En esta línea los autores del presente trabajo han elaborado un código R para el caso particular en que los datos puedan ser ajustados mediante una distribución Erlang con estructura Tipo Fase.

Como se ha comentado en el apartado anterior, la representación mostrada en (2.3) corresponde con una distribución Erlang con estructura Tipo Fase. El valor del parámetro λ es estimado mediante el cociente entre el número de fases m y la media de los valores experimentales $\overline{x}$, es decir, $\hat{\lambda} = \frac{m}{\overline{x}}$. De esta manera, es realmente sencillo obtener el vector α y la matriz $\mathbf{T}$, puesto que sería suficiente con elaborar un programa que, en función del número de fases fijado,

determine la estructura de (2.3). Para el número de fases óptimo basta con crear una función que determine el valor de la log-verosimilitud de una distribución Erlang clásica para distintos números de fases dados (esto es fijar el valor del parámetro forma de una distribución Gamma) y quedarse con el valor que mejor ajuste proporcione. Asimismo, ya se tendrían a disposición las expresiones que adoptan el vector α y la matriz $\mathbf{T}$, y a partir de ellas se podrían obtener de manera inmediata las expresiones de la función de distribución, la función de fiabilidad, etc., realizar representaciones con ellas, e incluso, realizar el test de Kolmogorov-Smirnov para comprobar si los datos experimentales se ajustan a dicha distribución.

4. Aplicación

Con el fin de mostrar el potencial de las distribuciones Tipo Fase en estudios de fiabilidad, se van a considerar datos experimentales provenientes de memorias RRAMS (resistive random access memories). Este tipo de memorias está presente en multitud de aparatos cotidianos (móviles, ordenadores, etc) gracias a sus excelentes propiedades físicas (tiempos de escritura/lectura más cortos, bajo consumo, alta retención, durabilidad,...), y son una de las fuentes de ingresos más importantes a nivel mundial en la industria de los semiconductores. El funcionamiento de las memorias RRAM está basado en los procesos de conmutación resistiva, que en la mayoría de los casos, crean y rompen un filamento conductor que provoca que la resistencia del aparato cambie drásticamente. Estos procesos de creación y ruptura del filamento conductor se conocen en los ámbitos de la investigación e industrial como procesos set y reset, respectivamente. Los cambios de resistencia provocan una muestra de curvas de tensión-intensidad correspondientes a los ciclos set y reset. Las curvas set/reset se caracterizan por la evolución del voltaje frente a la intensidad hasta que en un determinado punto (lo que se conoce como punto set/reset) se produce una subida/caída repentina de intensidad. Estos puntos set/reset son distintos en cada ciclo, lo que provoca diferentes intensidades set/reset y voltajes set/reset. A modo de ejemplo, nos centraremos únicamente en los ciclos reset. En la Figura 1 se muestran algunas curvas reset del estudio.

En esta área es de suma importancia ajustar una distribución a las intensidades o voltajes set/reset con el objetivo de estudiar el comportamiento de estos dispositivos. Por lo tanto, previo paso al análisis, se han obtenido todos los V_{reset} de todas las curvas reset de los 300 ciclos considerados. Una vez obtenidos estos puntos, el primer paso ha sido realizar el análisis estadístico habitual basado en la distribución Weibull que se aplica en la rama de la ingeniería cuando se está trabajando con datos experimentales correspondientes a memorias RRAM. En consecuencia, se han obtenido los Weibits calculados como $ln(-ln(1-F(t)))$ y se han representado frente a los $ln(V_{reset})$. Si el ajuste por la distribución

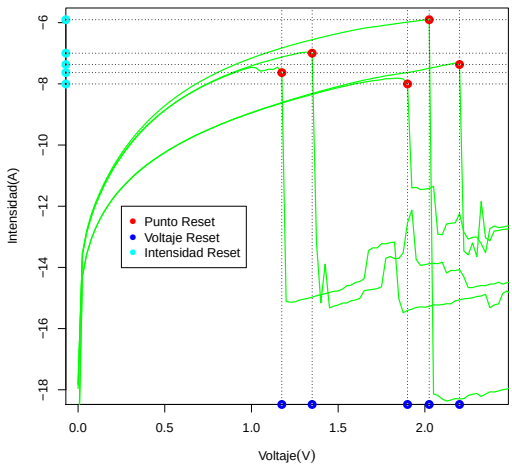


Figura 1: Corriente reset experimental frente a la tensión aplicada (mostrada en líneas verdes) para cinco curvas reset de una serie de 300 ciclos continuos. Los puntos reset, las tensiones reset correspondientes y las corrientes reset se muestran para mayor claridad.

Weibull fuera adecuado, los datos experimentales deberían seguir la línea correspondiente al ajuste, en el cual la pendiente corresponde al parámetro de forma β (β mide la dispersión estadística) y $V_{63\,\%}$ es el inverso del parámetro escala. Los resultados del ajuste se muestran en la Figura 2.

Como se observa en la Figura 2, los Weibits de los datos experimentales no son lineales. Por tanto, aunque se podría realizar una aproximación más o menos precisa bajo el contexto de la distribución Weibull, parece razonable intentar ajustar otra distribución, en cuyo caso se utilizarán las distribuciones Tipo Fase.

Después de un análisis progresivo basado en la estimación paso a paso de las distribuciones Tipo Fase, se ha concluido que la distribución Erlang proporciona un buen ajuste para la variable V_{reset}. Asimismo, se ha obtenido que el número de fases óptimo es 210, por lo que $\lambda = 203{,}415$. Una vez estimados los parámetros de la distribución Erlang con estructura Tipo Fase, se procede a comparar gráficamente la precisión del ajuste de esta distribución sobre los datos experimentales medidos. La tasa de fallo acumulada experimental estimada por las distribuciones Erlang y Weibull se muestran y se comparan en la Figura 3. A tenor de lo que muestra la gráfica, el mejor resultado se logra cuando se considera la distribución Erlang, obteniéndose un ajuste preciso que explica en gran medida los datos experimentales.

Por otro lado, la función de fiabilidad o la función de supervivencia, como se

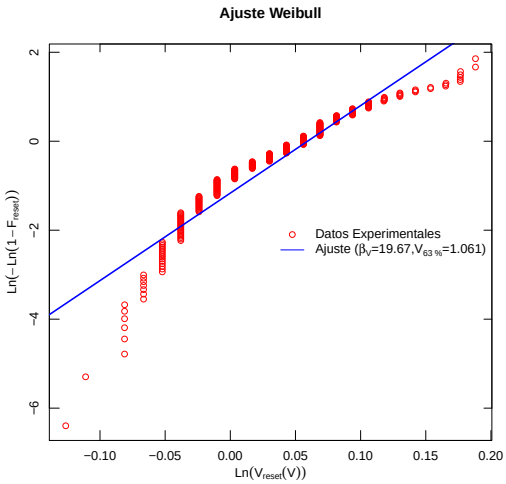


Figura 2: Ajuste lineal de la distribución Weibull para los V_{reset} de los 300 ciclos.

Tabla 1: Valor del estadístico y del p-valor asociado del test de Kolmogorov-Smirnov al considerar la distribuciones Weibull y Tipo Fase.

	Estadístico	**P.valor**
Weibull	0.13329	<0.001
Tipo Fase	0.070634	0.1002

la conoce en ramas científicas no relacionadas con la ingeniería, principalmente en el campo de la biomedicina, es interesante para analizar las propiedades estadísticas de los datos con los que estamos tratando. La función de fiabilidad describe la probabilidad de que el filamento conductivo no se rompa para tensiones más pequeñas que el voltaje de fallo. La función de fiabilidad ha sido representada en la Figura 4 para los datos experimentales, además del ajuste por Weibull y Tipo Fase. Aunque ninguna distribución muestra una reproducción cercana de los valores experimentales, la distribución Tipo Fase funciona mejor que la distribución Weibull y logra un rendimiento razonablemente bueno.

En consecuencia, podemos concluir que las distribuciones Tipo Fase, en particular, la distribución Erlang proporciona un ajuste más aproximado a los datos reales que la distribución Weibull. Finalmente se comprueba si estas distribuciones son aceptadas para los datos experimentales considerados. En la Tabla 1 se muestra el estadístico y el p-valor asociado del test de Kolmogorov-Smirnov para cada distribución. Se aprecia que la única distribución que puede ser considerada es la distribución Tipo Fase.

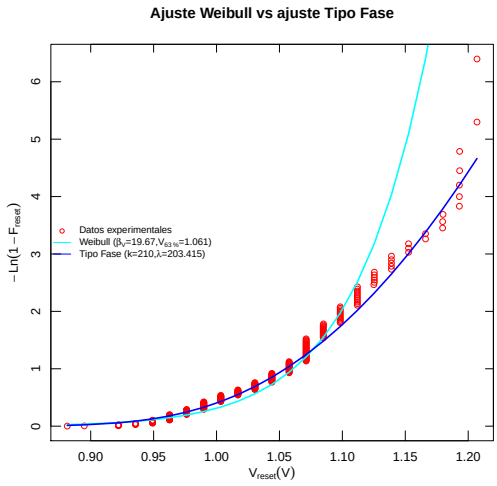


Figura 3: Tasa de fallo acumulada de los V_{reset} para los 300 ciclos y el correspondiente ajuste de las distribuciones Weibull y Erlang con estructura Tipo Fase (PHD).

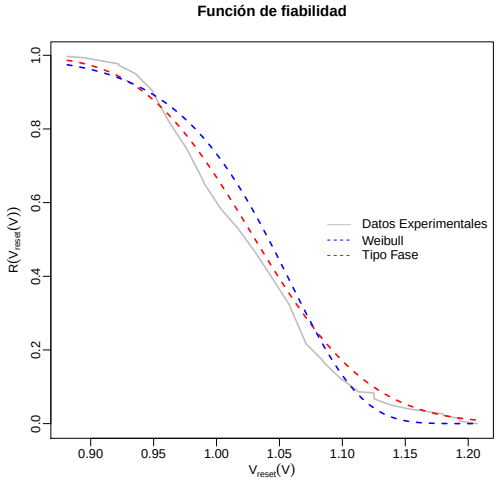


Figura 4: Función de fiabilidad de los V_{reset} para los 300 ciclos y el correspondiente ajuste de las distribuciones Weibull y Erlang con estructura Tipo Fase (PHD).

5. Conclusiones

El objetivo principal que persiguen los estudios de fiabilidad es analizar el comportamiento de sistemas mediante la modelización de los tiempos de vida (análogamente, los tiempos de fallo) o de otros valores relacionados directamente con el tiempo. En el ámbito de la docencia y en la mayoría de aplicaciones se suelen desarrollar metodologías que permiten un estudio detallado del funcionamiento de los sistemas a través del empleo de técnicas estadísticas clásicas basadas en el uso de las distribuciones de probabilidad clásicas (Weibull, Exponencial, Log-Normal, etc.). Sin embargo, en la vida cotidiana se plantean situaciones en las que estas distribuciones no se ajustan adecuadamente a los datos reales. Ante esta situación, se debe utilizar un enfoque diferente que solucione la falta de ajuste. Desde un punto de vista teórico y práctico, existen un tipo de distribuciones denominadas Tipo Fase con unas excelentes propiedades algebraicas que generalizan otras distribuciones clásicas, y facilitan el desarrollo metodológico mejorando considerablemente el ajuste que se obtiene al considerar las distribuciones clásicas de probabilidad.

La complejidad que presentan las distribuciones Tipo Fase se debe principalmente al desconocimiento que hay sobre ellas, que sumado a que el proceso de ajuste de las mismas es un problema de optimización difícil, dado que la representación de una distribución Tipo Fase es altamente redundante, hace que disponer de herramientas que permitan su estimación en fiabilidad sea de principal interés para el usuario.

En este trabajo se muestra una pequeña introducción a las distribuciones Tipo Fase, así como la bibliografía fundamental referida a ellas, y se revisan una serie de programas estadísticos (R, Matlab y la aplicación EMpht) que permiten trabajar directamente con ellas sin necesidad de disponer de unos amplios conocimientos sobre el tema. Por último, se realiza una aplicación a un conjunto de datos reales de memorias resistivas con el fin de demostrar el potencial que tienen las distribuciones Tipo Fase frente al análisis estadístico habitual empleado en esta área.

Agradecimientos

Este trabajo ha sido financiado por la Junta de Andalucía (grupo FQM-307) y por el Ministerio de Ciencia, Innovación y Universidad a través del proyecto MTM2017-87708-P (fondos FEDER incluidos). El trabajo de Christian Acal ha sido también subvencionado por la beca de doctorado FPU18/01779.

Referencias

[1] Acal C., Ruiz-Castro J.E. y Aguilera A.M. (2019a). Distribuciones tipo fase en un estudio de fiabilidad. *TEMat*, **3**, 63-74.

[2] Acal C., Ruiz-Castro J.E., Aguilera A.M., et al. (2019b). Phase-type distributions for studying variability in resistive memories. *J. Comput. Appl. Math.*, **345**, 23-32.

[3] Asmussen S., Nerman O. y Olsson M. (1996). Fitting phase-type distributions via the EM algorithm. *Scand. J. Stat.*, **23**(4), 419-441.

[4] Asmussen S. (2000). *Ruin Probabilities*, World Scientific, Hong Kong (Chinese).

[5] Aslett L. (2012). Package "PhaseType: Inference for Phase-type Distributions". En: https://cran.r-project.org/web/packages/PhaseType/

[6] Buchholz P., Kriege J. y Felko I. (2014). *Input modeling with phase-type distributions and Markov models, Theory and Applications*. Springer Cham Heidelberg New York Dordrecht London.

[7] BuTools Team. (2015). Tools for Phase-Type Distributions. En: http://webspn.hit.bme.hu/˜telek/tools/butools/doc/ph.html

[8] Epstein B. y Sobel M. (1953). Life Testing. *J. Am. Stat. Assoc.*, **48**(263), 486-502.

[9] Goulet V., Auclair S., Dutang C., et al. (2019). Package 'actuar: Actuarial Functions and Heavy Tailed Distributions'. En: https://cran.r-project.org/web/packages/actuar/

[10] He Q.M. (2014). *Fundamentals of Matrix-Analytic Methods*, Springer Science+Business Media, New York (EEUU).

[11] Long S., Cagli C., Ielmini D., et al. (2012). Analysis and modelling of resistive switching statistics. *J. Appl. Phys.*, **111**(7), 074508.

[12] Neuts M. F. (1975). *Probability Distributions of Phase Type*, Liber Amicorum Professor Emeritus Dr. H. Florin.

[13] Neuts M. F. (1994). *Matrix-Geometric Solutions in Stochastic Models: an Algorithmic Approach*, Courier Corporation.

[14] Okamura H. (2015). Package "mapfit: A Tool for PH/MAP Parameter Estimation". En: https://cran.r-project.org/web/packages/mapfit/

[15] Pérez E., Maldonado D., Acal C., et al. (2019). Analysis of the statistics of device-to-device and cycle-to-cycle variability in TiN/Ti/Al:HfO2/TiN RRAMs. *Microelectron. Eng.*, **214**, 104-109.

[16] The PSF (2019). *Python Software*, The Python Software Foundation (PSF). URL https://www.Python.org/.

[17] R Core Team (2019). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing. URL https://www.R-project.org/.

[18] The Math Works (2020). *MATLAB*, Natick, MA: The Math Works, Inc. URL www.mathworks.com

Acerca de los autores

Juan Eloy Ruiz Castro es Profesor Titular del Departamento de Estadística e I.O. de la Universidad de Granada y miembro del grupo de investigación Modelización y Predicción con Datos Funcionales de la Junta de Andalucía (FQM-307). Obtuvo la licenciatura y el doctorado en Ciencias Matemáticas por la Universidad de Granada y sus principales líneas de investigación son, análisis de datos de tiempo de vida con aspectos teóricos y aplicados en fiabilidad y supervivencia. Fruto de su dedicación docente e investigadora es la participación y dirección de proyectos financiados, autoría de artículos de investigación en revistas de alto impacto, libros docentes, y múltiples aportaciones en congresos nacionales e internacionales relevantes. Su actividad editorial es muy activa siendo en la actualidad editor asociado de diferentes revistas indexadas en Journal Citation Reports y revisor permanente en múltiples de ellas. Además, es miembro del Grupo de Trabajo GENAEIO de la SEIO, del International Group on Reliability (Gnedenko e-Forum), miembro del grupo Survival Analyisis del ERCIM WG Computational and Methodological Statistics y miembro del Instituto de Matemáticas de la UGR (IMAG).

Christian J. Acal González es estudiante de doctorado en el Programa de Doctorado en Estadística Matemática y Aplicada en el Departamento de Estadística e Investigación Operativa de la Universidad de Granada. Actualmente disfruta de una beca para la Formación de Profesorado Universitario (FPU) otorgada por el Ministerio de Ciencia, Innovación y Universidad de España y es miembro del grupo de investigación FQM-307 de la Junta de Andalucía y de los Grupos de Trabajo FDA y GENAEIO de la SEIO y del Instituto de Matemáticas de la UGR (IMAG). Fue el mejor expediente de la promoción 2013-2017 del Grado en Estadística en la Universidad de Granada y recibió el XII Premio Andaluz a la Mejor Trayectoria Académica en el Ámbito de la Estadística otorgado por la Fundación Bancaria Unicaja y la Academia de C.C. Sociales del Medio Ambiente de Andalucía. Su principal línea de investigación es el análisis de datos funcionales y sus aplicaciones en diversas áreas del conocimiento, aunque también trabaja habitualmente en temas de fiabilidad de sistemas.

Ana M. Aguilera del Pino es Catedrática del Departamento de Estadística e I.O. de la Universidad de Granada y miembro del grupo de investigación FQM-

307: Modelización y Predicción con Datos Funcionales de la Junta de Andalucía. Licenciada y Doctora en Ciencias Matemáticas, sus principales líneas de investigación son el análisis de datos funcionales, el análisis de datos categóricos y sus aplicaciones en diversas áreas como la economía, el medioambiente, las ciencias de la salud y la ingeniería. Fruto de su dedicación docente e investigadora es la coordinación de proyectos financiados, la autoría de artículos en revistas de alto impacto, además de libros sobre cursos de probabilidad, datos categóricos y modelización funcional, y las aportaciones a congresos relevantes. Realiza también una intensa actividad editorial siendo actualmente Co-editora Jefe de la revista Test de la SEIO y Editora Asociada de Computational Statistics. Además, es miembro activo de los Grupos de Trabajo FDA y GENAEIO de la SEIO, del grupo Statistics for Functional Data del ERCIM Working Group Computing and Statistics, y pertenece al Consejo de Investigadores y la Comisión de Relaciones Externas del IEMath-Granada.

Boletín de Estadística e Investigación Operativa
Vol. 37, No. 1, Marzo 2021, pp. 19-37

Investigación Operativa

Ranking of the Academic Departments of the Autonomous University of Aguascalientes

Luz Judith R. Esparza

Departamento de Matemáticas y Física
Cátedra CONACyT-Universidad Autónoma de Aguascalientes (México)
✉ judithr19@gmail.com

Julio C. Macías Ponce, Roberto A. Kú Carrillo, Sandra E. Delgadillo Alemán and Arturo E. Giles Flores

Departamento de Matemáticas y Física
Universidad Autónoma de Aguascalientes
✉ jlmacias@correo.uaa.mx, ✉ jraku@correo.uaa.mx,
✉ sedelgad@correo.uaa.mx, ✉ arturo.giles@edu.uaa.mx

Abstract

In this work, a ranking of the academic departments of the Autonomous University of Aguascalientes is carried out. A network is built that describes the interaction between departments considering the courses taught in the undergraduate educational programs. Once the network is generated, three ranking algorithms are applied: conservative ranking in graphs, weighted ranking in digraphs, and the analytical hierarchical process. The departments of Mathematics and Physics, Philosophy, Statistics and Law are the best ranked according to a dominance criterion. We provide an analysis by academic units of the teaching balance in terms of the ranking and the differences when compared to global results. This research could be useful for decision-making aimed at institutional development.

Keywords: Ranking, Digraph, Analytical Hierarchical Process, Dominance Analysis.
AMS Subject classifications: 62F07, 90C35, 91A12.

1. Introducción

La Benemérita Universidad Autónoma de Aguascalientes, creada el 19 de junio de 1973, tiene sus orígenes en la Escuela de Agricultura, fundada por el gobernador J. Jesús Gómez Portugal el 15 de enero de 1867[1]. Esta universidad funciona como un organismo público descentralizado del Estado con personalidad jurídica propia para adquirir y administrar bienes. Tiene por fines impartir la enseñanza media superior y superior en el Estado de Aguascalientes, realizar la investigación científica y humanística, y extender los beneficios de la cultura a los diversos sectores de la población.

