MONOGRAFÍAS FME

Estadística en acción

Qué es y para qué sirve la estadística a través de casos prácticos basados en proyectos final de carrera

Editado por Pere Grima

UPC
Facultat de Matemàtiques i Estadística
UNIVERSITAT POLITÈCNICA DE CATALUNYA

MONOGRAFÍAS FME

Estadística en acción

MONOGRAFÍAS FME

Estadística en acción

Qué es y para qué sirve la estadística a través de casos prácticos basados en proyectos final de carrera

FACULTAT DE MATEMÀTIQUES I ESTADÍSTICA
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Traducido y revisado del original en catalán
con la colaboración de Oriol Camps Lorente

FACULTAT DE MATEMÀTIQUES I ESTADÍSTICA
Pau Gargallo, 5 — 08028 Barcelona

Fotocomposición, impresión y encuadernación
BARCELONA DIGITAL
c/ Rosselló, 77. 08029 Barcelona
Tel. 93-3638610

ISBN: 978-84-7653-216-4***OJO***

Depósito legal: B-48384 -2008***OJO***

Printed in Spain

Contenido

V. *Nuevos medicamentos y mejora del sistema sanitario*

VI. *Prensa y televisión*

Índice de autores

Laura Marí (A): Análisis de patrones y tendencias en las votaciones del festival de Eurovisión

Lídia Montero (D): Aplicación de técnicas estadísticas al estudio del arte pictórico de los siglos XV al XIX

Maria Montón (A): La estadística en la prensa. Estudio crítico

Román Peñas (A): Previsión de la duración de les etapas de la *Vuelta Ciclista a España*

Elisabeth Peralta (A): El diseño emocional: cómo crear productos que sean atractivos (y un ejemplo sobre zumos de frutas)

Núria Porta (D): Análisis de la eficacia de un nuevo fármaco contra el SIDA

Xavier Puig (A): Análisis de la mortalidad por tumores malignos de mama y de estómago en Cataluña

Jordi Real (A): Comportamiento de la demanda de urgencias hospitalarias y factores meteorológicos asociados

Alexandre Riba (D): Reducción del número de defectos en el proceso de fabricación de un componente de automóvil | Previsión de la duración de les etapas de la *Vuelta Ciclista a España*

Lourdes Rodero (A) (D): Buscando productos más limpios y eficaces para luchar contra las plagas de los árboles frutales | Estudio de mercado para definir las características de una nueva variedad de galletas de chocolate

Miquel Romero (A): Aplicación de técnicas estadísticas al estudio del arte pictórico de los siglos XV al XIX

Josep Antón Sánchez (T) (D): Influencia de las condiciones climáticas en el crecimiento del pino silvestre en Cataluña | Comportamiento de la demanda de urgencias hospitalarias y factores meteorológicos asociados

Sara Solanes (A): Estudio de mercado para definir las características de una nueva variedad de galletas de chocolate

Elisenda Vila (A): Formación de precios en el mercado eléctrico: mejores estrategias para compradores y vendedores

Aureli Tobías (D): Comportamiento de la demanda de urgencias hospitalarias y factores meteorológicos asociados

(A): Autor. (D): Director. (T): Tutor.

Cuando el director de un proyecto no es profesor de la UPC (esto suele pasar cuando el proyecto se realiza en una empresa o institución en la que el estudiante realiza prácticas como becario) un profesor de la UPC actúa como Tutor y vela por el rigor académico y formal del proyecto.

Presentación

La percepción superficial que a menudo se tiene de la estadística no se corresponde con su relevancia en el mundo actual, ni con la demanda que hay de profesionales de esta disciplina. Es por lo tanto un motivo de satisfacción para la Facultad de Matemáticas y Estadística (FME) de la Universidad Politécnica de Cataluña poder ofrecer una obra que puede contribuir sustancialmente a cambiar esta percepción.

Esta contribución está en sintonía con el deseo del centro de ofrecer al alumnado la posibilidad de seguir alguno de sus estudios. En el caso de la estadística, la FME inició, hace quince años, los de la Diplomatura de Estadística, los cuales posteriormente se ampliaron con los de la Licenciatura en Ciencias y Técnicas Estadísticas y los del Máster en Estadística e Investigación Operativa.