La dinámica académica de la Universidad Autónoma de Aguascalientes (UAA) es muy particular. Una característica distintiva de esta universidad es su sistema departamental, en donde los profesores se encuentran adscritos a departamentos y estos a su vez forman parte de los centros académicos. Con la finalidad de ilustrar esto, su organigrama se puede consultar en la página web https://www.uaa.mx/portal/wp-content/uploads/2018/10/CO-010000-01-2-1.pdf. En este modelo organizacional, un departamento imparte todas las materias relacionadas con su área académica a todos los programas que lo requieran. Por ejemplo, el departamento de Filosofía imparte la asignatura de ética o de esta área tanto a ingenieros, economistas, administradores de empresas e incluso a matemáticos, si su plan de estudios tiene un curso de esta área. Podemos mencionar como ventajas de este sistema, que los catedráticos son especialistas en sus respectivas áreas y se optimiza la contratación de los recursos humanos. Por otro lado, este sistema plantea cuestiones interesantes para analizar, como la interdependencia entre los distintos departamentos. Además, surgen preguntas interesantes como: ¿Qué departamento está más o menos interconectado? ¿Cómo podemos clasificarlos de acuerdo a su impacto en un programa? Y si podría servir esta clasificación como criterio para balancear los presupuestos, plazas asignadas o carga administrativa.

Una manera de contestar estos interrogantes es usar técnicas de ordenamiento. Dado un conjunto finito de agentes (empresas, instituciones, inversionistas, deportistas) se presenta con mucha frecuencia la necesidad de ordenarlos, o equivalentemente listar a los agentes en orden monótono para identificar desde el mejor hasta el peor. Los ordenamientos usan como input información cuantitativa relacionada con el rendimiento de los agentes, el rendimiento puede ser individual o por interacción con los diferentes subconjuntos a los que pertenece cada agente. Con frecuencia el valor de Shapley [2] y otras soluciones de los juegos cooperativos son usadas para distribuir costos o beneficios, pero al mismo tiempo se están definiendo índices de poder de cada uno de los agentes. A su vez, un caso particular pero de gran interés matemático y económico es el de

[1] https://www.uaa.mx/nu/historia.php

ordenar los nodos en una red [10]. Una motivación para definir y estudiar este problema, se basa en identificar que la influencia de un individuo -en una red social por ejemplo- depende de la calidad y cantidad de sus contactos sociales [5]. El ordenamiento de nodos en una red es el principal interés de nuestro trabajo.

Por su parte, en estadística, un problema no supervisado es la agrupación en clústeres (o conglomerados) [7]. Éste se refiere a un conjunto muy amplio de técnicas para encontrar subgrupos en un conjunto de datos. Cuando agrupamos observaciones, tratamos de dividirlas en subgrupos distintos para que las observaciones dentro de cada subgrupo sean bastante similares entre sí, mientras que las observaciones en diferentes grupos son muy diferentes [1]. Es decir, la agrupación en conglomerados busca encontrar subgrupos homogéneos entre las observaciones. Lo cual ha sido usado por ejemplo en minería de texto y redes neuronales [6]. En este artículo, una vez obtenidos los ordenamientos, se utilizará esta técnica estadística para agrupar a los departamentos y poder identificar aquellos subgrupos de departamentos con características similares. Más aún, se utilizará un criterio de dominancia para calificar la robustez de los ordenamientos obtenidos por diferentes métodos [12]. Este tipo de criterios se han usado exitosamente en distintos problemas que abordan la comparación entre distintas clasificaciones [4, 18].

El objetivo principal de este artículo es aplicar técnicas de clasificación basadas en la teoría de juegos cooperativos y análisis estadístico de datos, para identificar a los departamentos con mayor impacto en los programas educativos de la UAA, considerando la cantidad de *cursos impartidos* como la variable de clasificación. Las técnicas usadas provienen de distintas áreas que se adecuan a la estructura del problema, buscando tener redundancia y soporte en los resultados obtenidos. En nuestro caso, la interacción departamental da lugar, de manera natural, a una red cuya construcción es inédita y considera la manera tan particular en la que se estructura la UAA. Este trabajo tiene como finalidad proveer de información a las autoridades universitarias para la toma de decisiones imparciales y relevantes, como la asignación equitativa de recursos humanos, presupuestales, etc.

Este artículo está organizado de la siguiente manera: en la Sección 2, se presentan los antecedentes de las técnicas de ordenamiento que se aplicarán en este estudio. En la Sección 3 se describe el problema de investigación. Mientras que en la Sección 4, se muestra la metodología usada para obtener los resultados que se presentan y discuten en la Sección 5. Los comentarios finales se encuentran en la Sección 6.

2. Antecedentes

El ordenamiento es ampliamente utilizado para clasificar a los elementos o agentes de un conjunto. En particular, una competición que se basa en los

resultados de comparaciones 2 a 2 se puede modelar a través de gráficas dirigidas. En ellas, los pesos nos sirven para comparar nodos o agentes, y dados los pesos iniciales de los nodos (personas, agentes, departamentos, etc.), se puede medir su importancia a través de teoría de juegos cooperativos [10]. En esta teoría podemos identificar el valor de Shapley como un ordenamiento natural, ya que el valor asignado a cada jugador (agente) es un valor promedio de las utilidades marginales que el jugador aporta a cada uno de los subconjuntos [17]. En las redes o gráficas dirigidas también aparece de manera natural el problema de listar los nodos de acuerdo a su interacción -importancia- en la estructura de la red correspondiente. En el contexto de las redes, un ordenamiento que aparece en la literatura y que se considera acorde al propósito de este artículo es el ordenamiento conservativo, llamado así porque se basa en aplicar el valor de Shapley al juego cooperativo conservativo inducido por la estructura de la red. Este juego conservativo se construye asignándole a cada subconjunto de nodos el conteo de todos los agentes, para los cuales se satisface que el conjunto de nodos predecesores junto con el nodo en cuestión esté contenido en la coalición. Una vez formado el juego se aplica el valor de Shapley. Recordemos que los juegos cooperativos son aquellos en los que dos o más jugadores forman coaliciones para conseguir un objetivo, se analizan las estrategias óptimas para grupos de individuos, asumiendo que pueden establecer acuerdos entre sí acerca de las estrategias más apropiadas[2]. Así pues, la clasificación -ordenamiento- de agentes, desde hace ya muchos años, viene siendo un tema muy importante.

En [3] con datos de la liga Premier 97-98 y del mundial de fútbol FIFA 1998, se ilustra una aplicación de este ordenamiento conservativo. En particular el ordenamiento conservativo involucra las incidencias entre los nodos -que en nuestro contexto sería el apoyo de impartir cursos entre los departamentos académicos-. En este ordenamiento se priorizan las direcciones de los arcos de los nodos en el siguiente sentido, es más importante para un nodo no ser incidido por otros nodos que incidir sobre ellos. Así pues, se estará ponderando mejor a un departamento que no tenga asignados cursos de otro departamento.

Sin embargo, existe una versión subjetiva de la clasificación, dada a través del Proceso Jerárquico Analítico, adicionalmente a las técnicas cuantitativas ya mencionadas. El Proceso Jerárquico Analítico (PJA), introducido por Thomas Saaty en 1980 [13], es una herramienta eficaz para tomar decisiones complejas, y puede ayudar al tomador de decisiones a establecer prioridades y elegir la mejor decisión al reducir las decisiones complejas a una serie de comparaciones por pares. El PJA ayuda a captar aspectos tanto subjetivos como objetivos de una decisión. Podemos mencionar su existosa aplicación en las siguientes referencias [8, 9, 11].

Operacionalmente ayuda a construir índices, reduciendo la complejidad a un

[2] https://economipedia.com/definiciones/teoria-de-juegos.html

esquema jerárquico simple. El proceso requiere que quien toma las decisiones, proporcione evaluaciones subjetivas respecto a la importancia relativa de cada uno de los criterios, y después especifique su preferencia con respecto a cada una de las alternativas de decisión y para cada criterio [15]. El PJA genera un peso para cada criterio de evaluación de acuerdo con la decisión del tomador de decisiones, el especialista. Se realizan comparaciones por pares de los criterios, cuanto mayor sea el peso, más importante será el correspondiente criterio. El objetivo de esta ponderación es llegar a expresar, en términos cuantitativos, la importancia de los distintos elementos. Asimismo, si bien es frecuente asignar pesos a los criterios, la especificación de éstos es una cuestión en la que no existe un método generalmente aceptado para su determinación, pudiéndose considerar este proceso como un aspecto que puede crear controversias acerca de la asignación de dichos pesos.

Por otro lado, siendo que la clasificación ha recobrado gran importancia en la investigación científica en las últimas décadas, otra vertiente de su estudio es a través del análisis de conglomerados, proporcionando técnicas estadísticas para realizarla. Existe una gran cantidad de métodos de agrupamiento, los dos más conocidos son el agrupamiento K-medias y el agrupamiento jerárquico, con sus ventajas y desventajas, pero ambos con el objetivo de clasificar individuos (que en nuestro caso serán los departamentos académicos) en grupos homogéneos. En particular, en el agrupamiento jerárquico se hace uso de una representación visual en forma de árbol de las observaciones, llamada dendograma. Éste nos permite visualizar los agrupamientos obtenidos para cada posible número de agrupaciones.

A continuación describimos con mayor detalle el problema al que aplicaremos estas técnicas de clasificación.

3. Descripción del Problema

La UAA cuenta actualmente con 10 centros académicos, donde 9 atienden a los programas de licenciatura o pregrado, y el décimo, se encarga del programa de enseñanza media superior. En este trabajo nos restringimos a los centros de nivel pregrado. Dichos 9 centros académicos agrupan a un total de 54 departamentos, cuya función principal radica en impartir 3,048 cursos[3] acorde con los planes de estudio de los 63 programas educativos de pregrado que actualmente oferta la institución. Los departamentos de los centros académicos, así como el número de programas educativos adscritos a cada departamento se pueden ver en la Tabla 1. También en esta tabla se muestra el número de asignaturas que cada departamento tiene a su cargo en los planes de estudio de dichos programas educativos de la universidad.

En la Figura 1, se muestra la representación porcentual de esta información.

[3]Dato recuperado el día 16 de junio 2020 a través www.uaa.mx.

Tabla 1: Centros y departamentos de la UAA con programas educativos y cursos.

CC. Agropecuarias (3 P.E.)	#A	**CC. de la Ingeniería** (6 P.E.)	#A
1. C. Agronómicas (1)	27	29. Ingeniería Automotriz (2)	51
2. C. de los Alimentos (1)	22	30. Ingeniería Biomédica (2)	38
3. C. Veterinarias (1)	40	31. Ingeniería Robótica (2)	54
CC. Básicas (10 P.E.)		**CC. Sociales y Humanidades** (12 P.E.)	
4. Biología (1)	34	32. Ciencias Políticas y Admon. Pública (1)	35
5. Ciencias de la Computación (1)	39	33. Comunicación (2)	76
6. Estadística (1)	95	34. Derecho (1)	89
7. Fisiología y Farmacología	29	35. Educación (1)	64
8. Ing. Bioquímica (1)	45	36. Filosofía (1)	124
9. Matemáticas y Física (1)	159	37. Historia (1)	52
10. Microbiología	17	38. Idiomas (2)	68
11. Morfología	23	39. Psicología (1)	64
12. Química (2)	86	40. Sociología (1)	65
13. Sistemas de Información (1)	63	41. Trabajo Social (1)	29
14. Sistemas Electrónicos (2)	108		
CC. de Diseño y de la Constr. (7 P.E.)		**CC. Económicas y Adtvas.** (9 P.E.)	
15. Arquitectura (1)	24	42. Administración (1)	81
16. Diseño de Interiores (1)	25	43. Contaduría (1)	60
17. Diseño de Moda (1)	35	44. Economía (2)	92
18. Diseño Gráfico (1)	74	45. Finanzas (1)	68
19. Diseño Industrial (1)	27	46. Mercadotecnia (1)	59
20. Ingeniería Civil (1)	64	47. Recursos Humanos (2)	100
21. Urbanismo (1)	58	48. Turismo (1)	23
CC. de la Salud (7 P.E.)		**CC. Empresariales** (4 P.E.)	
22. Cultura Física y Salud Pública (1)	38	49. Agronegocios (2)	46
23. Enfermería (1)	26	50. Comercio Electrónico (2)	50
24. Estomatología (1)	50	**C. de las Artes y la Cultura** (5 P.E.)	
25. Medicina (1)	87	51. Artes Escénicas y Audiovisuales (2)	85
26. Nutrición (1)	47	52. Arte y Gestón Cultural (1)	41
27. Optometría (1)	30	53. Letras (1)	79
28. Terapia Física (1)	25	54. Música (1)	58

La abreviatura C.C. significa Centro de Ciencias, mientras que P.E. significa Programa Educativo y #A indica el número de cursos adscritos a cada departamento. Entre paréntesis se encuentra el número de programas educativos de cada departamento. **Fuente:** Elaboración propia a partir datos recuperados de www.uaa.mx/portal/oferta_educativa/licenciaturas/

Observe que los departamentos de Matemáticas y Física, Filosofía, Sistemas Electrónicos, Recursos Humanos y Estadística, en términos porcentuales, son los

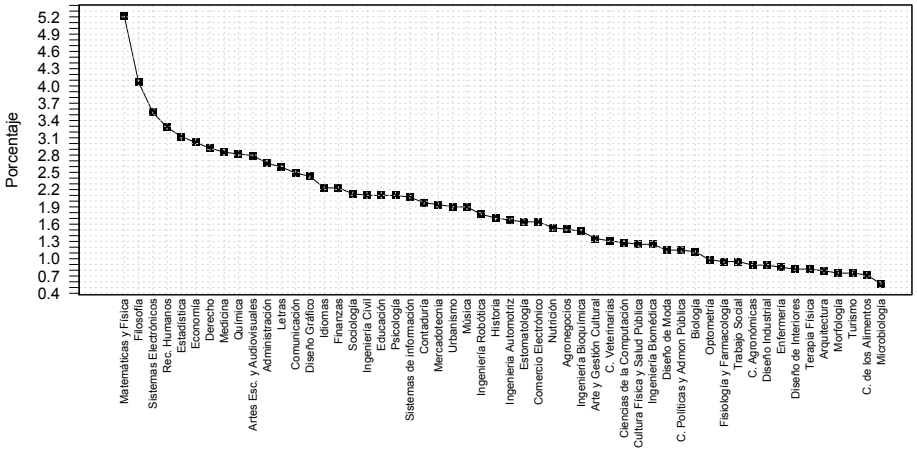


Figura 1: Porcentaje del total de cursos que se imparten en los programas educativos de pregrado.

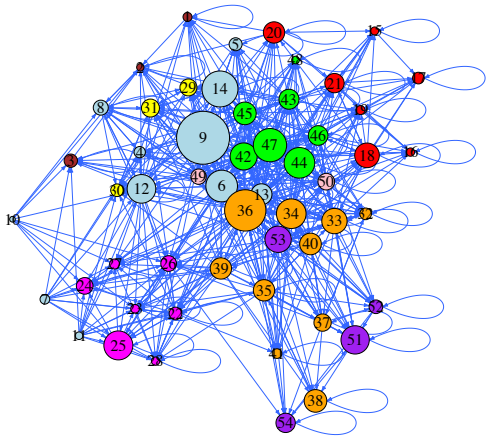


Figura 2: Red de interconexión entre los departamentos considerando los cursos impartidos a nivel pregrado en la UAA. El tamaño de los nodos (departamentos) es proporcional al número de cursos impartidos y se agrupan en 9 colores diferentes, de acuerdo al centro al que pertenecen.

departamentos que mayor número de cursos imparten en los planes de estudio de pregrado en la universidad. Ciertamente, la Figura 1 nos muestra el orden más básico al presentar los departamentos que más cursos imparten. Sin embargo, técnicas de ordenamiento más sofisticadas nos pueden permitir considerar la interrelación entre los departamentos o su importancia relativa como un nodo en la red departamental de la UAA, ver Figura 2. Esto se mostrará en la siguiente sección.

4. Metodología

En esta sección, se describen tres métodos de ordenamiento: el ordenamiento conservativo en grafos, el ordenamiento con peso en digráficas y el ordenamiento a través del Proceso Jerárquico Analítico. Esto además de un análisis de conglomerados y la aplicación de un criterio de dominancia a los métodos de ordenamiento utilizados.

4.1. Digráficas conservativas sin pesos (SP)

Una gráfica dirigida o digráfica es un par (N, D), donde N es un conjunto finito de nodos y $D \subset N \times N$ es una relación binaria sobre N, representando el conjunto de arcos dirigidos [3]. El arco dirigido del nodo i al nodo j, (i, j), representa que el nodo i implica al j (que ha tenido un mejor desempeño el nodo i que el nodo j). Supongamos que la relación binaria D es irreflexiva, es decir, $(i, i) \notin D$, para todo $i \in N$. Es decir, no admite comparaciones de los agentes

consigo mismos.

Sea $\mathcal{D}$ la colección de digráficas irreflexivas. Luego, para $i \in N$, llamamos los sucesores de i al conjunto $S_D(i) := \{j \in N : (i,j) \in D\}$ y los predecesores de i están dados por $P_D(i) := \{j \in N : (j,i) \in D\}$. Aplicando el valor de Shapley al juego cooperativo (N, v_D^C) con $v_D^C : 2^N \to \mathbb{R}$ dado por $v_D^C(E) =\mid \{j \in N \mid (P_D(j) \cup \{j\}) \subset E\} \mid$ para todo $E \subset N$, obtenemos el ordenamiento conservativo dado por:

$$\beta_i(D) = \sum_{j \in \{i\} \cup S_D(i)} \frac{1}{|P_D(j)| + 1}; \quad D \in \mathcal{D}, \tag{4.1}$$

donde $|P_D(j)|$ denota la cardinalidad del conjunto $P_D(j)$. Note que las $\beta_i(D)$ dependen del número de predecesores del nodo i en la estructura de la diagráfica. En [3] podemos ver más detalles del proceso mediante el cual se obtiene $\beta_i(D)$.

4.2. Digráficas con peso (CP)

Una digráfica Con Peso en N es una función $\omega : N \times N \to \mathbb{R}_+$, y siguiendo [3] asumimos que $\omega(i,i) > 0$ para todo $i \in N$. Este peso cuantifica la ventaja obtenida por el nodo i sobre el j representado por el arco (i,j). Sea $\mathcal{W}^N$ la colección de todas las digráficas con peso en N. Luego, tenemos que el peso para este tipo de digráficas está dado por la siguiente expresión:

$$\beta_i(\omega) = \sum_{j \in N} \left(\frac{\omega(i,j)}{\sum_{h \in N} \omega(h,j)} \right); \quad \text{para todo } i \in N, \omega \in \mathcal{W}^N. \tag{4.2}$$

En nuestro caso, las comparaciones de los departamentos se harán considerando que existe un mejor desempeño de uno sobre otro al impartir más cursos y su diferencia será el peso asignado inicialmente al nodo.

4.3. Proceso Jerárquico Analítico

En el PJA se establece una matriz de comparación entre pares de departamentos, identificando la importancia de cada uno de ellos con los demás. Posteriormente, se determina el autovector principal, el cual establece los pesos, que a su vez proporciona una medida cuantitativa de la consistencia de los juicios de valor entre pares de factores [13].

En general, sea $\boldsymbol{A}$ la matriz de comparación de dimensión $n \times n$. Cada entrada a_{jk} de la matriz $\boldsymbol{A}$ representa la importancia del j-ésimo departamento relativo al k-ésimo departamento. Si $a_{jk} > 1$, entonces el j-ésimo departamento es más importante que el k-ésimo departamento, mientras que si $a_{jk} < 1$, entonces el j-ésimo departamento es menos importante que el k-ésimo departamento. Si dos departamentos tienen la misma importancia, entonces $a_{jk} = 1$. Se debe satisfacer la condición $a_{jk}a_{kj} = 1$ y obviamente, $a_{jj} = 1$ para todo j. En nuestro caso, si $\boldsymbol{C}$ es la matriz que contiene todos los cursos de los 54 departamentos, es decir, $\boldsymbol{C}$ es una matriz de dimensión 54×54 donde la entrada (c_{ij}) es el número

de cursos que imparte el departamento i al j. Así pues, la matriz $\boldsymbol{A} = (a_{ij})$ se construirá de la siguiente manera: si $c_{ij} \geq c_{ji}$, entonces $a_{ij} := c_{ij}$ y $a_{ji} := 1/c_{ij}$, si $c_{ij} < c_{ji}$, entonces $a_{ji} := c_{ji}$ y $a_{ij} := 1/c_{ji}$. Finalmente $a_{ii} := 1$ para todo i. Definida la matriz de comparaciones, se normaliza por columnas y se promedia por filas para obtener el vector de pesos, también conocido como autovector principal [14]. Este vector muestra las ponderaciones de cada departamento y con esto el ordenamiento PJA.

4.4. Análisis de conglomerados

En el análisis de conglomerados, el dendograma de agrupamiento jerárquico se obtiene a través de un algoritmo simple. Se comienza por definir algún tipo de medida de disimilitud entre cada par de observaciones, por ejemplo, la distancia euclidiana. El algoritmo procede iterativamente comenzando en la parte inferior del dendograma, cada una de las n observaciones (departamentos) es tratada como su propio grupo. Los dos grupos que son más similares entre sí se fusionan para que ahora haya $n - 1$ agrupaciones. A continuación, los dos grupos que son más similares entre sí se fusionan de nuevo, de modo que ahora hay $n - 2$ grupos. El algoritmo prosigue de esta manera hasta que todas las observaciones pertenezcan a un solo grupo, y el dendograma se completa.

El concepto de disimilitud entre un par de observaciones se extiende a un par de grupos de observaciones. Esta extensión se logra desarrollando la noción de enlace, que define la disimilitud entre dos grupos de observaciones. Los cuatro tipos de enlace más comunes son: completo, promedio, simple y centroide. En este trabajo se utilizará el promedio. En éste se calculan todas las diferencias entre pares entre las observaciones de un grupo y las observaciones en otro grupo, y registra el promedio de estas disimilitudes.

4.5. Análisis de Dominancia

Para identificar el método de clasificación más robusto entre los ya propuestos, se utilizará el criterio de dominancia presentado en [12], el cual nos permite comparar los ordenamientos de los métodos de clasificación. Se construye una matriz $\boldsymbol{B}$ de dimensión $n \times n$. Decimos que un departamento i (ubicado en una fila) domina fuertemente a un departamento j (ubicado en una columna), si en todos los ordenamientos obtenidos siempre tiene una mejor posición. Cuando esto pasa, en la celda correspondiente (i, j) se escribirá 1, y 0 en otro caso. En la diagonal principal se escribirá 0 porque un departamento no se domina a sí mismo. Así pues, la matriz $\boldsymbol{B}^2$ contendrá en la celda (i, j) el número de veces que el departamento en la fila i domina fuertemente al de la columna j en dos pasos [16]. La matriz $\boldsymbol{B}^k$ tenderá a la matriz $\mathbf{0}$ cuando k tiende a infinito, ya que no existirá forma de que un departamento domine fuertemente a otro en k pasos, cuando k sea demasiado grande. De modo que las dominancias posibles se tendrán en $\boldsymbol{B}^1, \boldsymbol{B}^2, \boldsymbol{B}^3, \ldots, \boldsymbol{B}^k$ para una k, tal que $\boldsymbol{B}^{k+1} = \mathbf{0}$, es decir, la

matriz nula. Con todas estas matrices se construirá una matriz integradora $\boldsymbol{B}^F$ a partir de todas ellas, del modo siguiente: en la celda (i, j) de $\boldsymbol{B}^F$ se colocará un 1 si existe al menos una $\boldsymbol{B}^t$, con $t = 1, 2, \ldots, k$, en cuya celda (i, j) haya un valor diferente de 0 (es decir, estrictamente positivo), de otro modo en la celda (i, j) se colocará un 0.

Una vez encontradas las formas en que un departamento de una fila domina a uno de una columna, basta que haya una forma de dominación del primero sobre el segundo, manifiesta en alguna $\boldsymbol{B}^t$, para cualquier t, para que se declare que el departamento de dicha fila domina al de la columna. Ahora bien, la suma de la fila i en $\boldsymbol{B}^F$, denotado por H_i, proporcionará el número de departamentos dominados por el de esa fila y con la suma de la columna j en $\boldsymbol{B}^F$ se conocerá el número de departamentos que dominan al de esa columna, lo que se denotará con G_j. Para determinar el orden estabilizado se calculan los valores $H_i - G_i$, con $i = 1, 2, \ldots, n$ y se ordenan de mayor a menor, los empates se resuelven de acuerdo con los valores de H_i. A continuación presentamos los resultados obtenidos al aplicar esta metodología.

5. Resultados y Discusión

En esta sección se muestran los resultados obtenidos con los métodos de ordenamiento previamente descritos en la Sección 4, así como los resultados del análisis de conglomerados y de dominancia. Es importante mencionar que en el ordenamiento Sin Peso, el algoritmo sólo considera el hecho de que un departamento imparte cursos (sin importar el número) a carreras de otro departamento, y no considera los cursos que imparten a sus propios programas educativos. En el caso de los métodos de ordenamiento Con Peso y PJA, ambos algoritmos consideran el número de cursos impartidos como una medida de impacto o apoyo a otros departamentos, así como el número de cursos que imparte en su mismo departamento. Cabe mencionar que sólo se consideran las materias por plan de estudios y no el número de grupos que se abren simultáneamente para cada programa. Todos los resultados presentados en este artículo se han implementado en el paquete estadístico R.

5.1. Ordenamiento de los departamentos y análisis de conglomerados

Los resultados obtenidos por los tres métodos de ordenamiento Sin Peso, Con Peso y PJA se muestran en la Tabla 2, donde se puede observar la posición obtenida para cada uno de los departamentos, los cuales se identifican con la etiqueta asignada en la Tabla 1. De manera gráfica, podemos observar estos resultados de los ordenamiento Sin Peso, Con Peso y PJA en las Figuras 4, 5 y 6, respectivamente. Note que en estas gráficas también se han presentado los resultados del análisis de conglomerados [7]. En todos los casos se eligió una clasificación en 5 grupos, que se muestran en colores. Este número de grupos es

arbitrario y se utilizó el promedio entre grupos como medida de asociación, ver Figura 3.

Tabla 2: Clasificación de los departamentos considerando los ordenamientos Sin Peso (SP), Con Peso (CP) y PJA.

Posición	SP	CP	PJA	Posición	SP	CP	PJA	Posición	SP	CP	PJA
1	36	9	9	19	12	13	46	37	2	17	30
2	6	36	36	20	37	46	7	38	29	22	1
3	42	6	6	21	47	21	20	39	31	50	49
4	53	34	42	22	18	37	22	40	19	27	54
5	11	42	35	23	22	33	47	41	3	19	50
6	9	18	34	24	25	54	11	42	1	31	28
7	13	25	18	25	8	26	44	43	23	29	24
8	35	40	12	26	14	3	37	44	54	23	29
9	7	53	45	27	52	8	8	45	17	48	32
10	39	35	25	28	26	51	10	46	49	2	48
11	40	47	26	29	21	38	4	47	27	49	51
12	34	39	53	30	20	4	33	48	51	16	41
13	10	12	13	31	4	52	2	49	28	7	17
14	44	14	40	32	5	5	23	50	48	28	5
15	45	44	14	33	50	1	19	51	30	30	31
16	33	45	39	34	32	24	52	52	16	11	16
17	46	43	21	35	38	32	27	53	41	15	15
18	43	20	43	36	15	41	3	54	24	10	38

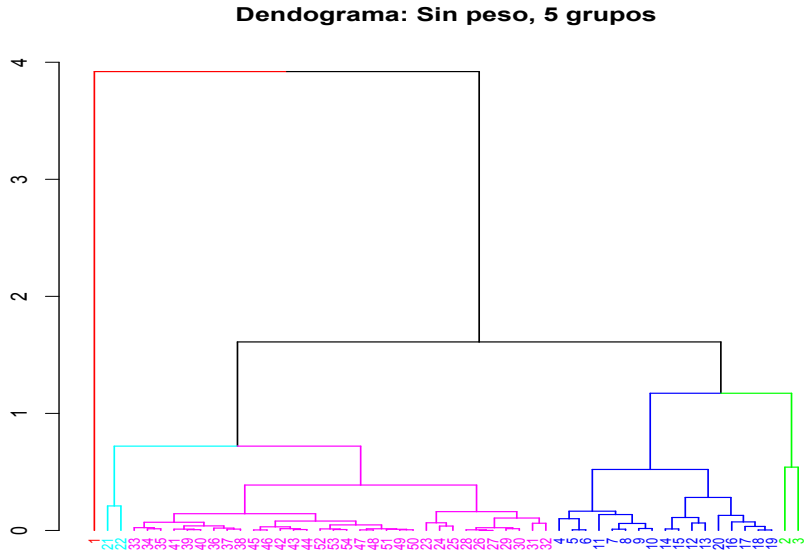


Figura 3: Dendograma utilizando datos del ordenamiento conservador Sin Peso y clasificando a los departamentos en 5 grupos. La función de enlace es el promedio.

Los resultados de la Figura 4 muestran que Filosofía encabeza el ordenamiento Sin Peso y se clasifica él solo en el primer grupo. Esto es consistente con el hecho que en todos los programas de pregrado se imparte la materia de Ética

pertenenciente a este departamento. Estadística y Administración le siguen y pertenecen al segundo grupo. En el tercer grupo se encuentran Letras, Morfología, Matemáticas y Física, Sistemas de Información, Educación, entre otros. Recordemos que este método sólo considera el hecho de que un departamento le imparte materias a otro, sin importar el número de las mismas. Luego, podemos afirmar que los cursos del segundo grupo, Estadística y Administración, son comunes en muchos programas educativos de la UAA y por ende, son apoyo para muchos departamentos. Note que para los departamentos del tercer grupo, se destaca menos esta característica.

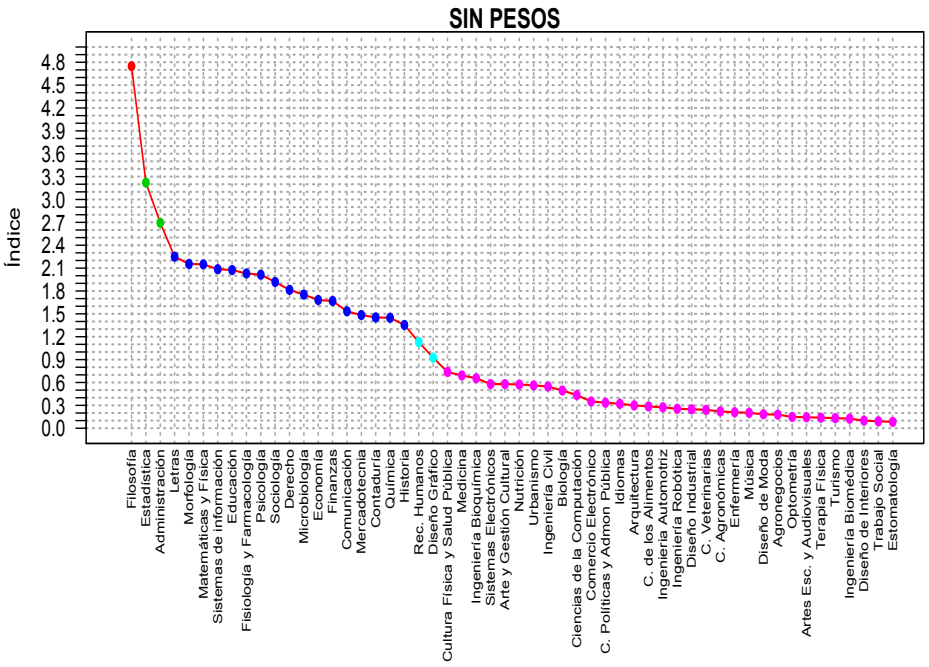


Figura 4: Clasificación utilizando el ordenamiento conservador Sin Peso.

Para el segundo ordenamiento Con Peso, ver Figura 5, observemos que el departamento de Matemáticas y Física junto con el departamento de Filosofía encabezan este ordenamiento y conforman el primer grupo. Le siguen en el siguiente grupo Estadística y Derecho, mientras que el tercer grupo está conformado por los departamentos de Administración, Diseño Gráfico, Medicina, Sociología, Letras, entre otros. Note que existe un cambio en las posiciones de los departamentos, así como en los conglomerados, con respecto al ordenamiento anterior. La aparición en los primeros lugares de los departamentos tales como Matemáticas y Física y Derecho, se debe principalmente a la gran de cantidad de cursos que imparten en los programas educativos de los departamentos. Sin

embargo, también se debe a que este método toma en cuenta la estructura de la red, pues bonifica a los departamentos que imparten cursos a departamentos con muchos cursos a su cargo.

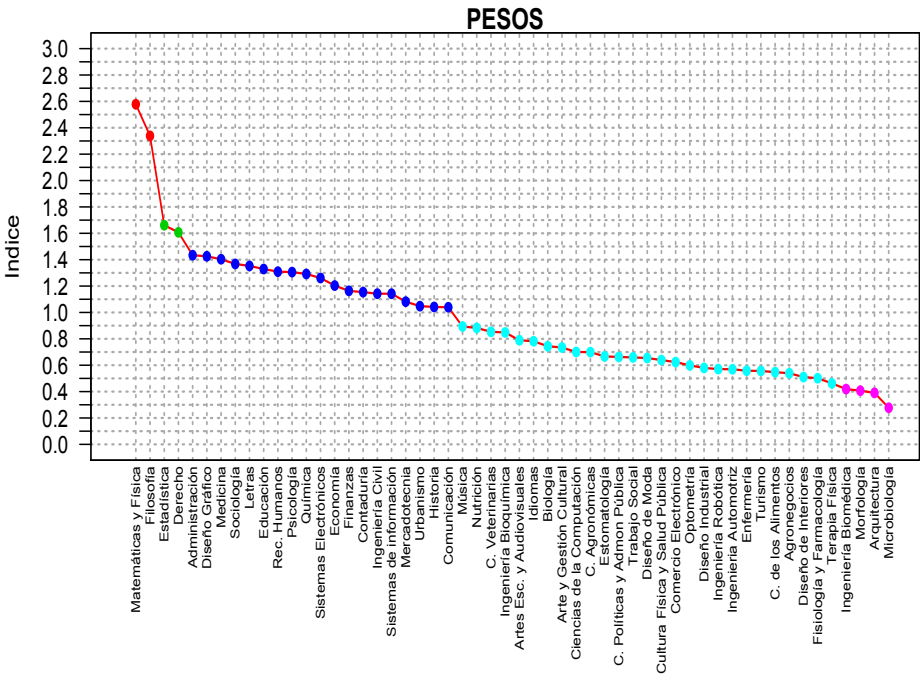


Figura 5: Clasificación y ordenamiento utilizando digráficas Con Peso.

Para el tercer ordenamiento, utilizando el PJA, ver Figura 6, Matemáticas y Física encabeza nuevamente el primer grupo de este ordenamiento, Filosofía el segundo grupo y Estadística el tercer grupo. Obsérvese que en esta clasificación, los departamentos mejor clasificados pertenecen a un solo grupo cada uno, es decir, no hay otro departamento que sea homogéneo a ellos respecto a la variable *cursos impartidos*. La mayoría de los departamentos en esta clasificación está en los grupos 4 y 5, lo cual indica que mediante la clasificación PJA, existe gran homogeneidad en la mayoría de los departamentos, con excepción de los tres primeros lugares. Esto se debe a que el rango de ponderaciones en este método es más amplio.

Retomando el ordenamiento individual de los departamentos, se puede observar que los métodos Con Peso y PJA arrojan clasificaciones muy similares, pues ambos toman en cuenta el número de cursos impartidos. De hecho, 8 de los 10 departamentos mejor clasificados son los mismos en ambas clasificaciones y 5 coinciden para los 3 métodos. Por otro lado, si consideramos el centro académico a los que pertenecen los departamentos, tanto para el ordenamiento

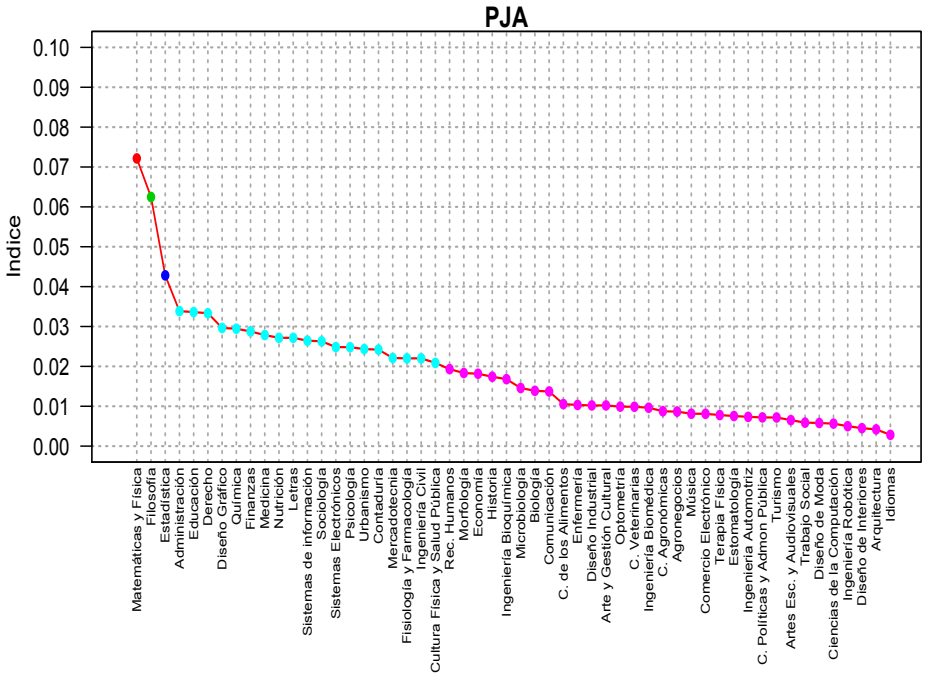


Figura 6: Ordenamiento y clasificación utilizando el PJA.

Sin Peso como para el PJA, el CC. Básicas y CC. Sociales y Humanidades son los que predominan en los primeros 10 lugares, mientras que en el ordenamiento Con Peso predomina el CC. Sociales y Humanidades. Cabe señalar que los centros que no aparecen entre los 10 mejores, son el CC. Agropecuarias, CC. de la Ingeniería y CC. Empresariales. Esto se puede interpretar como que estos últimos centros tienen menor influencia o impacto en los programas educativos de la universidad.

Ahora bien, una vez obtenidos los resultados de los ordenamientos por los tres métodos, se realizó un análisis de dominancia para identificar el método más robusto.

5.2. Análisis de dominancia y ordenamiento por centros académicos

Para analizar la robustez de los tres métodos de ordenamiento Sin Peso, Con Peso y PJA, se aplicó el análisis de dominancia expuesto en la Sección 4. Se obtuvo un valor de $k = 11$, es decir, se requirieron 12 iteraciones de la matriz original de dominancias para alcanzar a la matriz $\mathbf{0}$, y saber todas las formas en que un departamento de una fila dominaba a uno de una columna. De la matriz final $\boldsymbol{B}^F$ se obtuvo el orden estabilizado, que básicamente coloca en mejor posición al departamento que domina a los que están por debajo de él. El orden que tuvo mayor coincidencia con el generado por la matriz $\boldsymbol{B}^F$

fue el obtenido por el ordenamiento Con Peso, con 6 departamentos; le siguió ordenamiento Sin Peso con 5 departamentos y por último el ordenamiento PJA con 3 departamentos. En consecuencia, el ordenamiento Con Peso se consideró el más robusto, puesto que induce un ordenamiento más coincidente con el orden que considera todas las formas de dominancia que se pueden dar por los ordenamientos de los tres métodos de ordenamiento.

Así pues, una vez identificado el método de clasificación más robusto, el ordenamiento Con Peso, se procedió a obtener el ordenamiento de los departamentos dentro de cada centro académico utilizando este método. Cabe señalar que sólo se consideran los cursos que imparte un departamento a carreras de su mismo centro. Los resultados se presentan en la Figura 7.

El análisis para cada centro académico es el siguiente: en los departamentos de los centros CC. Agropecuarias, CC. de la Ingeniería y CC. Empresariales, no muestran gran variación respecto al ordenamiento departamental general, dado que sus departamentos corresponden al mismo conglomerado en el ordenamiento Con Peso. Lo que esto nos indica es que en estos centros la carga de cursos y su estructura es muy balanceada. Observemos que el ordenamiento al interior del CC. del Diseño y la Construcción también preserva el mismo orden que el ordenamiento general. Sin embargo, se nota una mayor diferencia en el nivel de carga de los departamentos. En el CC. Básicas, el departamento de Matemáticas y Física sigue resultando el mejor ranqueado, seguido por Sistemas Electrónicos y Estadística, quedando en cuarto lugar el departamento de Química. Esto resulta relevante, puesto que dado el análisis general, el departamento de Sistemas Electrónicos no aparecía en los primeros lugares. Esto nos indica que este departamento cobra mayor importancia para las carreras de su mismo centro. También ocurren casos en el sentido contrario, los departamentos de Derecho, Administración, Letras y Educación, tienen mayor orden o impacto a nivel general, que en sus propios centros.