La estadística tiene un papel fundamental en el avance del conocimiento en casi todos los ámbitos, desde la investigación sobre nuevos medicamentos hasta los estudios sobre salud pública, desde los estudios de mercado hasta el control de calidad, desde el análisis del medio ambiente hasta estudios sobre literatura, música o pintura.

En esta obra se describen dieciséis casos tratados en proyectos de final de carrera que dan una visión general de las posibilidades de aplicación de la estadística. Esperamos que el profesorado de secundaria lo encontrará útil para inspirar sus tareas y, habiendo procurado usar un lenguaje lo menos técnico posible, también esperamos que lo será para un amplio público, incluyendo a los estudiantes de bachillerato.

Sebastià Xambó
Decano de l'FME
Octubre de 2008

¡Esto es estadística!

Todos los que han cursado alguna de las titulaciones de estadística en la Facultad de Matemáticas y Estadística de la UPC (Diplomatura, Licenciatura o Máster) han realizado un proyecto final de carrera como parte de su formación. En este proyecto, que también se presenta en público, deben demostrar que son capaces de realizar un estudio estadístico como los que aquí se presentan.

Muchos de estos trabajos son excelentes[1], reflejan muy bien qué es y para qué sirve la estadística, y de algunos de ellos se derivan directamente publicaciones en revistas científicas, y otros forman parte de líneas de investigación más generales que han dado importantes resultados. Realmente es una lástima que estos trabajos duerman el sueño de los justos en la biblioteca de la Facultad, y solo sean consultados, de vez en cuando, por algún estudiante que cuando va a preparar su proyecto quiere saber si los anexos se deben poner antes o después de la bibliografía.

Por otro lado, muchas personas, y entre ellas los estudiantes de bachillerato que deben escoger los estudios que seguirán en la universidad, tienen una visión muy reducida ("medias, encuestas y porcentajes") o incluso deformada, de qué es y para qué sirve la estadística.

La estadística estudia cómo recoger datos (¿cuántos?, ¿de qué forma?) y cómo analizarlos para obtener la información que permita responder a les preguntas que nos planteamos, ya sea en el ámbito de la medicina, de la sociología, de la psicología o de la ciencia en general. Se trata de avanzar en el conocimiento a partir de la observación y el análisis de la realidad, de una forma inteligente y objetiva. Es la esencia del método científico.

Una buena forma de verlo es a través de estos proyectos. Lo que aquí se incluye no son sus resúmenes, sino una explicación que quiere dar una visión general de cuál es el contexto del problema, qué datos son necesarios, cómo se obtienen, y unas pinceladas de cómo se han analizado y qué conclusiones se han obtenido. La extensión de cada una de estas partes no es proporcional a su extensión o a la importancia que se le da en el proyecto. En algunos de ellos, el proyecto trata sobre cuáles son los mejores modelos o las mejores técnicas para analizar los datos pero aquí no se entra en estos detalles (difíciles de entender sino se es especialista) y se ha dado sólo una visión muy general de estos aspectos.

Los proyectos se han seleccionado sin un criterio muy definido, pero por un lado hemos querido que salgan representados diferentes ámbitos de aplicación, algunos

[1] Con datos: El 77% de los proyectos final de carrera de la Diplomatura i la Licenciatura de Estadística presentados en los 5 últimos cursos (2002-03 al 2007-08) han obtenido una calificación de 9 o superior.

que podríamos denominar clásicos, como la medicina o el marketing y otros más sorprendentes, como la pintura, la música y la literatura. Hay proyectos sobre temas que preocupan, como el que trata del análisis de la eficacia de un nuevo fármaco contra el SIDA, y otros sobre temas más lúdicos, como los patrones que siguen las votaciones en el Festival de Eurovisión, pero siempre con el mismo rigor en la recogida y el tratamiento de los datos.