Por otro lado, observamos que existen centros académicos donde el índice de ordenamiento Con Peso es muy similar para sus departamentos: CC. Agropecuario, CC. Ingeniería, CC. Empresarial y el C. de las Artes y la Cultura. Esto se puede interpretar como que estos centros tienen una carga docente balanceada. Llama la atención que estos centros son pequeños. Por el contrario, también existen centros académicos donde hay un desbalance de moderado a alto, con respecto al índice de ordenamiento: CC. Económicas y Administrativas y CC. Sociales y Humanidades, CC. de la Salud, CC. de Diseño y la Construcción y CC. Básicas; ordenados de menor a mayor. Esto sugiere que una reestructura en algunos centros podría ser conveniente para balancear las cargas departamentales y eficientar su administración. En el caso del CC. Básicas, el rango de variación del índice de ordenamiento es el más amplio de todos los centros. Observemos que en este centro se encuentra el departamento de Matemáticas y

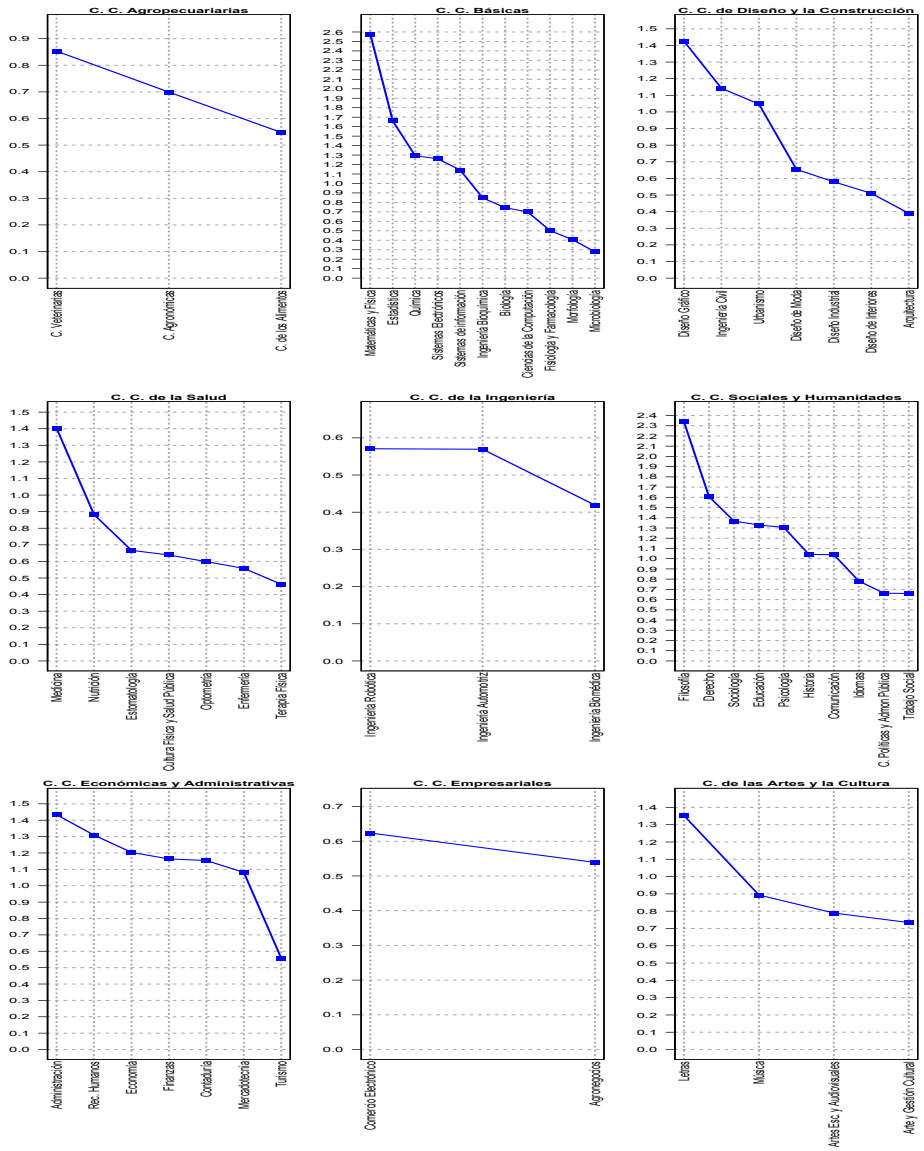


Figura 7: Clasificación utilizando el ordenamiento Con Peso dentro de cada Centro Académico.

Física, el cual tiene la mayor carga porcentual de cursos y es el primer lugar en el ordenamiento Con Peso. Esto sugiere de manera natural la necesidad de una reestructura en este centro o por lo menos del departamento de Matemáticas y Física.

6. Conclusiones

En este trabajo hemos presentado un análisis de las interrelaciones de la red departamental de la UAA, mediante la aplicación de varios tipos de ordenamiento. Esto nos permitió clasificar a los departamentos académicos de la Benemérita Universidad Autónoma de Aguascalientes utilizando teoría de juegos cooperativos y el Proceso Jerárquico Analítico. Se utilizaron dos métodos objetivos: un ordenamiento conservativo en grafos y el ordenamiento con peso en digráficas, y un método subjetivo: el PJA. La variable que se utilizó para aplicar los diferentes métodos de ordenamiento fue el número de cursos que imparten los departamentos en los programas educativos de nivel pregrado de la universidad. Siendo el método más robusto el de Digráficas con peso.

De manera general, los departamentos de Matemáticas y Física, Filosofía, Estadística y Derecho resultaron ser los departamentos mejor clasificados de la universidad. Es decir, que son los departamentos que más impactan dentro de los programas educativos de la institución. Estos departamentos pertenecen a los centros académicos CC. Básicas y CC. Sociales y Humanidades. Es importante remarcar que si bien los dos primeros departamentos son los que imparten más cursos en el UAA, Estadística y Derecho mejoraron su posición al considerar la estructura de la red. Además, al realizar el análisis por centro académico, nos permitió visualizar el balance docente e identificar a los departamentos con mayor relevancia dentro de su centro. Sin embargo, sería importante tomar en cuenta otras variables, tales como: el número de estudiantes atendidos, calidad de la investigación o actividades de difusión de la ciencia, entre muchos otros.

Así pues, el haber realizado un ordenamiento en los departamentos de la universidad nos provee de una imagen más transparente de la jerarquía departamental que podría orientar las políticas de desarrollo institucional y asignación de presupuesto.

Referencias

[1] Anderberg, M. R. (2014). Cluster Analysis for Applications: Probability and Mathematical Statistics: A Series of Monographs and Textbooks. *Elsevier Science*.

[2] Aumann, R., and Myerson, R. (1988). Endogenous formation of links between players and of coalitions: An application of the Shapley value. En:

The Shapley Value: *Essays in honor of Llyoyd Shapley*, A. Roth (Ed.), Cambridge University Press, Cambridge (UK), pp. 175-191.

[3] Borm, P., Van Den Brink, R., Slikker, M. (2002). An Iterative Procedure for Evaluating Digraph Competitions. *Annals of Operations Research*, **109**, pp. 61-75.

[4] Cherchye, L., Ooghe, E., Van Puyenbroeck, T. (2008). Robust human development rankings. *J Econ Inequal* **6**, pp. 287?321 .

[5] Criado R., Romance M., Solá L. (2014). Teoría de Perron-Frobenius: importancia, poder y centralidad. *La Gaceta de la RSME*, **17**(3), pp. 485-514.

[6] Hastie, T., Tibshirani, R. and Friedman J. (2009). The Elements of Statistical Learning. Data Mining, Inference, and Prediction. *Springer Series in Statistics*. New York (USA).

[7] James G., Witten D., Hastie T., and Tibshirani R. (2017). An Introduction to Statistical Learning with Applications in R. *Springer*, New York Heidelberg Dordrecht London.

[8] Khanmohammadi, S., and Rezaeiahari, M. (2014). AHP Based Classification Algorithm Selection for Clinical Decision Support System Development. *Procedia Computer Science*, **36**, pp. 328-334.

[9] Nawal Sael, Touria Hamim, Faouzia Benabbou. (2019). Implementation of the Analytic Hierarchy Process for Student Profile Analysis. *International Journal of Emerging Technologies in Learning,* **14**(15), pp 78-93.

[10] Myerson, R. B. (1977). Graphs and cooperation in games. *Mathematics of Operations Research*, **2**(3), pp. 225-229.

[11] Rodríguez-Esparza, L. J., Barraza-Barraza, D. Salazar-Ibarra, J., y Vargas-Pasaye, R. G. (2019). Index of suicide risk in Mexico using Twitter. *Journal of Social Researches*, **5**(15), pp. 1-13.

[12] Romo-Lozano, J. L., Zamudio-Sánchez, F. J., Martínez-Gómez, G., Rodríguez-Esparza, L. J. (2018). Uso de un criterio de dominancia para seleccionar un índice de desarrollo humano. *Región y Sociedad*, **30**(71), pp. 1-26.

[13] Saaty, T. (1980). The Analytic Hierarchy Process. *McGraw Hill.* International, Pittsburgh (USA).

[14] Saaty, T. (2003). Decision-Making with the AHP: Why is the principal eigenvector necessary. *European Journal of Operational Research;* **145**(1), pp. 85-91.

[15] Saaty, T. (2008). Decision making with the analytic hierarchy process. *Int. J. Services Sciences,* **1**(1), pp. 83-98.

[16] Searle, S. R. (1982). Matrix algebra useful for statistics. *Wiley Series in Probability and Mathematical Statistics*. New York (USA).

[17] Shapley, L. S. (1953). A Value for n-Person Games. En: *Contributions to the Theory of Games,* volume II, H.W. Kuhn and A.W. Tucker (eds.), Princeton University Press (USA), pp. 307-317.

[18] Sharpe, A., & Andrews, B.(2012). An Assessment of Weighting Methodologies for Composite Indicators: The case of the Index of Economic Well-being. *CSLS Research Reports 2012-10*, Centre for the Study of Living Standards.

Acerca de los autores

Luz Judith Rodríguez Esparza es Cátedra-CONACyT en la UAA y sus líneas de investigación son: Estadística, Probabilidad Aplicada y Procesos Estocásticos. Cuenta con publicaciones en revistas indizadas con aplicaciones en economía, actuaría, agronomía, salud, sociales y ciencias de la computación.

Julio César Macías Ponce es Profesor Investigador del Departamento de Matemáticas y Física de la UAA y sus líneas de investigación son: Optimización, Teoría de Juegos y Sistemas de Votación.

Roberto A. Kú Carrillo es Profesor Investigador del Departamento de Matemáticas y Física de la UAA y sus líneas de investigación son: Modelación matemática, Ecuaciones Diferenciales y Métodos Numéricos.

Sandra E. Delgadillo Alemán es Profesora Investigadora del Departamento de Matemáticas y Física de la UAA y sus líneas de investigación son: Modelación Matemática, Ecuaciones Diferenciales y Métodos Numéricos.

Arturo E. Giles Flores es Profesor Investigador del Departamento de Matemáticas y Física de la UAA y sus líneas de investigación son: Teoría de Singularidades en Geometría Algebraica y Geometría Analítica Compleja.

Boletín de Estadística e Investigación Operativa
Vol. 37, No. 1, Marzo 2021, pp. 38-46

Estadística Oficial

Coherence between surveys and register-based data in labour market statistics

Sara Carrascosa García
National Statistics Institute
✉ sara.carrascosa.garcia@ine.es

Abstract

The study of the labour market has always had a special interest in society, both economically and socially. In Spain, there are different sources of information that allow us to approach and measure this reality. Statistics based on surveys are available, such as the Labour Force Survey (LFS) - carried out by the National Statistics Institute of Spain (INE) -; and others using administrative data, such as those based on the social security affiliation register and the unemployed persons registered in public employment services, provided by the Ministry of Employment and Social Security. Therefore it is interesting to study the coherence among these sources of information as an aspect of improving the quality of statistics. Under the framework of the High Council on Statistics the working group of Short-Term Labour Market Statistics draws up periodically a report with the aim of analyzing and comparing the data provided by the LFS and those provided by the Ministry of Employment, using for that purpose, a harmonized methodology. In addition, a series of short-term analyses are also carried out to study the coherence between them from the point of view of the final published data. This project allows, on the one hand, to compare, study and review methodological differences, and on the other to analyze the information that users receive from different sources of information.

Keywords: Administrative data, Coherence, Labour market.

1. Coherence as a principle of quality

The quality management system of National Statistics Institute of Spain (INE) is based on the European Statistics Code of Practice (CoP) (see [3] and [5]), which sets the standard for developing, producing and disseminating European statistics through fifteen principles.

In this document we will focus on the study of the principle 14 of the CoP which establishes coherence and comparability as a dimension of the quality of the statistics.

This principle of coherence defined in the CoP covers the internal consistency, over time and regional comparability, but also the possibility of combining related data from different sources. The 14th principle is developed through five indicators and we are going to focus on Indicator 14.4 which looks at how statistics from different sources and of different periodicity are compared and reconciled. The concept of coherence is further broken down into more specific ones, but we will concentrate on cross-domain coherence. The aim of cross-domain coherence is to compare the extent to which statistics are reconcilable with those obtained through other data sources or statistical domains.

According to the latest User Satisfaction Survey (USS2016), carried out by INE, coherence is the most difficult quality dimension for users to understand, which is reflected in the low response rate of the question related to this aspect of quality. So, it could be interesting to carry out projects analyzing coherence among different sources. That being said, the coherence of the INE labour market statistics is rated positively or very positively by 94,9% of users.

2. Different sources of labour market data

Nowadays, official statistics has access to different data sources: surveys, administrate data, web data, etc.

On the one hand, there is labour market data from **surveys**. The information that comes from surveys, allows us to define concepts such as **employment** or **unemployment** according to the International Labour Organization (ILO) definitions.

On the other hand, there is data from **administrative registers**. The purpose of administrative registers is always administrative management, but data could be the information source of a statistic, as well. Therefore, although their concepts do not coincide with those defined by the ILO, they might be very similar.

- In the case of **employment**, there is the term **"affiliated"**, which refers to those who pay social security contributions.
- In the field of **unemployment**, there is the term **"job seeker"** which refers to someone who is registered in a public employment office.

3. Labour Market Statistics in Spain

One of the main pillars of the Official Statistics in Spain is the four-year National Statistical Plan (PEN) which lists all statistical operations and statistical activities conducted by the institutions of the National Statistical System

(NSS), and defines the dissemination of the results. All the statistics included in the PEN are considered statistics for state purposes and must be completed.

The sources of information on the labour market in Spain included in the PEN are numerous and diverse:

- **The Economically Active Population Survey (EAPS)**: is a continuous investigation of quarterly periodicity directed to families and has been carried out since 1964. Its main objective is to obtain data on the population in relation to the labour market: active, unemployed and inactive people. The EAPS follows the concepts defined by the ILO and Eurostat. This allows EAPS to be internationally comparable and serves as a basis for the development of the Labour Force Survey (LFS).

- **Statistics on the affiliation of workers to Social Security**: Its objective is to obtain and disseminate data on affiliated workers and movements of entering and leaving of the social security system. These statistics are based on the administrative register of **Social Security Affiliates**. This is where the additions and deletions of social insurance contributions are registered.

- **Registered Labour Movement Statistics**: Its objective is to obtain and disseminate data on the labour movement registered with the public employment services. These statistics are based on the administrative register of the **job seekers register**.

These three statistical operations measure similar realities (employment and unemployment) within the scope of the labour market. This means that they inevitably interact with each other, giving rise to a variety of situations within an employment and unemployment scenario.

Focusing on **employment**:

- Typically, employed persons, as defined by the ILO, are registered in a Social Security Administration.

- There are also people who are employed, as defined by the ILO, and are not registered with the Social Security Administration. We are referring, for example, to the black economy.

- Finally, there are people registered in the Social Security registers who are not considered employed, according to the ILO definition, e.g. people covered by Special Health Care Provision or people not working but contributing to Social Security in a special scheme in order to top up their state pension.

Focusing on **unemployment**:

- Most commonly, unemployed people, as defined by the ILO, actively seek employment and are registered as job seekers in a public employment office.

- However, many of the unemployed, as defined by the ILO, are not registered with public employment offices for various reasons. The most common reason is because they do not trust that it is a good way to find work.

- Finally, there are people who are registered as job seekers but who are not classified as unemployed according to the ILO definition, either because they do not actively seek employment or because they actually do work but in the black economy.

4. Conciliation projects in Spain

In Spain, the High Council on Statistics and the Working Group on Current (Short-term) Labour Market Statistics have supported and encouraged the study of the coherence between sources in this field.

In order to implement the mentioned study, the Working Group carries out two conciliation projects with the objective of studying the coherence between sources through the reconciliation of their figures. These projects are aimed at the end user and a final report and the results are published on INE's website every two years.

The studies focus on two levels: Macro and Micro analysis; and on two concepts: Employment and Unemployment.

- **Macro reconciliation** compares the results of the statistics with each other, trying to create conceptual identities between them.

- **Micro-conciliation** consists in comparing the statistical data in order to check whether it is consistent with each other. If the data is consistent, it is foreseeable that the statistics created from it will also be coherent.
- The term **"employment"** is legally linked to the right to be affiliated to the Social Security or other equivalent mutual insurance companies (MUFACE, MUGEJU and ISFAS). This makes the concepts of occupation and affiliation similar and therefore, the reconciliation process possible.
- The term **"unemployment"** is not legally linked to any administrative register. According to the ILO's definition, "being unemployed" implies a lack of work, willingness to work and an active search of work. "Active job search" involves a subjective perception of the respondent and, in this sense, is very difficult to measure reliably.

The reports begin with the definitions and peculiarities between methodologies of the information sources involved. Then they explain the reconciliation methodology of the publication and, finally, they comment on the most important results. The publication provides data that crosses the following classification variables: sex, age, nationality, regional breakdown (NUTS2), and activity sector. Currently, there are two periodic publications: the analysis of unemployment at micro level (published the odd-numbered years) and the analysis of employment at macro level (published the even-numbered years). In addition to that, a report analysing the employment at micro level was published in October 2019. It is important to note that the reconciliation of the figures will always be done under the prism of the EAPS methodology. They are all available on the INE website in both English and Spanish.

4.1. Unemployment - Micro Level

The micro-conciliation of unemployment data is carried out by cross-checking individual data from the EAPS with the databases of the Public State Employment Service (PSES), see [2]. In this publication we work with the concept of job seeker and unemployed person. For the matching process, we obtain the ID number of the EAPS sample though record linkage techniques.

The reconciliation process is structured as follows:

- Match the EAPS sample with the PSES database.
- From those who were found in the PSES database, we check if they are job seekers registered in the reference period.
- Match persons of the EAPS sample, who state in the survey that they are registered as job seekers in a Public Employment Office, with the job seekers registered of the PSES database.

- Match persons of EAPS sample, who are classified as unemployed person, with the job seekers registered who are consider as unemployed according to the Registered Labour Movement Statistics.
- Match persons of EAPS sample, who state in the survey that they receive unemployment benefits, with the job seekers registered that actually receive benefits.

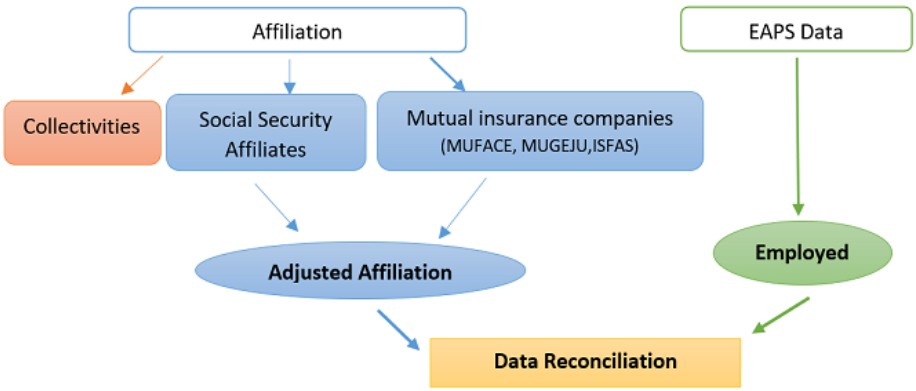
In overall terms, despite the fact that the time evolution is coherent and the two concepts have a similar name, a direct comparison of the unemployment registered figures and the unemployment estimated by the EAPS is not possible, as they correspond to different concepts and come from different sources.

4.2. Employment - Macro Level

This study is carried out in even-numbered years with the aim of comparing, in a homogeneous way, the employment data obtained from the EAPS and the Social Security affiliation. The affiliation data comes from the statistical use of the databases maintained by the General Social Security Treasury (Ministry of Employment and Social Security-MEYSS), see [1].

Both Affiliation and EAPS data have their own particularities, derived from their own methodology and purposes, and offer a reflection of the evolution of employment in Spain filtered through their respective specificities. For a comparison of results, it is necessary to make adjustments to the groups covered in each case and to homogenise the reference period of the data.

These adjustments are made by adding other mutual societies specific for civil servants, such as MUFACE, ISFAS or MUGEJU to the Social Security schemes. The sum of Social Security affiliates and those affiliated to MUFACE, ISFAS and MUGEJU correspond to the term Adjusted affiliation. Adjusted affiliation is a similar concept to that of employed persons in the EAPS and will be used to reconcile the figures from both sources.



As stated above, the EAPS is the source of data that serves as a reference for the coherence study projects. In this sense, the figures used in this publication are the same as those used in the statistics. Only in the regional breakdown are there differences such as, the fact that instead of taking into account the place of residence of the respondent, the province of the place of work is considered. In this way we gain comparability.

On the other hand, when analysing the results, a distinction is made between the agricultural and non-agricultural sectors (including all mutualities).

In the comparative analysis, the information is broken down by regional scope, sex, age, professional situation, economic activity, type of employment relationship, type of working day and nationality.

The latest available publication compares data for the second quarter of 2016. The EAPS occupancy rate exceeds the adjusted affiliation rate (531600 people in favor of the EAPS occupation). This situation is maintained when analysing the figures by sex and nationality (both for Spaniards and foreigners). By type of working day, there are major differences, which may be due to the fact that to work less than 40 hours per week is considered to be part-time work in affiliation data. Overall, the employment study show more coherence among the two statistics studied.

4.3. Employment - Micro Level

As in the case of unemployment, we could also carry out a comparative analysis at the micro level, using the data from the EAPS as a basis and searching for each person surveyed in the register of Social Security affiliation (see [4]).

This study is fundamentally concerned with the characterization of dissenting cases and with the analysis of the extent to which the relationship with the labour market obtained from the EAPS and the Social Security situation can be explained.

This micro analysis is more complex than the unemployment analysis because the administrative register has more functions than the management of the people who work. The register also includes non-working members, such as some special agreements aimed at supplementing pension rights.

On the other hand, as with the micro unemployment reconciliation, when the reference period is extended to more than the EAPS reference week, the percentage increases in both sources. In addition, the full micro study should also involve the members of the mutual insurance companies.

This project was published for the first time in October 2019 and it is focused on the difference found in data of 2016. The results indicate very high consistency between the two data sources, 92,6% in the first automatic approach. Regarding discrepancies, a high percentage of cases can be explained by specific situations. Indeed, taking these cases into account the coherence arises to 96,0%.

Within the discrepancy cases of those employed persons in the EAPS and not found in the Social Security register, we found again civil servants, who are covered through mutual societies specific to these groups, special professions that have alternative mutual societies (such as lawyers, architects or doctors), and people who reside in Spain but work abroad. Within the group of those found in the register but not being employed in the EAPS, we found people affiliated in a special system for employed agricultural workers, and those with a partial retirement contract. Due to the characteristics of these laboral relationships, it is possible that they are not working at the time they are surveyed in the EAPS.

At the end, we have 4,0% of not explained incoherence. In this residual inconsistency, we find a greater number of people who are self-employed or who are assisted by a family member, with temporary contracts with fewer hours with elementary training and who are foreigners. In addition to that, there are cases that would correspond to people who contributed without working or who simply did not respond to the survey truthfully.

4.4. Quarterly Coherence

To complete the micro/macro reconciliation studies, a comparative study is carried out, each quarter, between the figures published in the EAPS, the average affiliation of Social Security Affiliates Statistics, and the registered unemployment of the Registered Labour Movement.

In this case, it is not intended to make a methodological approximation or reconciliation of figures, but rather to analyse the differences received by the end user, that is, to know the perception of cross-domain coherence that the user has.

At the same time, it allows us, as producers of statistics, to better understand the aspects of the labour market that EAPS cannot collect, and that an administrative register can better analyse.

5. Conclusion

As this document has shown, users of Spanish labour market statistics have not only information provided by the EAPS but also statistics based on administrative data. In this situation, users may think that the information that they receive is contradictory or inconsistent. Consequently, users may perceive that the quality of those statistics is lower. In this sense, these projects are fundamental, because they provide users a better comprehension of labour market statistics, their similarities and differences.

Furthermore, from the statistics producer perspective, the study of coherence improve the quality of their statistics as it allow them to better know the concepts and the reality that they study.

While the cross-domain coherence analysis of employment figures outcomes in a high consistent between the two sources, the unemployment analysis is more difficult. This discrepancy could be explained by the difference in the definition of the concept made by the ILO. The unemployment as is defined by the ILO, implies a lack of job and a willingness of working, which is more subjective and difficult to find in an administrative register.

References

[1] Comparison of Employment according to The Economically Active Population Survey and to affiliation data. January, 2018.

[2] Comparison of Unemployment according to The Economically Active Population Survey and to Public Employment Service data. September, 2017.

[3] European statistics Code of Practice. Revised edition. 2017.

[4] Micro conciliation of employment as measured by the Economically Active Population Survey and Social Security Affiliation. October, 2019.

[5] Quality Assurance Framework of the European Statistical System. Version 1.2.

About the author

Sara Carrascosa García has a degree in Mathematics from the Complutense University of Madrid and she is specialized in Statistics and Operation Research. She is a Goverment Statistician (Cuerpo Superior de Estadísticos del Estado) and works in the Spanish National Statistics Institute (INE) from 2015. She has worked in several different departments of the INE and she has participated in numerous national and international projects. She currently applies her professionnal skills in the S.G. for Labour Market Statistics.

Boletín de Estadística e Investigación Operativa
Vol. 37, No. 1, Marzo 2021, pp. 47-37

Historia y Enseñanza

An introduction to Markov Chain Monte Carlo

Ricardo Medel Esquivel and Isidro Gómez-Vargas
CICATA-Legaria
Instituto Politécnico Nacional (México)
✉ rmedele1500@alumno.ipn.mx,, ✉ igomezv0701@alumno.ipn.mx,

J. Alberto Vázquez
Instituto de Ciencias Físicas
Universidad Nacional Autónoma de México
✉ javazquez@icf.unam.mx

Ricardo García Salcedo
Departamento de Ingeniería Civil, División de Ingeniería[1]
Universidad de Guanajuato (México)
✉ rigarcias@ipn.mx

Abstract

In this paper we present a didactic introduction to the Markov Chain Monte Carlo method. We do not assume that the reader has prior knowledge about the subject, so we also present the necessary fundamentals of probability theory, a historical outline of the evolution of the Monte Carlo method and practical examples on numerical integration and statistical inference implemented in Python.

Keywords: Markov chains, Monte Carlo method, Random numbers.
AMS Subject classifications: 65C05.

1. Introducción

Bajo el nombre de método de Monte Carlo (MC) se designa a una familia de técnicas computacionales basadas en el muestreo aleatorio (en la generación de números pseudoaleatorios) y empleadas para hallar soluciones aproximadas de una gran variedad de problemas matemáticos surgidos en diversos campos de la ciencia (por ejemplo en la investigación, en genética y biología, en astronomía y

[1]Estancia sabática.

cosmología, en los métodos bayesianos de la estadística, la minería de datos y el aprendizaje automático, entre otros). No hay pues un método MC sino muchos de ellos y puede decirse que, en general, se usan en tres modalidades distintas como [30]:

1. técnica de integración numérica;

2. método de muestreo aleatorio para simulación u optimización;

3. procedimiento de prueba de hipótesis.

La historia del método MC es fascinante. El contexto en el cual Stanislaw M. Ulam lo ideó y se empleó por vez primera coincide con dos hechos científicos de importancia mayor en la historia mundial del siglo XX: la construcción de la primera bomba atómica (el proyecto Manhattan) y la disponibilidad de una de las primeras computadoras de uso general: la ENIAC (Electronic Numerical Integrator And Computer). Coincidieron entonces la necesidad práctica de un método de cálculo poderoso y la herramienta adecuada para llevarlo a cabo.

En la actualidad, sobre todo en relación con la estadística bayesiana, las versiones del método MC más utilizadas se basan en la construcción de cadenas de Markov para obtener muestras de distribuciones de probabilidad específicas; estas extensiones del método MC se conocen como métodos de Monte Carlo vía Cadenas de Markov (MCMC, por sus siglas en inglés).

Presentamos aquí una primera aproximación a los métodos MCMC que pretende ser concisa, accesible y práctica para cualquier persona interesada en estos métodos y que cuente con un mínimo de requerimientos técnicos. Deseamos mostrar las ideas básicas del método y señalar porqué es plausible su funcionamiento, más que discutir consecuencias, alcances y dificultades del mismo. La estructura del artículo es la siguiente. En la Sección 2 se presenta una breve reseña del desarrollo histórico de los métodos MC y MCMC. En la Sección 3 se citan los teoremas de probabilidad fundamentales para comprender el método MC; en el Apéndice se ofrece una referencia más detallada, para aquellos lectores con menos bagaje sobre teoría de probabilidad. La Sección 4 es una recopilación de algunas ideas básicas sobre la generación de números pseudoaleatorios, que son el fundamento de todos los métodos MC. En la Sección 5 se ponen en práctica las ideas esenciales de los métodos de integración MC, mediante un ejemplo resuelto por tres aproximaciones distintas. La Sección 6 es un resumen de la teoría de cadenas de Markov, relevante para la implementación moderna de los métodos MC. Finalmente, en la Sección 7 se presenta otro ejemplo, esta vez para encontrar los parámetros de un ajuste lineal mediante un muestreo MCMC realizado con un algoritmo Metropolis-Hastings. Los códigos utilizados están disponibles en [8].

2. Reseña histórica

Stanislaw Ulam cuenta que la idea del método de Monte Carlo se le ocurrió cuando jugaba al solitario con un mazo de cartas, mientras se recuperaba de una enfermedad en 1946 [3, 18, 29]. Analizando el juego, se percató de que para tener una idea de la probabilidad de que salga un solitario (cosa en la que no influye la habilidad del jugador) era más práctico ir echando las cartas o experimentar con el proceso para tomar nota de cuántas veces salía, que tratar de determinar todas las combinaciones posibles, pues crecen exponencialmente y son difíciles de calcular, salvo en casos excepcionales. Esto le llevó a darse cuenta de que en problemas complejos es mejor un muestreo aleatorio que el análisis completo de todas las posibilidades.

En 1946, Ulam le planteó esta idea a John von Neumann –quien había jugado un papel muy importante en el desarrollo de la ENIAC– durante un largo viaje en automóvil y lo convenció de las posibilidades de los esquemas de cálculo probabilistas. Juntos, Ulam y von Neumann desarrollaron las matemáticas del nuevo método y publicaron el primer artículo sobre éste en 1949 [20]. El nombre de Monte Carlo, al parecer, fue sugerido por Nicholas Metropolis, en alusión al casino de Monte Carlo, lugar al cual era aficionado un tío de Ulam [18].

Sin embargo, las investigaciones históricas evidencian que el método ya era conocido y se había usado previamente para resolver algunos problemas de estadística. Incluso podría remontarse hasta el año 1773 con el experimento de la aguja de Buffon, que sirve para estimar el número π (aunque esta historia podría ser apócrifa) [12]. Lo que sí está documentado es que hacia 1901 Kelvin usó una simulación MC rudimentaria, valiéndose de cartas numeradas para estimar una distribución de velocidades; también Fermi se valió de cálculos manuales tipo MC en su trabajo sobre la fisión nuclear en la década de 1930 [23]. Y existen informes de que el método fue propuesto independientemente por J. E. Mayer para tratar problemas de la física del estado líquido [19].

En 1952 Metropolis y sus colaboradores propusieron el primer algoritmo de muestreo MCMC y lo implementaron en el MANIAC (siglas en inglés de Mathematical Analyzer, Numerical Integrator and Computer) de los Álamos para derivar, de la distribución de Boltzmann y del modelo de esfera dura, algunas propiedades físicas de un sistema simulado compuesto por 224 partículas ubicadas en una retícula de área unitaria [17, 19]. La aplicación del algoritmo de Metropolis al modelo que describe las transiciones de fase de un ferromagneto, conocido como modelo de Ising bidimensional, es quizá la más difundida, tanto por su utilidad pedagógica como por resultar fundamental en los trabajos de reconstrucción de imágenes [23].

Físicos y químicos, principalmente, desarrollaron generalizaciones y alternativas del algoritmo de Metropolis, sobre todo en relación con el modelo de Ising. Este trabajo de exploración y generalización culminó en 1970 cuando W.

K. Hastings logró fundamentar matemáticamente el algoritmo de Metropolis [10, 11, 25], y éste pasó a ser conocido como algortimo Metropolis-Hastings. El artículo de Hastings mostró el potencial del algoritmo como herramienta de muestreo de propósito general, más allá de los ámbitos de la física estadística, sin embargo pasarían varios lustros hasta que todas sus implicaciones fuesen valoradas acertadamente.

En 1983 S. Kirkpatrick y colaboradores introdujeron el algoritmo de Recocido Simulado [13] para abordar problemas de optimización combinatoria, en los cuales se desea minimizar una función objetivo, llamada costo, en cierta región, denominada espacio de búsqueda. Un ejemplo típico es el problema del vendedor viajero, en el cual se busca minimizar la distancia de un recorrido que ha de pasar por un conjunto de vértices dado. Kirkpatrick y colaboradores decidieron pensar la función objetivo como un tipo de energía, de esta manera pudieron aprovechar las intuiciones de la física estadística y aplicar el algoritmo de Metropolis. Con este algoritmo el uso de los MCMC se extendió a la investigación de operaciones, la biología, economía e ingeniería eléctrica [23].

El algoritmo de Recocido Simulado, así como el modelo de Ising, trabajan sobre una retícula dividida en celdas. Esto inspiró a los hermanos Donald y Stuart Geman a aplicar las ideas del algoritmo de Metropolis y el Recocido Simulado, en combinación con un enfoque estadístico bayesiano, para atacar el problema de la restauración de imágenes digitales; su enfoque los llevó a encontrar en 1984 una variante del algoritmo de Metropolis que culminó dos siglos de historia del enfoque estadístico bayesiano: el algoritmo de Gibbs [6].

La restauración de imágenes digitales por medio del algoritmo de Gibbs constituyó la primera aplicación popular conjunta de los métodos estadísticos bayesianos y los MCMC, aunque ya en 1968 los MCMC habían sido utilizados secretamente como una herramienta de la estadística bayesiana en la localización de un submarino de combate de propulsión atómica, el U.S.S. Scorpion. Este episodio histórico, así como un relato pormenorizado de los avatares del teorema de Bayes, desde sus inicios hasta la síntesis que supuso el algortimo de Gibbs, pueden encontrarse en el libro de Sharon Bertsch McGrayne [16].

En la década de 1990 el método de Monte Carlo experimentó una completa renovación al combinarse con los métodos bayesianos de inferencia, así como por la disponibilidad creciente de computadoras potentes y económicas; fue en esta época, evidenciado ya su potencial, cuando los MCMC se popularizaron como herramienta invaluable de la inferencia estadística de propósito general. El software especializado en análisis estadístico bayesiano BUGS (Bayesian Inference Using Gibbs Sampling), lanzado en 1992 y aún disponible en internet como OpenBUGS [28] contribuyó enormemente a la popularidad de los MCMC al hacerlos accesibles al público, principalmente para aplicaciones en Bioestadística.

Los lenguajes de programación modernos han permitido generar una ga-

ma muy amplia de innovaciones en relación con los MCMC y también formas alternas de muestreo aleatorio, cada vez más eficientes. R es un lenguaje de programación orientado a la estadística; por otro lado, Python se ha revelado como un lenguaje de programación muy versátil, popular en muchas áreas de la ciencia y la ingeniería: módulos científicos desarrollados en Python, R y otros lenguajes modernos se crean constantemente.

PyMC [22], por ejemplo, es un módulo de Python especializado en análisis estadístico bayesiano y contiene métodos MCMC de propósito general. Su desarrollo comenzó en 2003 con la intención de hacer accesibles los MCMC a los investigadores no especializados en estadística y estaba dirigido principalmente a los ecologistas. La publicación de su primera versión ocurrió en 2005; actualmente (2020), la versión 4.0 de PyMC está en desarrollo. Además existen otros módulos, muchos de ellos con aplicaciones a disciplinas específicas, como MontePython [1], especializado en Cosmología.

En la actualidad los métodos MCMC tienen un uso generalizado en el análisis intensivo de datos en áreas de la ciencia que trabajan con grandes cantidades de datos. Sin embargo, la complejidad de los modelos y las enormes cantidades de datos a tratar han orientado la búsqueda de técnicas de muestreo más eficientes que las derivadas de los métodos MC.

A manera de resumen de esta sección, la Tabla 1 enlista algunos de los acontecimientos más importantes en la historia del Método de Monte Carlo.

Nota: La lista de acontecimientos no es totalizadora.
Fuente: Elaboración propia.

Año	Acontecimiento
1949	Método de Monte Carlo
1953	Algoritmo Metropolis
1970	Algoritmo Metropolis-Hastings
1983	Algoritmo de Recocido Simulado
1984	Muestreo de Gibbs
1992	BUGS
2000-2020	Módulos especializados en Python y R

Tabla 1: Hitos en la historia de los métodos de Monte Carlo.

3. Antecedentes de probabilidad

En esta sección se citan las definiciones y los teoremas de probabilidad indispensables para comprender el funcionamiento del método MC. En el apéndice, siguiendo el excelente resumen que hace S. M. Ross [26], ofrecemos informa-

ción más completa para aquellos lectores poco familiarizados con la teoría de probabilidad.

Definición 3.1. *El valor esperado de $g(x)$, función de la variable aleatoria X, para el caso en que X adopte valores discretos de una función densidad de probabilidad $p(x)$, se define como*

$$\mathbb{E}[g(X)] := \sum_x g(x)p(x),$$

y para el caso en que X sea una variable aleatoria continua con función densidad de probabilidad $f(x)$

$$\mathbb{E}[g(X)] := \int_{-\infty}^{\infty} g(x)f(x)dx. \tag{3.1}$$

Definición 3.2. *Para una variable aleatoria X con media μ se define la varianza como*

$$Var(X) := \mathbb{E}[(X-\mu)^2].$$

Definición 3.3. *La covarianza de dos variables aleatorias, X de media μ_x y Y de media μ_y, se define como*

$$Cov(X,Y) := \mathbb{E}[(X-\mu_x)(Y-\mu_y)].$$

La desigualdad de Chebyshev (ver Apéndice para mayores detalles) permite arribar a los teoremas más importantes para el método de Monte Carlo ordinario, que citamos a continuación.

Teorema 3.1. ***Ley débil de los grandes números.***
Sea $X_1, X_2, ...$ una sucesión de variables aleatorias independientes e idénticamente distribuidas con media μ. Entonces, para cada $\epsilon > 0$

$$P\Big\{\Big|\frac{X_1+...+X_n}{n} - \mu\Big| < \epsilon\Big\} \to 0, \quad \textit{cuando } n \to \infty.$$

La generalización de este resultado es:

Teorema 3.2. ***Ley fuerte de los grandes números.***
Bajo las condiciones del teorema anterior, se tiene con probabilidad 1, que

$$\lim_{n\to\infty} \frac{X_1+...+X_n}{n} = \mu. \tag{3.2}$$

Y un resultado de gran utilidad para realizar inferencia estadística es el:

Teorema 3.3. ***Teorema central del límite.***
Dada una sucesión de variables aleatorias independientes e idénticamente distribuidas, $X_1, X_2, ...$, con media finita μ y varianza finita σ^2 se cumple que

$$\lim_{n\to\infty} P\Big\{\frac{X_1+...+X_n-n\mu}{\sigma\sqrt{n}} < x\Big\} = \phi(x), \tag{3.3}$$

donde $\phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-x^2}{2}} dx$*, para* $-\infty < x < \infty$*, es la distribución normal o gaussiana.*

La ley fuerte de los grandes números es el fundamento de la integración por el método de Monte Carlo. Nótese, en primer lugar, que la Ecuación (3.1) permite interpretar una integral $\int_a^b g(x)dx$ como el valor esperado, μ, de la función $g(x)$ respecto a una distribución de probabilidad uniforme definida para $a \leq x \leq b$; es decir, si $f(x) = Unif(a, b) = \frac{1}{b-a}$, multiplicando la integral anterior por el factor $(b - a)$, se tiene

$$\int_a^b g(x)dx = (b-a)\int_a^b g(x)f(x)dx = (b-a)\mathbb{E}[g(X)] \text{ con } X \sim Unif(a, b).$$

Por otro lado, la ley fuerte de los grandes números (Ecuación 3.2), permite aproximar este último valor esperado mediante el promedio de una muestra de tamaño n de la variable aleatoria distribuida uniformemente, si se toma n suficientemente grande. Así:

$$\int_a^b g(x)dx = (b-a) \lim_{n\to\infty} \frac{X_1 + ... + X_n}{n} \text{ con } X \sim Unif(a, b).$$

La cuestión relevante, ahora, es cómo generar las variables aleatorias uniformemente distribuidas, $X_1, ..., X_n$. Exploramos esto en la siguiente sección.