Evidentemente, el mérito de este trabajo corresponde a sus autores (y directores y tutores), pero si alguna explicación ha quedado a un nivel demasiado elevado y no se entiende muy bien, o si otras han quedado demasiado descafeinadas y no se llega a apreciar el trabajo o las aportaciones que se han realizado, esto es culpa nuestra, que hemos forzado un estilo, y hemos corregido y retocado algunos originales explicando las cosas como nosotros pensábamos que se debían explicar (naturalmente, al final los autores han dado el visto bueno, seguramente alguna vez pensando que no les quedaba más remedio…)

Cuando hablo de nosotros, en este apartado de culpas, me refiero a mí mismo, pero también (nadie quiere todas las culpas para él solo) a María Montón, una estudiante que mientras ha realizado el último curso de la Licenciatura, ha colaborado como becaria en las tareas de coordinación de esta publicación y también ha aportado iniciativas y propuestas que han mejorado mucho su contenido.

Las tres titulaciones de estadística de las que hablaba al principio están inmersas en un proceso de cambio. El Máster, que nació como una titulación de la UPC, se ha convertido este curso 2008-2009 en un Máster Interuniversitario impartido conjuntamente con la Universidad de Barcelona (UB). Y con el proceso de adaptación al Espacio Europeo de Educación Superior (acuerdos de Bolonia) desaparecen los estudios de diplomatura y licenciatura, convirtiéndose en el los nuevos estudios de grado. Estamos trabajando en el diseño de un grado de estadística organizado e impartido conjuntamente con la UB, donde, al igual que en el Máster, se aprovechen las áreas de especialidad y los recursos de las dos universidades, creando una titulación de referencia, con una orientación aplicada y que dé unas bases sólidas para formar excelentes profesionales.

Esperamos que estos textos, que abordan problemas típicos donde la estadística tiene un papel protagonista, ayuden a entender mejor sus posibilidades y cuáles son los campos donde se puede aplicar.

Pere Grima
Vicedecano Jefe de Estudios de Estadística de la FME
Octubre de 2008

Estudio de la literatura, la música y la pintura

1

Análisis estadístico del estilo literario. Discusión sobre la autoría del *Tirant lo Blanc*

Proyecto realizado por: **Susanna Cabos Ruiz**
Dirigido por: **Josep Ginebra Molins**

El objetivo del análisis estadístico del estilo literario es buscar características cuantificables del estilo de un texto –de las que seguramente el mismo autor no es consciente– y aprovecharlas para compararlo con el estilo de otros textos. Normalmente el objetivo final es determinar su autoría, y en este caso las características escogidas deben ser propias del autor y no del género o de la época en que el texto fue escrito, para no confundir el efecto del autor con el que puedan tener otros factores.

El análisis del estilo literario se puede realizar de diferentes formas ya que las herramientas de resolución del problema todavía no están del todo estandarizadas. Hasta la realización de este proyecto prácticamente toda la investigación se había realizado sobre aplicaciones a textos ingleses, alemanes y, en menor medida, franceses y clásicos.

Con la llegada de los escáneres y los programas de reconocimiento óptico de caracteres (OCR) se produjo una explosión en la aplicación de estas técnicas y uno de los objetivos del trabajo que aquí se comenta era adaptarlas al catalán y aplicarlas a la cuestión de la autoría del "Tirant lo Blanc" una de las obras más relevantes de la literatura catalana, escrita en el siglo XV.

Estilometría y autoría

Los problemas de determinación de la autoría de un texto de escritor desconocido o disputado se pueden clasificar en tres grandes familias

1. Casos en los que se disputa la autoría de un texto entre dos o más autores y disponemos de textos de estos autores que son "comparables" con el texto de paternidad discutida. Un ejemplo de esta categoría es el estudio de los *Federalist Papers*, una colección de artículos escritos entre 1787 y 1788 por Hamilton, Jay y Madison para apoyar la aprobación de la constitución americana y publicados todos ellos bajo un mismo seudónimo. Se conoce quien es el autor de 65 de los 77 artículos, pero los otros doce se pueden atribuir tanto en Madison como a Hamilton. Utilizando otros artículos escritos por Madison y por Hamilton, se llega a la conclusión que los artículos disputados se deberían atribuir a Madison.