4. Números pseudoaleatorios

Existen dispositivos mecánicos capaces de generar muestras o secuencias de números aleatorios, usualmente en relación con algunos juegos de azar: dados, ruletas, tómbolas, etcétera. En una computadora es posible implementar algoritmos para generar secuencias de números que, bajo ciertas pruebas estadísticas, sean prácticamente indistinguibles de las muestras generadas por procesos realmente aleatorios. Para hacer evidente la distinción, se llama números pseudoaleatorios a dichas secuencias producidas en una computadora, y a los algoritmos utilizados en esta tarea se les denomina generadores de números aleatorios.

Técnicamente, se dice que una secuencia de n números es aleatoria si no puede predecirse el valor de la entrada n_{k+1} dadas las entradas anteriores n_i, con $i = 0, 1, ..., k$.

Además, una secuencia pseudoaleatoria es una secuencia de números generada de manera determinista que debe resultar, bajo ciertos criterios estadísticos, indistinguible de una verdadera secuencia aleatoria.

Nótese que esta definición no hace referencia alguna a la distribución de probabilidad (que llamaremos distribución padre) generadora de la muestra de números pseudoaleatorios. Sin embargo, los números pseudoaleatorios fundamentales para fines prácticos son los que siguen una distribución uniforme, que se conocen como números rectangulares o, sencillamente, números aleatorios.

En 1946 John von Neumann propuso el primer algoritmo generador de números pseudoaleatorios: el método del cuadrado central. Consiste en proponer un número de n dígitos, al que se le llama semilla, se eleva éste al cuadrado y del resultado se toman los n dígitos centrales; el procedimiento se repite con el nuevo número de n dígitos, hasta generar toda una secuencia de números de n dígitos. La Tabla 2 esboza un ejemplo.

Nota: Números generados de la semilla $X_0 = 1981$.
Fuente: Elaboración propia.

N	X_N	$(X_N)^2$	X_{N+1}
0	1981	39**2436**1	2436
1	2436	59**3409**6	3409
2	3409	11**6212**81	6212
3	6212	38**5889**44	5889
4	5889	34**6803**21	6803
...	...	...	...

Tabla 2: Método del cuadrado central.

Este algoritmo funciona a cierto nivel, pero puede tener fallas severas. Por ejemplo, si se toma el número 3792 como semilla y se eleva al cuadrado, se obtiene 14379264; luego, esta semilla produce la secuencia no aleatoria 3792, 3792, 3792,... [5].

Otro método para producir números pseudoaleatorios, el generador congruencial multiplicativo, consiste en comenzar con una semilla x_0 y calcular de manera recursiva x_n, para $n \geq 1$:

$$x_n = ax_{n-1} \text{ módulo } m,$$

siendo a y m enteros positivos y donde la operación módulo m indica que x_n es el residuo de dividir ax_{n-1} entre m.

Existen también generadores congruenciales lineales y mixtos, del tipo:

$$x_n = ax_{n-1} + c \quad \text{módulo} \quad m.$$

En estos generadores los números a, c y m se deben elegir siguiendo ciertas reglas basadas en la aritmética modular y de acuerdo con las características de la memoria de la computadora para evitar que la secuencia alcance rápidamente el período que es característico de las estructuras congruenciales y, en consecuencia, deje de ser una muestra aleatoria [2].

Los generadores anteriores producen números pseudoaleatorios distribuidos uniformemente. Para generar números distribuidos bajo otras funciones de probabilidad se usan técnicas más sofisticadas, como el método de la transformación

inversa o el algoritmo de aceptación y rechazo, ambos ampliamente usados hasta nuestros días y propuestos también por John von Neumann, en una carta dirigida a Ulam fechada en 1947 [3]. El método de la transformada inversa opera bajo el siguiente resultado [26]:

Teorema 4.1. *Sea U una variable aleatoria uniforme en $(1, 0)$. Para cualquier función de distribución continua F, la variable aleatoria X definida como*

$$X = F^{-1}(U),$$

donde F^{-1} es valor de x tal que $F(x) = u$, tiene distribución F.

Este método aplica cuando puede determinarse la forma analítica de la inversa de la distribución de probabilidad padre de los números pseudoaleatorios. Sin embargo, esto no siempre es posible, incluso para funciones de distribución tan usuales como la distribución normal. En todos los casos en que no puede aplicarse el método de la transformada inversa es necesario un método particular. La referencia [15] es un manual que recopila técnicas para generar números pseudoaleatorios para muchas distribuciones de probabilidad.

El algoritmo de aceptación y rechazo, por otro lado, es una forma eficaz para generar muestras de una variable de distribución padre $f(x)$ y contiene el germen de las ideas para técnicas más elaboradas, así como para introducir el uso de las cadenas de Markov. Consiste en proceder indirectamente, a partir de una distribución $g(x)$, que previamente ya se sepa generar; cada valor se genera con una probabilidad proporcional a $\frac{f(y)}{g(y)}$, según el siguiente esquema. Se toma una constante c, de modo que para toda y se cumpla $\frac{f(y)}{g(y)} \leq c$ y se procede de manera recursiva, según el Algoritmo de Aceptación y Rechazo, mostrado en la Tabla 3.

Fuente: [26, 27].

Aceptación y Rechazo
Paso 1: Generar $Y \sim g(\cdot)$.
Paso 2: Generar un número aleatorio $U \sim Unif(0, 1)$.
Paso 3: Si $U < f(Y)/cg(Y)$, $X = Y$. Otro caso, ir a Paso 1.

Tabla 3: Algoritmo 4.1.

Al final de la sección anterior vimos que para estimar numéricamente la integral $\int_a^b g(x)dx$ podemos usar una muestra de una variable con distribución de probabilidad uniforme. Los números pseudoaleatorios con esta distribución son, por ello, fundamentales para el método de integración MC; también lo son, como indica el método de la transformada inversa, fundamentales para

generar muestras de distribuciones de probabilidad no uniformes. Muestrear una distribución de probabilidad más complicada puede tener un objetivo más allá de la integración, y estar más relacionado con conocer la distribución de probabilidad de los parámetros de un modelo. Los métodos MCMC suelen estar más enfocados a esta segunda tarea, como veremos en la Sección 7.

Los lenguajes de programación modernos tienen incorporados generadores de números pseudoaleatorios (ver la Figura 1). Por esta razón se les suele asumir como algo dado, sin embargo para aplicaciones específicas es una buena práctica analizar su procedimiento de generación para evitar errores estadísticos [14].

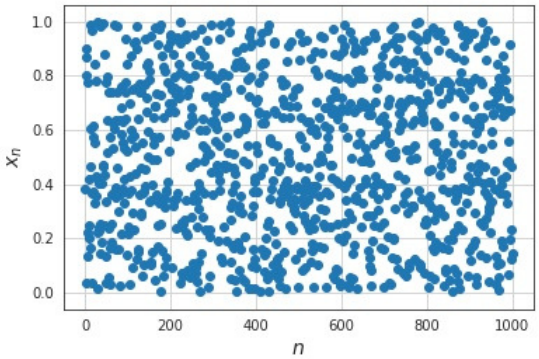


Figura 1: Representación gráfica de una muestra de 1000 números aleatorios en el intervalo $(0, 1)$ generados en Python.
Fuente: Elaboración propia.

5. Método de Monte Carlo ordinario

Como ya se dijo en la Sección 3, la idea central del método de integración Monte Carlo es la siguiente: si deseamos calcular la integral $I = \int_a^b g(x)dx$ para alguna función $g(x)$ difícil de integrar analíticamente, podemos hacer la siguiente sustitución:

$$I = \int_a^b g(x)dx = \int_a^b w(x)f(x)dx,$$

con $w(x) = g(x)(b-a)$ y $f(x) = \frac{1}{(b-a)}$. Por tanto, la integral transformada puede interpretarse como un múltiplo del valor esperado de $g(x)$ bajo la distribución uniforme:

$$\mathbb{E}(g(X)) = \frac{I}{(b-a)}, \quad \text{con} \quad X \sim Unif(a, b). \tag{5.1}$$

Si tomamos una muestra de números pseudoaleatorios uniformemente distribuidos, digamos $X_1, ..., X_N \sim Unif(a, b)$, por la ley de los grandes números

(Ecuación 3.2), para N suficientemente grande tendremos que, con probabilidad 1:

$$\frac{1}{N}\sum_{i=1}^{N} g(X_i) \to \mathbb{E}(g(X)) = \frac{I}{(b-a)}. \tag{5.2}$$

Ejemplo 5.1. *Para ilustrar el método de integración MC calculemos*

$$\int_0^1 \sqrt{arctan(x)}dx,$$

que no es integrable por métodos analíticos [24]. Programas orientados al cálculo científico, como Maple y Mathematica, no dan una solución analítica de la integral anterior. Se puede estimar esta integral numéricamente de varias maneras: por ejemplo, mediante sumas de Riemann, implementando el MC de acuerdo a la Ecuación (5.2) o implementando un MC en la región bidimensional que acota la función $f(x) = \sqrt{arctan(x)}$ *en el plano cartesiano.*

Con sumas de Riemann la aproximación numérica a la integral se puede elaborar con base en una partición regular, $b_0 = 0, b_1, b_2, ..., b_n = 1$*, del intervalo de integración, según la conocida fórmula del Cálculo:*

$$I = \frac{1}{n}\{g(b_1) + ... + g(b_n)\}, \tag{5.3}$$

donde $g(x) = \sqrt{arctan(x)}$ *y* $b_i = i/n$ *para* $i = 1, 2, ..., n$.

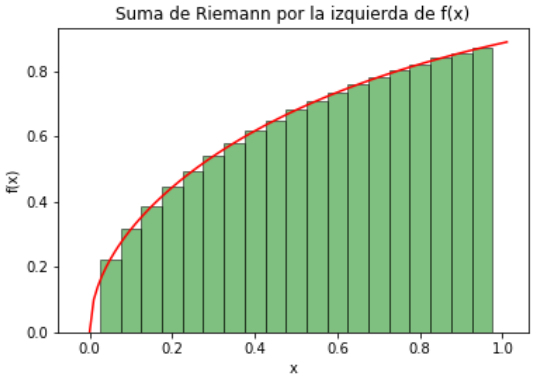


Figura 2: Sumas de Riemann para el área bajo la curva $f(x) = \sqrt{arctan(x)}$. Fuente: Elaboración propia.

La Figura 2 muestra el esquema gráfico de esta aproximación.

El MC basado en la ley de los grandes números, Ecuación (5.2), suele llamarse método MC de la Media Muestral. La aproximación numérica se obtiene de la siguiente expresión:

$$I = \frac{1}{n}\{g(x_1) + ... + g(x_n)\}, \tag{5.4}$$

donde $g(x) = \sqrt{arctan(x)}$ *y* $x_i \sim Unif(0,1)$. *Las Ecuaciones (5.3) y (5.4) tienen aspectos muy similares, sin embargo su interpretación es muy distinta, pues en el primer caso los valores de la variable representan puntos igualmente espaciados en el intervalo de integración* $(0,1)$ *mientras que en el segundo caso los valores de la variable consisten en una muestra de números pseudoaleatorios uniformes para el mismo intervalo* $(0,1)$*; entonces, de la Ecuación (5.1):*

$$I = (b-a)\mathbb{E}[g(X)].$$

Y un estimador insesgado (ver Apéndice) de esta integral es:

$$\theta_1 = (b-a)\frac{1}{n}\Sigma_{i=1}^{n} g(X_i) \;\; con \;\; X_i \sim Unif(0,1), \tag{5.5}$$

el cual tiene una varianza dada por:

$$Var(\theta_1) = \frac{1}{n}\left[(b-a)\int_a^b g^2(x)dx - I^2\right].$$

Entonces este método puede implementarse con el Algoritmo MC de la Media Muestral, presentado en la Tabla 4, para estimar $\int_a^b g(x)dx$ *[27].*

Fuente: [27].

Monte Carlo de la Media Muestral
Paso 1: Generar una secuencia $\{U_i\}_{i=1}^n \sim Unif(0,1)$.
Paso 2: Calcular $X_i = a + U_i(b-a)$.
Paso 3: Calcular $g(X_i)$, para $i = 1, ..., n$.
Paso 4: Estimar I calculando la media muestral θ_1 (Ecuación 5.5).

Tabla 4: Algoritmo 5.1.

La Figura 3 muestra el histograma de los resultados de esta integración MC repetida 1000 veces. Para hacer evidente la distribución estadística subyacente a este experimento, en este caso hemos tomado un valor pequeño para el tamaño de las muestras de números pseudoaleatorios $(n = 50)$.

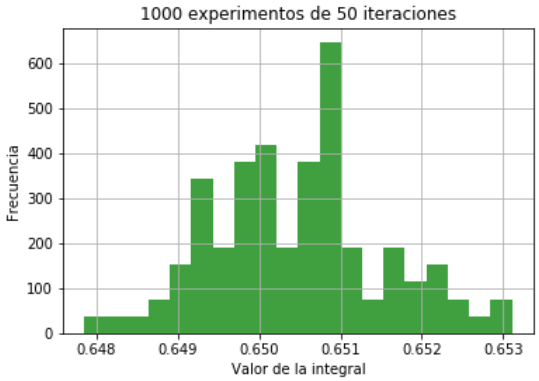


Figura 3: Histograma de resultados para la aproximación de $\int_0^1 \sqrt{arctan(x)}dx$ por el método MC de la Media Muestral.
Fuente: Elaboración propia.

Obsérvese que el histograma sugiere una distribución normal para los resultados de este experimento, lo que está de acuerdo con el teorema central del límite (Ecuación 3.3).

Nótese que la ley fuerte de los grandes números lleva a pensar que en la estimación numérica expresada por la Ecuación (5.4) el error, la diferencia entre I y μ, inversamente proporcional a $\sqrt{n}$, puede prácticamente anularse si n se hace suficientemente grande, y esta idea parece respaldada por el teorema central del límite; sin embargo, esto no se verifica en la práctica porque el método de MC trabaja no sobre números pseudoaletorios sino sobre transformaciones de ellos y los números pseudoaleatorios tienen períodos definitivamente finitos, por tanto suele ser necesario recurrir a diversas estrategias estadísticas para estimar los errores de aproximación y los intervalos de confianza [4].

El siguiente método MC es el más difundido porque tiene una interpretación geométrica, lo cual incrementa su valor pedagógico. Suele ser llamado método MC de Acierto y Error. Consiste en interpretar la integral como el área bajo la curva en una región rectangular, en nuestro ejemplo: $\{0 \leq x \leq 1\} \times \{0 \leq g(x) \leq c\}$, para algún valor c que acote la función en todo el intervalo de integración, y generar n puntos aleatorios (x, y) sobre dicho rectángulo. Luego, la probabilidad de acertar a la región bajo la curva es:

$$p = \frac{\int_a^b g(x)dx}{c(b-a)} = \frac{I}{c(b-a)}.$$

El parámetro p puede ser estimado de:

$$\hat{p} = \frac{n_a}{n},$$

donde n_a es el número de puntos bajo o sobre la curva, es decir, los aciertos del

muestreo. Y la integral puede ser, a su vez, estimada por:

$$I \approx \theta_2 = c(b-a)\frac{n_a}{n}, \tag{5.6}$$

Como la generación de cada uno de los puntos aleatorios es independiente de los otros, este procedimiento define un experimento de Bernoulli con probabilidad p de acertar; entonces θ_2 es un estimador insesgado de I porque:

$$\mathbb{E}(\theta_2) = c(b-a)\mathbb{E}(\frac{n_a}{n}) = pc(b-a) = I.$$

En este caso la varianza es:

$$Var(\theta_2) = \frac{1}{n}\Big[c(b-a) - I\Big],$$

mientras que la desviación estándar es:

$$\sigma_{\theta_2} = \frac{\sqrt{I(c(b-a)-I)}}{\sqrt{n}},$$

y puede observarse que la precisión del estimador es de orden $n^{-\frac{1}{2}}$.

Cuando n se hace suficientemente grande y se aplica el teorema central del límite a la variable $\hat{\theta} = \frac{\theta_2 - I}{\sigma_{\theta_2}}$ se tiene que el intervalo de confianza (ver Apéndice) con nivel $1-2\alpha$ para I es [27]:

$$\theta_2 \pm z_\alpha \frac{\sqrt{\hat{p}(1-\hat{p})}(b-a)c}{\sqrt{n}},$$

donde $z_\alpha = \phi^{-1}(\alpha)$ es la inversa de la distribución normal. Este método se puede implementar de acuerdo al Algoritmo MC de Acierto y Error, mostrado en la Tabla 5, para estimar $\int_a^b g(x)dx$ [27]:

Fuente: [27].

Monte Carlo de Acierto y Error
Paso 1: Generar una secuencia $\{U_i\}_{i=1}^{2n} \sim Unif(0,1)$.
Paso 2: Conformar pares (U_i, U_{n+i}), para $i = 1, ..., n$.
Paso 3: Calcular $X_i = a + U_i(b-a)$, para $i = 1, ..., n$.
Paso 4: Calcular $g(X_i)$, para $i = 1, ..., n$.
Paso 5: Contar los aciertos n_a tales que $g(X_i) > cU_{n+i}$.
Paso 6: Estimar I calculando la media muestral θ_2 (Ecuación 5.6).

Tabla 5: Algoritmo 5.2.

La Figura 4 muestra el resultado de este algoritmo para la función $g(x) = \sqrt{arctan(x)}$. Los tres métodos descritos se programaron en Python para tres valores de n y se obtuvieron los resultados registrados en la Tabla 6. Los códigos están disponibles en [8].

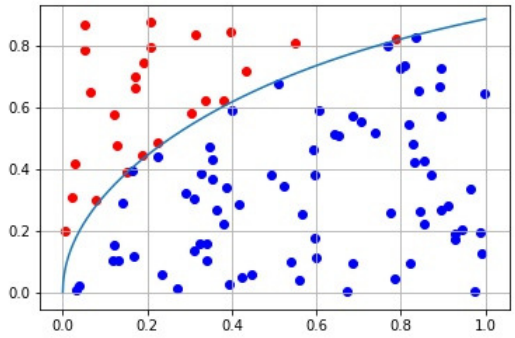


Figura 4: Estimación de $\int_0^1 \sqrt{arctan(x)}dx$ por el método MC de Acierto y Error. El área bajo la curva se estima como la razón del número de puntos bajo la curva (en azul) al número total de puntos (rojos más azules).
Fuente: Elaboración propia.

Nota: Los códigos están disponibles en [8].
Fuente: Elaboración propia.

Método	$n = 1000$	$n = 10000$	$n = 100000$
Sumas de Riemann	0.62937	0.62978	0.62982
MC media muestral	0.62560	0.63297	0.62989
MC acierto y error	0.62756	0.63001	0.62946

Tabla 6: Comparación de tres aproximaciones distintas de $\int_0^1 \sqrt{arctan(x)}dx$.

6. Cadenas de Markov

La fuente principal de esta sección es el libro de Wasserman [31], y su objetivo es brindar alguna familiaridad con las definiciones y los teoremas sobre cadenas de Markov más útiles en los métodos MCMC.

Definición 6.1. *Un proceso estocástico es una colección de variables aleatorias* $\{X_t : t \in T\}$*, donde* X_t *toma valores de un espacio de estados,* $\mathcal{X}$*, indexados por el conjunto* T*, llamado tiempo, que puede ser discreto o continuo.*

Definición 6.2. *Un proceso estocástico* $\{X_n : n \in T\}$ *es una cadena de Markov si*

$$P(X_n = x|X_0, ..., X_{n-1}) = P(X_n = x|X_{n-1}),$$

para todo n *y todo* x*.*

Es decir, en una cadena de Markov la probabilidad de alcanzar el valor X_n depende únicamente del valor previo X_{n-1}. Esta condición se conoce como la *propiedad de Markov* y representa un modelo probabilístico sencillo pero de gran potencia.

Definición 6.3. *Las cantidades*

$$p_{ij} \equiv P(X_{n+1} = j | X_n = i),$$

se llaman probabilidades de transición y la matriz $\mathbf{P}$ *cuyas entradas* (i, j) *son los elementos* p_{ij} *se llama matriz de transición.*

Una cadena de Markov puede imaginarse como una sucesión de valores de una variable aleatoria:

$$X_0 \to X_1 \to X_2 \to ... \to X_n \to ...,$$

en la cual cada valor de la variable tiene un solo padre. La cantidades clave de las cadenas de Markov son las probabilidades de ir de un estado de la cadena (un valor de la variable) al siguiente.

Definición 6.4. *Un estado se llama recurrente si*

$$P(X_n = i \text{ para algún } n \geq 1 | X_0 = i) = 1.$$

En otro caso, se dice que el estado es transitorio.

Las características importantes al generar cadenas de Markov, y por ende para aplicarlas a los métodos MC, son las referentes a su convergencia.

Definición 6.5. *Una cadena de Markov es irreducible si para cada par de estados* i *y* j *hay una probabilidad positiva de que el proceso transite del estado* i *al estado* j*.*

Definición 6.6. *Supongamos que* $X_0 = i$*. El tiempo de recurrencia se define como*

$$T_{ij} = min\{n > 0 : X_n = j\},$$

si X_n *siempre retorna al estado* i*, de lo contrario se define* $T_{ij} = \infty$*.*

Definición 6.7. *El tiempo medio de recurrencia de un estado* i *se define como*

$$m_i = \mathbb{E}(T_{ij}) = \Sigma_n n f_{ii}(n),$$

donde $f_{ij} = P(X_1 \neq j, X_2 \neq j, ..., X_{n-1} \neq j, X_n = j | X_0 = i)$*.*

Definición 6.8. *Un estado recurrente se llama nulo si* $m_i = \infty$*, en otro caso es llamado no nulo o positivo.*

Definición 6.9. *Una cadena de Markov se denomina aperiódica si no oscila regularmente entre valores de la variable.*

Definición 6.10. *Un estado se llama ergódico si es recurrente, no nulo y aperiódico. Y se dice que una cadena en sí misma es ergódica si todos sus estados son ergódicos.*

Ahora, un vector $\pi = (\pi_i : i \in \mathcal{X})$ cuyas entradas no negativas sumen 1 puede pensarse como un función discreta de probabilidad. El siguiente par de definiciones son las más relevantes para los fines de los métodos MCMC.

Definición 6.11. *π es una distribución estacionaria si $\pi = \pi\mathbf{P}$. Y se dice que la cadena tiene una distribución límite si*

$$\mathbf{P}^n \to \left\{ \begin{array}{c} \pi \\ \pi \\ \vdots \\ \pi \end{array} \right\},$$

para algún π, esto es, si $\pi_j = lim_{n\to\infty} P_{ij}^n$ existe y es independiente de i.

Definición 6.12. *Se dice que π satisface la condición del balance detallado si*

$$\pi_i p_{ij} = p_{ji}\pi_j. \tag{6.1}$$

Y los teoremas principales sobre convergencia de cadenas de Markov son los siguientes:

Teorema 6.1. *Una cadena de Markov ergódica e irreducible tiene una única distribución estacionaria π. Si g es una función acotada, entonces, con probabilidad 1:*

$$lim_{N\to\infty} \frac{1}{N} \Sigma_{n=1}^{N} g(X_n) \to \mathbb{E}_\pi(g) \equiv \Sigma_j g(j)\pi_j. \tag{6.2}$$

Teorema 6.2. *Si π satisface el balance detallado, entonces π es una distribución estacionaria.*

7. Monte Carlo vía Cadenas de Markov

La Ecuación (6.2), el teorema ergódico, es una generalización de la ley fuerte de los grandes números para cadenas de Markov. Es natural, entonces, pensar en utilizar estas últimas para calcular numéricamente integrales, en analogía al procedimiento mostrado en la Sección 5. La diferencia radica en que ahora al calcular numéricamente la integral:

$$\int_a^b g(x)f(x)dx,$$

la distribución de probabilidad $f(x)$ puede ser mucho más compleja y no importa su forma, las cadenas de Markov nos ayudarán a encontrar una muestra de valores X_i que sigan tal distribución. Esto se logra construyendo una cadena de Markov que tenga a $f(x)$ como distribución límite. Entonces, después de m iteraciones, una vez que la cadena de Markov ha convergido, se pueden desechar o *quemar* los primeros m valores de la cadena (en inglés se llama *burn in* a este segmento de la cadena) y el resto puede usarse para estimar el valor de expectación (y por ende la integral deseada), mediante la aproximación [7]:

$$\mathbb{E}_f(g(X)) = \frac{1}{n-m} \sum_{i=m+1}^{n} g(X_i) \text{ con } X_i \sim f(x).$$

Entonces, para implementar el método MCMC necesitamos un algoritmo para generar cadenas de Markov que tengan como distribución límite a $f(x)$ y también un criterio para determinar que las cadenas de Markov han convergido y que nos guíe en la elección del valor m en la quema de los primeros valores de la cadena.

El procedimiento más popular para generar cadenas de Markov es el algoritmo Metropolis-Hastings [7, 26], que tiene el siguiente esquema general mostrado en la Tabla 7:

Fuente: [7].

Metropolis-Hastings
Paso 1: Inicializar X_0, $t = 0$.
Paso 2: Repetir {
Generar un candidato $Y \sim q(\cdot\|X_t)$
Generar $U \sim U(0,1)$
Si $U \leq \alpha(X_t, Y)$, tomar $X_{t+1} = Y$
otro caso, tomar $X_{t+1} = X_t$
Incrementar t
}

Tabla 7: Algoritmo 7.1.

La función $q(\cdot|X_t)$ es una distribución que ya se sepa simular (por su simetría, suele elegirse la distribución normal para facilitar el proceso de simulación), $\pi(\cdot)$ es la función objetivo, es decir $f(\cdot)$, y $\alpha(X_t, Y)$, la probabilidad de aceptación, está definida como:

$$\alpha(X,Y) = min\left(1, \frac{\pi(Y)q(X|Y)}{\pi(X)q(Y|X)}\right).$$

El algoritmo Metropolis-Hastings es una versión general para el MCMC, de la cual pueden derivarse otros algoritmos de muestreo. El algoritmo original de

Metropolis (1953), por ejemplo, se recupera al considerar solamente distribuciones de probabilidades simétricas, donde $q(Y|X) = q(X|Y)$, de modo que la probabilidad de aceptación se reduce a:

$$\alpha(X,Y) = min\left(1, \frac{\pi(Y)}{\pi(X)}\right).$$

Y como caso especial de este algoritmo están las caminatas aleatorias de Metropolis, donde $q(Y|X) = q(|X - Y|)$.

En todos los casos, es necesario notar que la cantidad $\alpha(X,Y)$ es fundamental para la construcción de las cadenas de Markov. La elección de la forma de $\alpha(X,Y)$, que es muy sencilla, garantiza que $\pi(\cdot)$ satisface la condición del balance detallado (Ecuación 6.1), y por tanto que es $\pi(\cdot)$, en sí misma, la distribución estacionaria de la cadena de Markov.

Por otro lado, un criterio muy utilizado para analizar la convergencia de las cadenas de Markov es la Prueba de Gelman-Rubin, que consiste en los siguientes pasos:

1. Generar $M \geq 2$ cadenas de Markov para generar las variables aleatorias θ que se desea muestrear, cada una con $2N$ iteraciones, partiendo de distintos puntos iniciales.

2. Descartar las primeras N iteraciones de cada cadena.

3. Calcular la varianza de cada cadena:

$$W = \frac{1}{M}\Sigma_{j=1}^{M} s_j^2,$$

   donde s^2 es la varianza de cada cadena, calculada tras descartar las primeras N iteraciones. Y calcular la varianza entre las cadenas:

$$B = \frac{N}{M-1}\Sigma_{j=1}^{M} s_j^2.$$

4. Calcular la varianza estimada de θ:

$$var(\theta) = (1 - \frac{1}{N})W + \frac{1}{N}B,$$

5. Calcular el factor:

$$R = \sqrt{\frac{var(\theta)}{W}}.$$

   Se acepta que las cadenas han convergido cuando $0.97 < R < 1.03$.

Los métodos MCMC son, pues, generadores de números pseudoaleatorios utilizados para simular funciones de probabilidad complicadas.

Aunque pueden utilizarse para integración numérica (que en ciertas aplicaciones puede ser algo muy complejo debido a la alta dimensión de las funciones a integrar), es más común que los MCMC se usen como métodos de exploración de las distribuciones estadísticas en sí mismas, generalmente para determinar sus valores óptimos globales (máximos y mínimos) o sus promedios. A continuación mostramos un ejemplo donde se explora una distribución estadística. En [9] se aborda el mismo tipo de análisis, enfatizando su relación con otras técnicas computacionales.

Ejemplo 7.1. *Para ilustrar la aplicación de los métodos MCMC vamos a considerar un ejemplo de ajuste de parámetros de un modelo lineal: dado un conjunto de datos encontraremos los dos parámetros (pendiente y ordenada al origen) del modelo lineal que mejor se ajusten estadísticamente para describirlos. El ejercicio consta de los siguientes pasos (el código utilizado está disponible en [8]):*

1. *Planteamos un modelo lineal $y = m_0x + b_0$, como prueba del método, eligiendo los parámetros $m_0 = 2$ y $b_0 = 3$. La prueba consiste en generar datos sintéticos a partir de este modelo y luego hacer inferencia estadística sobre esos datos, mediante el MCMC, para comprobar que podemos recuperar los parámetros elegidos.*

2. *Generamos datos sintéticos agregando una dispersión aleatoria a los valores de y: $y = m_0x + b_0 + \epsilon$, donde $\epsilon \sim N(0, \sigma)$, y también barras de error aleatorias hasta un tamaño elegido. El resultado se muestra en la Figura 5.*

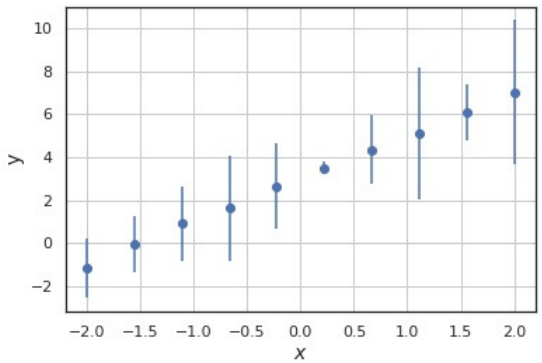


Figura 5: Datos sintéticos generados agregando dispersión y barras de error aleatorias a puntos de la recta $y = 2x + 3$.
Fuente: Elaboración propia.

3. *Ajustamos los datos al modelo valiéndonos de una prueba Chi-cuadrada, que es una generalización de la técnica de mínimos cuadrados [21]. Dado*

un conjunto de datos D_i y un modelo de estos datos, $y(x_i|\theta)$, que depende de un conjunto de parámetros θ, el ajuste del modelo está determinado por aquellos valores de los parámetros que minimizan la distribución:

$$\chi^2 \equiv \Sigma_{ij}(D_i - y_i(x_i|\theta))Q_{ij}(D_j - y_j(x_j|\theta)),$$

donde Q denota la inversa de la matriz de covarianza de los datos. En nuestro ejemplo, tenemos 10 datos que deseamos ajustar a un modelo lineal $y = mx + b$ con dos parámetros: m y b.

4. *Mediante un algoritmo Metropolis-Hastings realizamos un muestreo de la función χ^2 definida por los datos sintéticos y, tras una prueba de convergencia Gelman-Rubin, se determinan los valores de m y b que minimizan la función χ^2. La Figura 6 muestra las cadenas resultantes.*

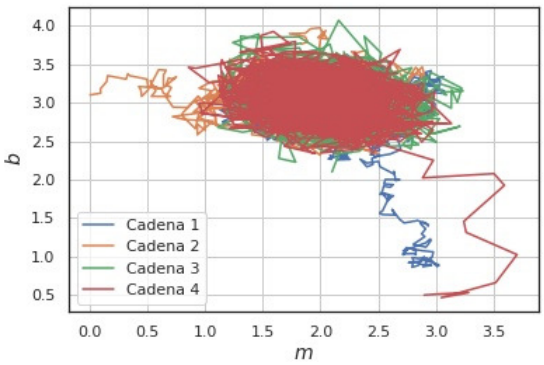


Figura 6: Cadenas de Markov en dos dimensiones ejecutadas para ajustar los datos sintéticos de la Figura 5 a un modelo lineal $y = mx + b$.
Fuente: Elaboración propia.

5. *Construimos las elipses de confianza (ver Apéndice) a partir de las cadenas generadas, las cuales se muestran en la Figura 7. Las elipses de confianza son regiones del espacio de parámetros (que en nuestro ejemplo es 2-dimensional) alrededor del punto de valores medios que contienen un porcentaje dado de la distribución de probabilidad. Es común utilizar la desviación estándar para cuantificar los niveles de confianza, de manera que las regiones de 1σ, 2σ y 3σ corresponden al $68.3\,\%$, $95.4\,\%$ y $99.5\,\%$, respectivamente. El mejor ajuste de los parámetros se obtuvo para $\theta_1 = m = 2.040$ y $\theta_2 = b = 3.029$ como valores que minimizan la función χ^2. Mientras que los valores medios son: $m = 1.951422$, $b = 3.027087$ con desviaciones medias 0.389912 y 0.280212, respectivamente, y matriz de covarianza:*

$$\begin{pmatrix} 0.1520314 & -0.00954926 \\ -0.00954926 & 0.07851859 \end{pmatrix}$$

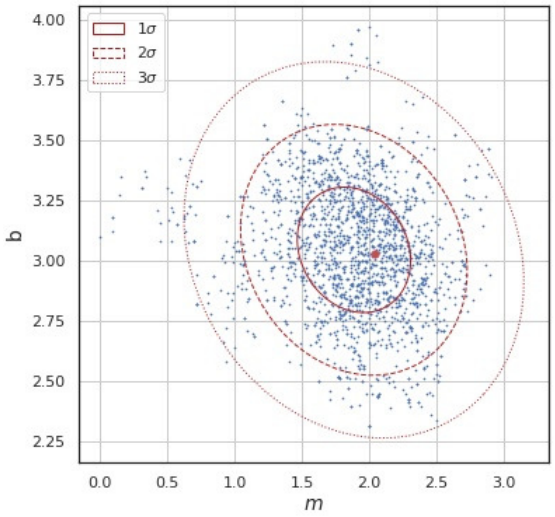


Figura 7: Elipses de confianza centradas en los valores medios $(1.951, 3.027)$ en el espacio de parámetros asociadas al ajuste de datos sintéticos de la Figura 5 para un modelo lineal $y = mx + b$. El punto $(2.040, 3.029)$ corresponde al mínimo de la distribución chi-cuadrada encontrado por el muestreo MCMC. Fuente: Elaboración propia.

8. Conclusión

Presentamos los métodos MC y MCMC de manera unificada, como resultantes de la aplicación de una misma idea esencial: la generación de números pseudoaleatorios para obtener muestras de distribuciones estadísticas. No obstante, mientras los métodos MC pueden requerir algoritmos muy particulares para simular distribuciones específicas, los métodos MCMC cuentan con algoritmos de aplicación mucho más general: entre ellos, el Metropolis-Hastings es uno de los más populares.

Las principales aplicaciones, integración numérica y exploración del espacio de una distribución estadística, se ilustran mediante ejemplos no triviales acompañados de códigos en Python que pueden implementarse y modificarse fácilmente, por lo cual éstos constituyen un primer paso en una exploración que puede iniciar el lector motivado.

Los autores sugerimos que se use este trabajo como material didáctico en cursos donde se estudien los métodos Monte Carlo, en vista de que la presentación unificada de estos métodos y la implementación práctica propuestos pueden contribuir a mejorar la comprensión del tema.

Agradecimientos

Este trabajo fue parcialmente apoyado por: CONACYT, ICF-UNAM, CICATA - Legaria del Instituto Politécnico Nacional (IPN) y el proyecto SIP20210500 del IPN. R.G.S. agradece el apoyo de las becas COFAA, EDI y Estancia Sabática del IPN, así como al proyecto FORDECYT-PRONACES-CONACYT No. CF-MG-2558591. J.A.V. agradece el apoyo proporcionado a los proyectos FOSEC SEP-CONACYT Investigación Básica A1-S-21925 y UNAM-DGAPA-PAPIIT IA102219. R.M.E. e I.G. V. agradecen el apoyo de las becas de posgrado CONACYT.

Referencias

[1] Audren, B. (2015). Monte Python. En: `https://baudren.github.io/`.

[2] Coss, R. (2017). *Simulación: Un enfoque práctico*, Limusa, México (México).

[3] Eckhardt, R. (1987). Stan Ulam, John Von Neumann, and the Monte Carlo Method. *Los Alamos Science*, **Special Issue(15)**, 131-136.

[4] Fishman, G. S. (1996). *Monte Carlo: concepts, algorithms and applications*, Springer-Verlag, New York (USA).

[5] Gardner, M. (1980). *Carnaval Matemático*, Alianza, Madrid (España).

[6] Geman, S., y Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-6(6)**, 721-741.

[7] Gilks, W. R., Richardson, S., y Spiegelhalter, D. J. (1998). *Markov Chain Monte Carlo in practice*, Chapman & Hall/CRC, Boca Raton (USA).

[8] Gómez, I. (2019). IntroMCMC. En: `www.github.com/igomezv/`.

[9] Gómez, I., Medel, R., Vázquez, J. A., y García, R. (2019). Una Aplicación de las Redes Neuronales Artificiales en la Cosmología. *Komputer Sapiens*, **Año XI**(Vol. II), 12-16. En: `www.smia.mx/komputersapiens/`.

[10] Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, **57(1)**, 97-109.

[11] Hitchcock, D. B. (2003). A History of the Metropolis-Hastings Algorithm. *The American Statistician*, **57(4)**, 254-257.

[12] Johansen, A. M., y Evers, L. (2007). *Monte Carlo Methods: Lecture Notes*, University of Bristol. En: `www.warwick.ac.uk`.

[13] Kirkpatrick, S., Gelatt Jr., C. D., y Vecchi, M. P. (1983). Optimization by Simulated Annealing. *Science*, **220(4598)**, 671-680.

[14] Kneusel, R. T. (2018). *Random Numbers and Computers*, Springer (USA).

[15] Krishnamoorthy, K. (2006). *Handbook of statistical distributions with applications*, Chapman & Hall/CRC, Boca Raton (USA).

[16] McGrayne, S. B. (2012). *La teoría que nunca murió*, Crítica, Barcelona (España).

[17] Metropolis, N., y Ulam, S. (1952). A Property of Randomness of an Arithmetical function. *AECU*, **Technical Report(2038)**.

[18] Metropolis, N. (1987). The Beginning of the Monte Carlo Method. *Los Alamos Science*, **Special Issue(15)**, 125-130.

[19] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., y Teller, A. H. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, **21(6)**, 1087-1092.

[20] Metropolis, N., y Ulam, S. (1949). The Monte Carlo Method. *Journal of the American Statistical Association*, **44(247)**, 335-341.

[21] Press, W. H., Teukolsky, S. A., Vetterling, W. T., y Flannery, B. P. (2007). *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, New York (USA).

[22] PyMC3 (2019) En: `https://docs.pymc.io/`.

[23] Richey, M. (2010). The Evolution of Markov Chain Monte Carlo Methods. *The American Mathematical Monthly*, **117(5)**, 383-413.

[24] Riley, K. F., Hobson, M. P., y Bence, S. J. (2006). *Mathematical Methods for Physics and Engineering: A Comprehensive Guide*, Cambridge University Press, New York (USA).

[25] Robert, C., y Casella, G. (2011). A Short History of Markov Chain Monte Carlo: Subjective Recollections from Incomplete Data. *Statistical Science*, **26(1)**, 102-115.

[26] Ross, S. M. (1999). *Simulación*, Prentice Hall, México (México).

[27] Rubinstein, R. Y., y Kroese, D. P. (2008). *Simulation and The Monte Carlo Method*, John Wiley & Sons, New Jersey (USA).

[28] Thomas, A. (2010). Overview. En: `www.openbugs.net/`.

[29] Ulam, S. M. (2002). *Aventuras de un Matemático*, Nivola, Madrid (España).

[30] Upton, G., y Cook, I. (2002). *A Dictionary of Statistics*, Oxford University Press, New York (United States of America).

[31] Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*, Springer, New York (USA).

Acerca de los autores

Ricardo Medel Esquivel es licenciado en física y matemáticas. Actualmente desarrolla un proyecto de análisis de datos cosmológicos en el programa de doctorado del Centro de Investigación en Ciencia Aplicada y Tecnología Avanzada Unidad Legaria (CICATA-Legaria) del Instituto Politécnico Nacional (IPN).

Isidro Gómez Vargas es licenciado en física y matemáticas por el IPN, ubicado en la Ciudad de México. En la actualidad, realiza su doctorado en Tecnología Avanzada (también en el IPN), donde estudia la estimación de parámetros cosmológicos mediante técnicas estadísticas y computacionales.

J. Alberto Vázquez es investigador del Instituto de Ciencias Físicas (ICF) de la UNAM. Previo a su incorporación al ICF, fue catedrático CONACYT asociado al CINVESTAV, IPN. Estancia postdoctoral en el Laboratorio Nacional Brookhaven, NY. Doctorado y Maestría otorgados por el Kavli Institute for Cosmology de la Universidad de Cambridge, UK. Su investigación se enfoca en la exploración y análisis de observaciones cosmológicas, en el estudio de modelos de Energía oscura e Inflacionarios, a través de la estadística Bayesiana y el computo científico.