2. Problemas donde hay un candidato a ser el autor, del cual se dispone textos "comparables" al texto disputado, y candidatos alternativos de los cuales no se dispone de textos reconocidos, o bien se dispone de textos que no sirven para comparar, porque son de generaciones o épocas diferentes. Un ejemplo de este problema son los análisis para investigar si las obras de Shakespeare podrían haber sido escritas por Bacon. A partir del análisis estadístico de ciertas características del lenguaje se encuentran diferencias entre los estilos de Shakespeare y Bacon, pero estas diferencias parecen ser debidas a que se trata de dos géneros diferentes, la poesía y la prosa, y no a la posible existencia de dos autores. Un ejemplo de la literatura catalana es el planteado por Josep Guia en torno al *Tirant lo Blanc* y la hipotética autoría de Joan Roís de Corella. En este caso, el trabajo del estadístico es determinar si la variabilidad de las características estilísticas estudiadas es mayor entre textos escritos por dos autores diferentes que entre textos escritos por un mismo autor.

3. Por último, un tercer grupo de problemas son los que tratan de estudiar la homogeneidad en el estilo de un texto, es decir, intentar detectar cambios dentro del mismo texto. Estos cambios podrían ser debidos a cambios de autoría, a una evolución temporal en el estilo del autor, o a un acontecimiento puntual en su biografía. Dos ejemplos clásicos son los planteados en torno a la hipotética existencia de más de un autor en el Libro de Isaías y en las Epístolas de San Pablo. En la literatura catalana, el caso más estudiado es la posible existencia de una frontera de estilo dentro del *Tirant lo Blanc*. Estos tipos de problema son los de más difícil aproximación y los menos tratados en la literatura estadística.

Cuantificación del estilo literario

La manera más habitual de cuantificar el estilo de un texto es utilizar las frecuencias de uso de unidades lingüísticas fáciles de identificar, fáciles de contar y difíciles de controlar conscientemente por el autor. La identificación y medida de estas unidades debe estar muy bien definida, de forma que no haya ambigüedades en su recuento y éste se pueda llevar a cabo de forma automática utilizando un ordenador con el software adecuado. Algunas unidades que se pueden utilizar son:

- *Número de letras por palabra*
  La ventaja de esta unidad es su facilidad de conteo de forma automática, pero algunos especialistas consideran que es poco fiable cuando se comparan dos autores porque las diferencias a menudo son mayores entre géneros que entre autores.

- *Número de sílabas por palabra*
  Es más difícil de contar automáticamente que el número de letras, especialmente en lengua catalana, que tiene reglas para diptongos y diéresis complicadas de reconocer. Los estudios publicados hasta la realización de este proyecto mostraban una fuerte correlación entre los resultados de esta medida y los de la anterior, por lo que se decidió utilizar sólo la primera.

- *Longitud de la frase*
  La longitud se suele medir a través del número de palabras por frase, y es habitual considerar como frase todo lo que acaba con un punto, un signo de interrogación o un signo de exclamación. Pero los textos medievales se escribían sin ningún tipo de puntuación, siendo esta introducida muy posteriormente por algún editor. Por tanto, esta unidad de cuantificación no será útil en nuestro caso.

- *Distribución de nombres, verbos, adjetivos,…*
  Otra posibilidad es comparar las proporciones de uso de nombres, verbos, adjetivos, preposiciones, conjunciones, artículos y otras partes del lenguaje. La gente cultivada utiliza más sustantivos, y una actitud más activa se traduce en un porcentaje de verbos más alta. El problema de esta variable es que no resulta fácil reconocer automáticamente la función de las palabras.

- *Frecuencia de uso de "palabras herramienta"*
  La proporción de uso de algunas palabras varía mucho en diferentes obras de un mismo autor, mientras que otras palabras presentan mucha estabilidad en todas sus obras. Para discriminar entre autores se necesitan palabras tales que su presencia sea tan independiente como se pueda del contexto, para no confundir el efecto autor con el efecto contexto. Estas palabras en francés se llaman "*mots*

outil", en inglés "*function words*" y en este proyecto se tradujeron al catalán como "*paraules eina*" de forma que aquí las denominamos “palabras herramienta”. En el problema de los *Federalist Papers* los dos candidatos eran bien conocidos y se pudieron elegir las palabras herramienta que discriminaban mejor los estilos de Hamilton y de Madison, partiendo de textos que se sabía quién de los dos había escrito.

- *Uso de una palabra entre dos posibles alternativas*
  Algunos autores compararon la frecuencia relativa de uso entre parejas de palabras, a menudo sinónimas. En inglés, por ejemplo, se ha comparado la proporción de uso de "*that*" respecto a "*this*", "*and the*" respeto a "*and*", "*to be*" respecto de "*to*", o bien "*not only*" respecto a "*not*". La utilidad de estas variables depende de la elección de las parejas a estudiar. En el caso de los *Federalist Papers*, se descubrió que Hamilton tenía tendencia a utilizar "*while*", mientras que Madison se decantaba por "*whilst*", y esta circunstancia se aprovechó para atribuir la autoría de los textos disputados. Sería bueno disponer de listas de parejas de palabras potencialmente útiles para estudios de estilo en catalán.

También se pueden utilizar otras formas más sofisticadas de cuantificación. Aunque en este proyecto no se utilizó, se puede caracterizar la riqueza y la diversidad del vocabulario de un texto o de un autor haciendo un inventario de todas las palabras utilizadas y contabilizando el número de ocurrencias de cada tipo, o modelando el orden de aparición del vocabulario y de otras unidades lingüísticas.

Obtención de los datos para el análisis del *Tirant lo Blanc*

Se desea aplicar el análisis estadístico del estilo literario al problema de autoría del *Tirant lo Blanc*, una novela de caballerías escrita en valenciano por Joanot Martorell (1414-1468). Se cree que la obra la acabó de escribir Martí Joan de Galba (?-1490), pero no se tiene seguridad sobre este hecho.

Elección de los textos

Un primer problema importante era decidir qué texto convenía utilizar para este estudio. La primera idea fue trabajar con la edición original de 1490, pero ésta edición no tiene ningún tipo de puntuación y, por tanto, no es viable el estudio de la longitud de las frases ya que no se puede determinar cuándo acaba una y empieza otra. Además, utilizar el texto original sin tener conocimiento del vocabulario de aquella época presentaba una gran dificultad para escoger los campos que compondrían la frecuencia de uso de las palabras herramienta. Por último, en el texto original una

misma palabra aparecía escrita de maneras muy diferentes debido a que los impresores transcribían los textos sin fijarse en la ortografía propia de la palabra. Por estas razones, al final fue escogida la edición aparecida en la colección "Mejores Obras de la Literatura Catalana" (MOLC) de Edicions 62, editada en 1983 por Martí de Riquer.

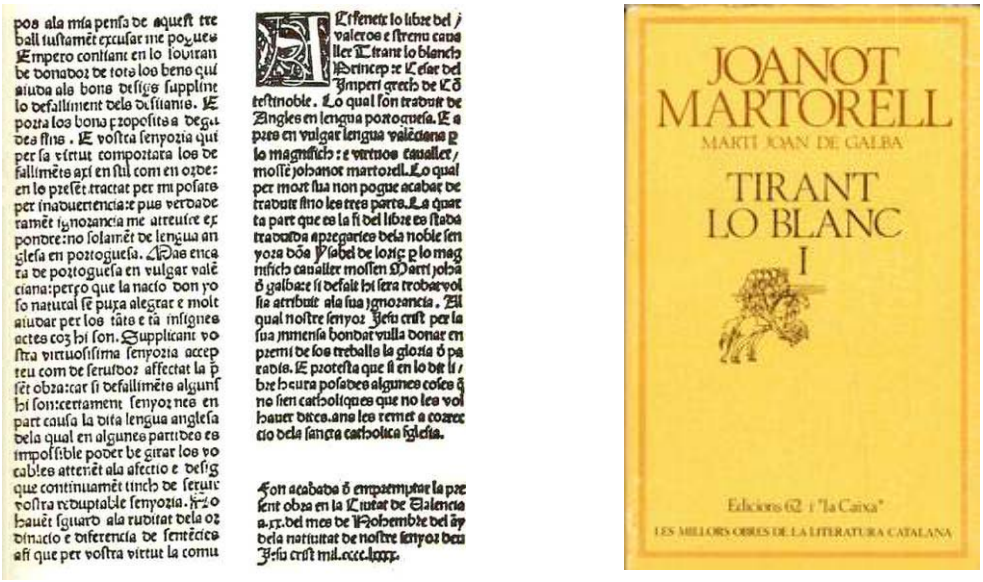


Figura 1: *Página introductoria de la edición original del Tirant lo Blanc (Fuente: Wikipedia)* y *portada de la edición que se ha utilizado en este proyecto, con tipografía actual.*

Para comparar el estilo del *Tirant lo Blanc* con la obra de Joan Roís de Corella (1435-1497) se escogió la obra en prosa religiosa *Història de Josef*, editada por R. Miquel i Planas en el año 1913. Posiblemente esta obra no tiene el lenguaje normalizado de acuerdo con la normativa actual, como en el caso de la edición utilizada del *Tirant lo Blanc,* y esto podía ocasionar problemas al interpretar diferencias entre resultados. Pero el profesor Josep Guia, reconocido experto en estos temas, nos sugirió utilizar esta obra, suponiendo que ésta sería una de las obras de Corella más plagiadas en el *Tirant*.