Ricardo García Salcedo es profesor titular en el Departamento de Física Educativa en el CICATA-Legaria del IPN. Licenciado y maestro en Ciencias por la Universidad Autónoma del Estado de México, doctorado en Ciencias por el Depto. de Física del CINVESTAV-IPN. Su investigación se centra en la cosmología, especialmente los modelos de energía oscura y TIC en la enseñanza de la Física.

A. Apéndice

Esta sección es un brevísimo repaso de los fundamentos de la teoría de probabilidad; está basado en [26], que recomendamos ampliamente a los lectores interesados en los fundamentos de la simulación.

Asociado a cada evento A de un espacio muestral S hay un número $P(A)$, la probabilidad de ocurrencia del evento A, que satisface los siguientes axiomas:

1. $0 \leq P(A) \leq 1$.
2. $P(S) = 1$.
3. $P\Big(\bigcup_{i=1}^{n} A_i\Big) = \sum_{i=1}^{n} P(A), \qquad n = 1, 2, ..., \infty$.

La probabilidad conjunta de los eventos A y B (la probabilidad de que sucedan simultáneamente) se denota como $P(AB)$. La probabilidad condicional del evento A dado el evento B se escribe $P(A|B)$, y se cumple la relación:

$$P(A|B) = \frac{P(AB)}{P(B)}.$$

Y cuando los eventos A y B son independientes:

$$P(AB) = P(A)P(B).$$

Definición A.1. *El valor esperado, o media, de una variable aleatoria discreta X que asume alguno de los valores posibles $x_1, x_2, ..., x_n$, es*

$$\mathbb{E}[X] := \sum_i x_i P\{X = x_i\}.$$

Definición A.2. *El valor esperado, o media, para una variable continua X que puede tomar todos los valores x en $(-\infty, \infty)$, con función densidad de probabilidad f, se define como:*

$$\mathbb{E}[X] := \int_{-\infty}^{\infty} x f(x) dx.$$

Teorema A.1. *El valor esperado y la varianza (Definición 3.2) tienen las siguientes propiedades*

1. $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$.
2. $\mathbb{E}[X_1 + X_2] = \mathbb{E}[X_1] + \mathbb{E}[X_2]$.
3. $Var(X) = \mathbb{E}[X^2] - \mu^2$.
4. $Var(aX + b) = a^2 Var(X)$.

donde a y b son constantes y X_1 y X_2 son variables aleatorias.

Teorema A.2. *La varianza y la covarianza (Definición 3.3) satisfacen las siguientes propiedades*

1. $Cov(X, Y) = \mathbb{E}[XY] + \mathbb{E}[X]\mathbb{E}[Y]$.
2. $Var(X \pm Y) = Var(X) + Var(Y) \pm 2Cov(X, Y)$.

3. $Var(X) \geq 0$.

Así, cuando X y Y son independientes, $Cov(X, Y) = 0$ y se tiene

$$Var(X + Y) = Var(X) + Var(Y).$$

Algunos de los resultados más importantes de la teoría de probabilidad son las desigualdades, pues con base en ellas se pueden hacer inferencias y determinar intervalos de confianza. Aquí nos interesan principalmente los resultados para variables continuas. Uno muy útil es el siguiente.

Teorema A.3. ***Desigualdad de Markov.***
Si X toma solo valores no negativos, entonces para cualquier $a > 0$

$$P\{X \geq a\} \leq \frac{\mathbb{E}[X]}{a}. \tag{A.1}$$

Demostración. Si X es una variable aleatoria no negativa, entonces

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x)dx = \int_{0}^{\infty} xf(x)dx,$$

y separando en dos segmentos el intervalo de integración:

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^a xf(x)dx + \int_a^\infty xf(x)dx \\
&\geq \int_a^\infty xf(x)dx \\
&\geq \int_a^\infty af(x)dx, \text{ pues } xf(x) \geq af(x) \text{ si } x \geq a \\
&= a\int_a^\infty f(x)dx = aP\{X \geq a\},
\end{aligned}$$

de donde se obtiene el resultado deseado. ■

La desigualdad de Markov permite probar fácilmente el siguiente resultado, de mayor utilidad práctica.

Teorema A.4. ***Desigualdad de Chebyshev.***
Si X es una variable aleatoria con media μ y varianza σ^2, entonces para cualquier $k > 0$

$$P\{|X - \mu| \geq k\sigma\} \leq \frac{1}{k^2}. \tag{A.2}$$

Demostración. Se obtiene aplicando la desigualdad de Markov (Ecuación A.1) a la variable no negativa $\frac{(X-\mu)^2}{\sigma^2}$, que tiene media igual a 1, como puede comprobarse a partir de la definición de la varianza y las propiedades del valor esperado (Teorema A.1). ■

La desigualdad de Chebyshev (Ecuación A.2) permite, considerando la hipótesis de que la varianza siempre tiene valores finitos, arribar a los resultados más importantes para el método de Monte Carlo ordinario: las leyes de los grandes números (Teoremas 3.1 y 3.2).

Definición A.3. *Un estimador es una función de la muestra usada para estimar un parámetro desconocido de la población. El sesgo de un estimador es la diferencia entre el valor esperado del estimador y el verdadero valor del parámetro a estimar. Un estimador es insesgado cuando su valor esperado es igual al parámetro a estimar.*

Definición A.4. *Se llama intervalo de confianza a un par de números en la recta real entre los cuales se estima que estará cierto valor desconocido con un determinado nivel de confianza. Las elipses de confianza generalizan este concepto al caso bidimensional.*

Boletín de Estadística e Investigación Operativa
Vol. 37, No. 1, Marzo 2021, pp. 75-81

Opiniones sobre la profesión

Matemáticas y Estadística. Al César lo que es del Cesar...

Roberto Behar Gutiérrez
Escuela de Estadística, Universidad del Valle, Cali (Colombia)
✉roberto.behar@correounivalle.edu.co

Pere Grima and Xavier Tort-Martorell
Departamento de Estadística e Investigación Operativa
Universitat Politècnica de Catalunya - BarcelonaTech
✉pere.grima@upc.edu, ✉xavier.tort@upc.edu

Abstract

This article discusses whether statistics is a part of mathematics, as some teachers –of statistics in college and of mathematics in secondary education– often take for granted. It can be said that the important thing is that it is explained well –that the question is not if the cat is white or black, it is that it hunts mice– and therefore that this discussion is not very interesting. However, it is very relevant because the vision that one has of statistics influences how it is taught, the approach that is given, and what is prioritized... and, also relevant, in which group appropriates the discipline giving it its imprint and personality. The paper highlights the importance of being clear that statistics goes far beyond the mathematical methods it uses.

1. La estadística, ¿es una parte de las matemáticas?

En la educación secundaria suele decirse que la estadística es un tema que no siempre da tiempo de ver porque está al final del libro. Al final del libro de matemáticas, no del libro de matemáticas y estadística. Esto ya crea la idea de que la estadística es una parte de las matemáticas y de que está bajo la competencia del profesor de esta asignatura.

Fuera del ámbito académico, si en una librería uno quiere ver qué libros de estadística tienen deberá buscar en las estanterías donde están los libros de matemáticas y quizá allí –entre los de álgebra, cálculo y geometría– encuentre

alguno. Aunque ahora seguramente es más habitual buscar en internet. Si entramos en Amazon y buscamos libros por categorías, los de estadística están dentro de la categoría de matemáticas, por supuesto.

La UNESCO ha definido una nomenclatura para los campos de las ciencias y la tecnología. Matemáticas es uno de los campos y de ahí cuelgan Lógica, Álgebra, Análisis... y Estadística. Esta revista incluye un código AMS (de la American Mathematical Society) en cada artículo para ayudar a clasificarlo. Ahí se encuentra una exhaustiva relación de las áreas de las matemáticas. Una de ellas es –como no– la estadística. En fin ¿quién duda de que la estadística es una parte de las matemáticas?

Pues parece que no somos nosotros los únicos. En la universidad de dos de los autores tenemos una Facultad de Matemáticas y Estadística. Con lo meticulosos que son los matemáticos con la terminología y la notación que utilizan no dejarían que la facultad se llamara "de Matemáticas y Estadística" si consideraran que la estadística es una parte de las matemáticas. No hay facultades de "Matemáticas y Álgebra" ni de "Matemáticas y Geometría". Sí las hay de "Física y Química" o de "Geografía e Historia", que son disciplinas afines, pero no es una parte de la otra.

2. Estadística versus matemáticas

El pensamiento matemático es deductivo. Se parte de unos axiomas y mediante la lógica se deducen unos teoremas que se cumplen siempre. Este proceso deductivo persigue la resolución de problemas que se sitúan en el ámbito de los modelos abstractos, de lo teórico, y su resolución exige prestar mucha atención a la notación que se usa y a la aplicación de las reglas, propiedades y otros teoremas demostrados previamente.

El pensamiento estadístico es inductivo. Se parte de unos datos y a partir de ellos se estiman características de la población de la que provienen. El cómo seleccionar y evaluar la calidad de esos datos también forma parte del problema. Mientras que las matemáticas buscan encontrar soluciones exactas en el mundo de lo simbólico, en estadística estamos intentando buscar soluciones aproximadas pero útiles conociendo una medida de la bondad de la aproximación. En matemáticas un solo caso que no se cumpla ya es suficiente para declarar que una proposición es falsa. En estadística sabemos que el hecho de que un fumador de cajetilla diaria llegue a los 90 años en buenas condiciones no invalida la teoría de que el tabaco perjudica la salud.

La estadística sirve para responder preguntas en el terreno de la investigación empírica, preguntas del tipo: ¿Qué tipo de resina da mejores resultados para depurar el agua? ¿qué principio activo es más eficaz para curar una enfermedad? ¿Qué porcentaje de ciudadanos está de acuerdo con la política del gobierno? Estas preguntas no se pueden responder desde las matemáticas. Hay que hacer

un experimento o una encuesta y sabemos que las conclusiones que se extraigan no serán un teorema matemático. Si se repite el experimento/encuesta saldrán otros resultados, pero tenemos herramientas matemáticas que, bajo ciertos supuestos, nos permiten responder a las preguntas planteadas informando también sobre una medida de la confianza con la que damos nuestras respuestas.

Solo con las herramientas e instrumentos de las matemáticas no siempre es posible resolver problemas de la realidad fáctica, de la misma forma que un gran físico, que sabe las leyes de la mecánica, la termodinámica y la electricidad, no garantiza que cuando se le estropee el coche lo sepa arreglar. La aplicación de la estadística está más relacionada con la forma de hacer un buen diagnóstico sobre lo que le pasa al coche para dar una buena solución, que en aumentar la teoría sobre la mecánica o la termodinámica.

Pero desde luego, la estadística tiene en la matemática una de sus herramientas más útiles. Por ejemplo, la teoría de la probabilidad –uno de los pilares de la estadística– se desarrolla íntegramente con el proceso deductivo de la matemática. La teoría de la probabilidad sí es una parte de las matemáticas. Necesitamos la teoría de las distribuciones de probabilidad para calcular probabilidades en el terreno de lo práctico. Si no existiera la matemática que se ha desarrollado para demostrar los teoremas que usamos en estadística (ley de los grandes números, teorema central del límite, teorema de Glivenko-Cantelli...) poca cosa se podría hacer.

Desde luego que matemáticas y estadística no son disciplinas independientes, como no lo son las matemáticas y la física. Pero sus objetivos y metodología son distintos y el problema no está en cómo clasificamos la estadística sino en qué enfoque le damos y cómo la enseñamos.

3. Cuando los medios se confunden con los fines

Todos tenemos claro que una cosa es hacer cálculos (sumas, divisiones, raíces cuadradas) y otra son las matemáticas, que van mucho más allá tanto en sus métodos como en sus objetivos. De la misma forma, construir tablas de frecuencias, calcular medias, medianas y modas o representar datos en un diagrama de sectores no es hacer estadística. Poner el énfasis en unos procedimientos que parece que solo se hacen por hacer, sin tener claro cuál es el problema que queremos resolver o a qué pregunta relevante queremos responder, seguramente se percibe como una tarea tediosa y de escasa utilidad práctica.

El objetivo de la estadística es aumentar nuestro conocimiento sobre el mundo que nos rodea, ya sea en el terreno de la ciencia, la sociología o los negocios, a partir de la observación y el análisis de la realidad de una forma inteligente y objetiva. La estadística estudia cómo recoger datos o cómo valorar si aquellos de que disponemos tienen una calidad suficiente. También estudia cómo analizarlos para obtener la información que permita responder a les preguntas que

nos planteamos.

Naturalmente también hay que saber lo que es una tabla de frecuencias, pero construirla no es nunca el objetivo de la estadística. Poner el énfasis solo en los métodos o en las herramientas que se usan en el análisis de los datos sin el contexto del tipo de problemas –reales, cercanos y prácticos–, que se pueden resolver da una imagen equivocada de la utilidad y de las posibilidades de la estadística. Seguramente, más que una asignatura o una disciplina independiente, la estadística puede ser vista como una competencia transversal –igual que hacer cálculos, o saber expresarse bien– que está presente en todas las asignaturas. La estadística tiene que ver con el análisis objetivo de la realidad para crear teorías e interpretaciones que se dan como buenas (o como las más razonables con la información disponible) aunque es posible que –con más datos– las conclusiones sean otras. Está en la esencia del método científico.

En la universidad, en las asignaturas de estadística, seguramente lo más importante es transmitir interés por el valor de los datos, por cómo recogerlos para que sean útiles, por saber valorar su calidad. Claro que para analizar los resultados utilizaremos modelos matemáticos, o los que convengan, pero el protagonismo, el interés genuino, no debería estar centrado en los modelos sino en los problemas que cada uno en su campo tiene interés en resolver.

A veces se administra una sobredosis de matemáticas en los primeros años de los estudios de grado, también en el Grado de Estadística. Esto puede ser debido al deseo de que los estudiantes tengan una buena base matemática, y también –porque no decirlo– al interés de los diferentes departamentos por meter horas de clase de sus asignaturas con criterios que no priorizan las necesidades y el interés del estudiante sino el copar horas de docencia para poder contratar o estabilizar nuevos profesores, o consolidar otros con un puesto precario.

Esto hace que estudiantes con curiosidad e interés por la estadística se encuentren con que en los primeros cursos lo que menos se estudia es estadística y, sin embargo, deben tomar varias asignaturas de matemáticas, tema que –en principio– no les despierta especial interés. Es como si a un joven al que le gusta el futbol se apunta a un grado para ser futbolista pero el primer año las asignaturas tienen que ver con: darle vueltas al campo corriendo, hacer abdominales, ejercicios de estiramiento... Estamos todos de acuerdo en que estas "asignaturas" son útiles para ser futbolista, pero seguramente el estudiante se sentirá decepcionado ya que no es eso lo que a él le gusta. Y otro problema, siguiendo con el símil, sería que los profesores de abdominales o de dar vueltas al campo no supiesen explicar la conexión de esas disciplinas con el futbol porque a esos profesores el futbol nunca les ha interesado ni lo han practicado.

¿No sería más adecuado que los jóvenes que quieren ser futbolistas dediquen –ya desde el principio– una buena parte del tiempo a jugar al futbol? Seguro que ellos mismos se darían cuenta de que la resistencia física es muy importante

para aguantar en buenas condiciones hasta el final del partido, y de que los ejercicios de estiramientos son fundamentales para evitar lesiones, y seguramente los harían encantados y motivados, una vez vista su necesidad e importancia.

4. Estadística e Informática

También la informática parece estar desdibujando el terreno de la estadística, incluso con la aparición de grados de lo que se ha dado en llamar "Ciencia de Datos" o directamente "Data Science" sin traducir del inglés.

Hasta hace unos años, uno de los mantras que se repetían en nuestras clases de estadística era que los datos siempre son un bien escaso, que siempre quisiéramos tener más pero tenemos que conformarnos con los que tenemos y administrarlos lo mejor que podemos. En muchos campos esto ha dejado de ser cierto. Ahora recoger datos puede ser relativamente fácil, y lo difícil es analizarlos para obtener de ellos información relevante.

Los expertos en gestionar y manejar grandes volúmenes de datos son los informáticos. Los informáticos son también expertos en crear algoritmos (eso es lo suyo) y hablan de redes neuronales, *machine learning*, *deep learnig*, *random forest*,... y otras técnicas que se salen de la ortodoxia clásica de las técnicas de modelado. Las fortalezas de ese nuevo punto de vista son la gestión de grandes bases de datos, la visualización de la información y las predicciones. Existen paquetes de software especializados en estos menesteres. Pensamos que nosotros, los estadísticos clásicos (por llamarnos de alguna manera) también debemos meternos en este tema ya que, sin duda, tiene muchas posibilidades.

Sin embargo, esos algoritmos que tan bien funcionan para hacer predicciones, no sirven para crear modelos explicativos. No sirven para detectar las causas y así poder actuar sobre ellas. Parece que todo el ámbito del diseño de experimentos queda fuera del *Data Science*. Por otro lado, desde la informática se insiste en encontrar la manera de obtener información de grandes bases de datos que muchas veces están mal estructurados, tomados sin mucho cuidado y muchos de ellos de dudosa veracidad, aunque quizá aun así se puede encontrar algún patrón o tendencia que puede ser útil conocer.

En muchos casos tenemos la sensación de que ese esfuerzo en sacar información de donde no la hay (o hay muy poca) sería mejor orientarlo a planificar la recogida de los datos con el rigor, la meticulosidad y la estructura requeridos para obtener –a partir de ese momento– la información que se busca de una manera mucho más fácil.

5. Ingeniería Estadística

También está apareciendo una visión de la estadística como instrumento para abordar situaciones complejas. Es la llamada ingeniería estadística (www.isea-

change.org).

Por analogía podemos pensar en la química, que trata de las moléculas, enlaces, reacciones... y otra cosa es la ingeniería química, que utiliza esos conocimientos, pero su objetivo es –por ejemplo– construir una instalación para la fabricación de determinado tipo de plástico. O la física, que trata de electrones, campos magnéticos o las ecuaciones de Maxwell y otra cosa es la ingeniería eléctrica, que se preocupa del funcionamiento de las locomotoras del AVE.

En la práctica, no nos encontramos con problemas enunciados de forma clara –tal como nosotros se los ponemos a nuestros estudiantes– sino situaciones enredadas, embrollos que hay que aclarar. No nos encontramos ante un problema de regresión o de comparación de dos tratamientos. Cuando el problema está claramente planteado ya es mucho más fácil de resolver o de saber que no tiene solución.

Reconocer, identificar y acotar esos embrollos es una habilidad que conviene cultivar. Para convertirlos en un problema de enunciado conviene aplicar una metodología dentro de la cual habrá que identificar cual es la herramienta o técnica estadística que resulta más adecuada.

6. Conclusión

Reducir la estadística a una parte de las matemáticas, centrando el interés solo en los modelos que se utilizan y en los métodos matemáticos que se aplican, es una visión muy limitada de sus posibilidades y de sus ámbitos de aplicación. A la hora de clasificarla la podemos poner donde convenga, o donde es tradicional hacerlo, pero eso no nos debe hace perder de vista que nos encontramos ante una disciplina diferente, con un entorno y unas prioridades diferentes.

Olvidar estos aspectos en la enseñanza y, especialmente en la enseñanza secundaria, dejar la estadística como algo que solo interesa en la clase de matemáticas y que está bajo la jurisdicción exclusiva del profesor de esa asignatura, es un error que puede conducir a que nuestros científicos, técnicos, y población en general, tenga una visión deformada y antipática de lo que son herramientas muy útiles para el avance del conocimiento y una parte fundamental del método científico.

Acerca de los autores

Roberto Behar Gutiérrez es profesor de la Escuela de Estadística en la Universidad del Valle, en Cali, Colombia, donde ha sido director de la carrera de Estadística. Después de graduarse en Ciencias de la Educación, especialidad Matemáticas, realizó una Maestría en Estadística en el Colegio de Postgraduados de México y su tesis doctoral en la Universidad Politécnica de Cataluña. Ha escrito artículos y realizado numerosas conferencias sobre la enseñanza de la estadística, una de sus pasiones.

Pere Grima es profesor en la Universidad Politécnica de Cataluña. Imparte sus clases en la Escuela de Ingeniería Industrial de Barcelona y en la Facultad de Matemáticas y Estadística, donde ha sido Jefe de Estudios de Estadística. Sus áreas de investigación están relacionadas con la estadística industrial y la gestión de la calidad, pero también dedica parte de su tiempo a temas relacionados con la divulgación de la estadística.

Xavier Tort-Martorell es catedrático de Estadística en la Universidad Politécnica de Cataluña y director del departamento de Estadística e Investigación Operativa. Sus áreas de interés son la estadística industrial y la consultoría sobre temas estadísticos en general. Ha escrito numerosos artículos en revistas especializadas y actualmente es editor asociado de algunas de ellas. Ha sido presidente de de la European Network for Business and Industrial Statistics (ENBIS) en 2012 y 2013.

http://www.seio.es/BEIO