Otra dificultad que se tuvo que sortear fue el hecho de no disponer del texto en un soporte informático sobre el cual poder realizar los estudios estadísticos necesarios. Por tanto, se tuvieron que entrar muestras del texto vía escáner y basar en ellas el estudio.

En la edición utilizada de *Tirant lo Blanc*, Martí de Riquer clasificó los 487 capítulos de la obra en cinco partes. Como muestra de la obra, se introdujeron y trataron en soporte informático los diez bloques que corresponden a los inicios y finales de cada una de las cinco partes. Estos bloques vienen etiquetados como 1a, 1b, 2a, 2b, 3a, 3b,

4a, 4b, 5a y 5b, y están descritos en la Tabla 1. De los capítulos estudiados se eliminaron los títulos y todo lo que en la edición de las MOLC aparece en cursiva. Para comparar los estilos de *Tiran lo Blanc* y de *Història de Josef*, se agruparon las diez unidades del *Tirant* en tres grandes bloques (Ti_123, Ti_4567, Ti_890), tal como se indica en la misma Tabla 1. Además, se añadió un bloque con los capítulos en torno al 154, (Ti_fron) donde, según Josep Guía, acabaría la influencia de Joanot Martorell sobre la obra.

Tabla 1: *Descripción de las partes del Tirant lo Blanc que se han utilizado en este estudio*

Partes estudiadas	Bloques	Capítulos por bloque	Núm. ocurrencias	Núm. frases	Núm. ocurrencias	Núm. frases
Tirant_1	1a	1-6	4279	109	12449	334
	1b	87-97	3977	97		
	2a	98-99	4193	129		
Tirant_2	2b	111-114	4613	151	17504	549
	3a	116-119	4668	155		
	3b	292-295	4007	120		
	4a	296-300	4217	124		
Tirant_3	4b	388-395, 397, 399, 400	4167	112	13418	356
	5a	408-411	4613	109		
	5b	481-487	4638	135		
Tirant_4	F1	148, 149, 151	3874	160	12829	455
	F2	154	4297	130		
	F3	155, 156, 157	4658	166		

En total, las diferentes partes que se introdujeron representan aproximadamente un 15% del libro. El escáner de que se disponía (de funcionamiento lento y no muy moderno) realizaba la lectura de la grafía como un dibujo que se transformada a texto mediante un software de OCR ("*Optical Character Recognition*"). Un problema importante es que este programa de conversión no reconoce algunas grafías de la lengua catalana y fue necesaria una corrección manual de todo el texto introducido, trabajo al que se tuvieron que dedicar muchas horas.

Elección de las unidades analíticas

De las posibilidades comentadas como unidades para cuantificar el estilo literario se eligieron el número de letras por palabra, el número de palabras por frase, la frecuencia de aparición de ciertos signos de puntuación y la frecuencia de aparición de cada una de las palabras herramienta escogidas. La razón más importante para escoger estas unidades y no otras fue el hecho de que son fáciles de contar.

Durante el diseño del programa informático que realiza el recuento automático fue necesario tomar muchas decisiones de carácter operativo. Por ejemplo, se decidió no distinguir entre mayúsculas y minúsculas, ni entre palabras con acento o sin acento, de forma que dos palabras como "què" y "que" eran estudiadas como una sola forma: "que".

Antes de realizar el recuento del número de letras por palabra fue necesario definir claramente el término "palabra". Evidentemente, esta definición afecta también al recuento de la segunda variable: número de palabras por frase. Así pues se definió "palabra" como una secuencia de caracteres que finaliza en espacio en blanco, apóstrofo, guión o cualquiera de los signos de puntuación.

Medida de las unidades utilizadas. Programa informático

Una vez se tuvo el texto totalmente depurado, se elaboró un programa en lenguaje Pascal que hacía todo el recuento necesario para realizar el estudio. El programa era lo suficientemente flexible para que se pudiera adaptar a otras variables parecidas de forma sencilla y no tenía límites de longitud máxima ni mínima de texto.

Comparación de *Tirant lo Blanc* e *Història de Josef*

Los datos obtenidos para cada aspecto cuantificado se pueden presentar como en la Tabla 2, que muestra el número de palabras según su número de letras en cada uno de los 4 trozos estudiados de *Tirant lo Blanc* y en la *Història de Josef*.

Tabla 2: *Número de palabras según su número de letras en las cuatro partes estudiadas de 'Tirant lo Blanc' y de 'Història de Josef'*

Letras por palabra	Tirant_1	Tirant_2	Tirant_3	Tirant_4	Hist. Josef	Total
1 letra	1374	1820	1688	1347	1003	7232
2 letras	2761	3928	2757	2917	2588	14951
3 letras	2432	3588	2628	2683	2036	13367
4 letras	1326	1811	1358	1342	1286	7123
5 letras	1271	1873	1372	1228	1458	7202
6 letras	1264	1824	1344	1443	1451	7326
7 letras	714	988	671	702	997	4072
8 letras	590	768	650	525	830	3363
9 letras	366	498	502	429	533	2328
10 letras	203	221	292	117	262	1095
11 letras	102	86	84	60	116	448
Más de 11	46	99	72	36	83	336
Total	12449	17504	13418	12829	12643	68843

A partir de los datos de esta tabla se puede determinar si la distribución de las palabras según su número de letras en la *Història de Josef* es "significativamente" distinta de dicha distribución en los cuatro trozos del *Tirant lo Blanc*. En primer lugar hay que darse cuenta de que el número total de palabras en cada una de las 5 partes estudiadas es diferente y por tanto no se pueden comparar directamente estos valores. Por ejemplo, en la parte Tirant_1 aparecen 1.374 palabras de una sola letra y en Tirant_2 aparecen más: 1.820. Por otra parte, se observa que en Tirant_1 el número total de palabras es de 12.449 y, por lo tanto, la proporción de palabras de una sola letra es 0,110 (=1.374/12.449) mientras que en Tirant_2 la proporción de palabras de una sola letra es menor: 0,104 (=1.820/17.504).

Es útil presentar estas proporciones gráficamente, como en la Figura 2, en la cual se observa como los valores de la columna correspondiente al texto de Corella (*Història de Josef*) tienden a desmarcarse de los perfiles de los otros cuatro textos.

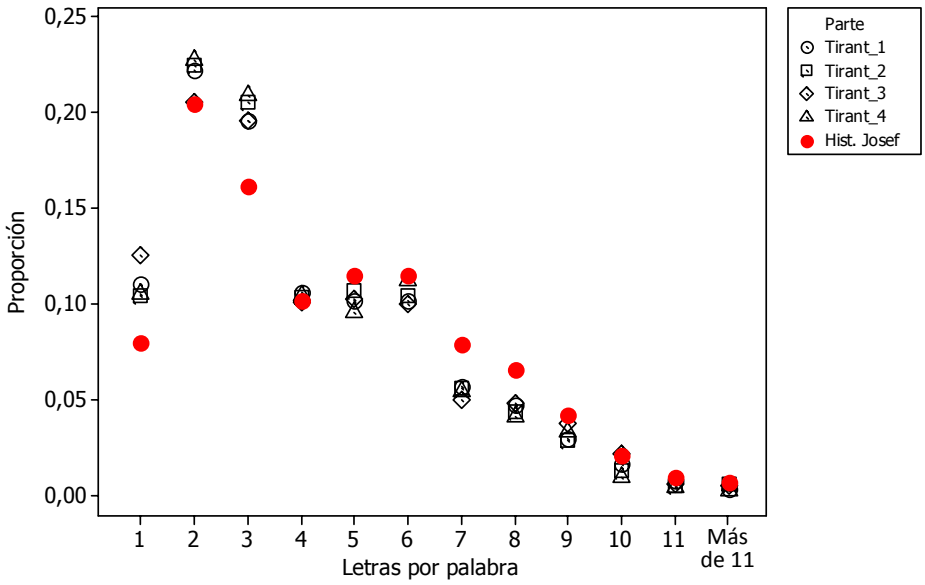
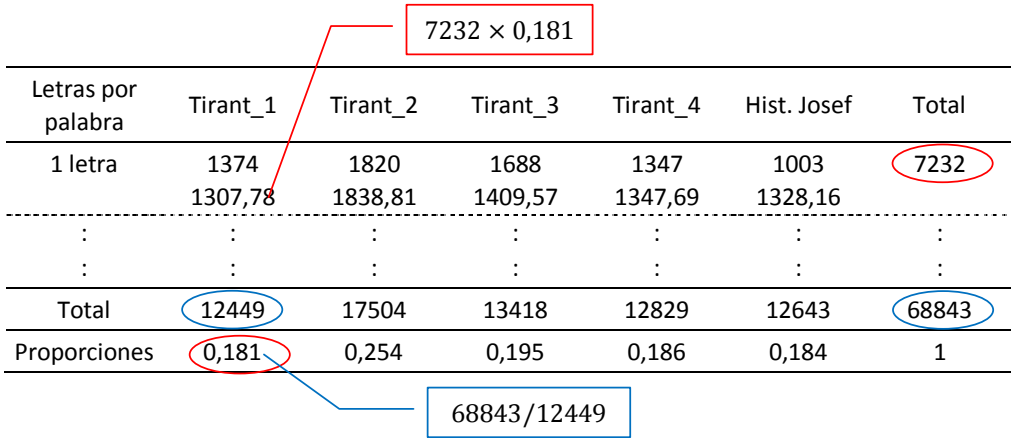


Figura 2: *Proporción de letras por palabra en los cuatro trozos estudiados de 'Tirant lo Blanc' y de 'Història de Josef'*

En la práctica, dos perfiles columna nunca serán idénticos, y lo que se quiere averiguar es si el perfil columna de *Història de Josef* se podría corresponder al perfil de alguna de las partes de *Tirant lo Blanc*, o bien si las diferencias entre ambos son demasiado grandes. La estadística formaliza esta pregunta en torno a lo que se denomina contraste de hipótesis, en el cual hay una hipótesis nula y otra alternativa. En este caso, la hipótesis nula afirma que los valores esperados para los cinco perfiles columna son iguales, y la alternativa postula que son diferentes.

Si las proporciones de palabras según su número de letras son las mismas en todos los trozos estudiados, la frecuencia esperada en cada celda se puede calcular dividiendo el total de su fila por la proporción que representa la suma de su columna con respecto al total de ocurrencias. Por ejemplo, el total de la primera fila es 7.232 y el total de la primera columna es 12.449, que representa una proporción de 0,181 respecto al total de palabras (68.843). Por lo tanto, si la proporción en las celdas se reparte de la misma manera que en los totales, esperaremos encontrar en la primera celda una frecuencia de: 7.232 x 0,181 = 1.307,78.



Letras por palabra	Tirant_1	Tirant_2	Tirant_3	Tirant_4	Hist. Josef	Total
1 letra	1374 1307,78	1820 1838,81	1688 1409,57	1347 1347,69	1003 1328,16	7232
:	:	:	:	:	:	:
:	:	:	:	:	:	:
Total	12449	17504	13418	12829	12643	68843
Proporciones	0,181	0,254	0,195	0,186	0,184	1

Figura 3: *Cálculo de las frecuencias esperadas para construir una tabla de contingencia*

La discrepancia entre las frecuencias observadas y las esperadas se mide a través de la llamada distancia chi cuadrado (χ^2). La aportación de cada celda a esta distancia se calcula a través de la fórmula.

$$\text{Aportación de cada celda a la distancia } \chi^2 = \frac{(\text{Valor observado} - \text{Valor esperado})^2}{\text{Valor esperado}}$$

En la Tabla 4 el primer valor de cada celda es el valor observado, el segundo es el valor esperado suponiendo que no hay diferencias en las proporciones de aparición en los diferentes fragmentos que se comparan. Por último, el tercer valor es la medida de la discrepancia entre el valor observado y el valor esperado; es decir, la contribución de esta celda a la distancia χ^2.

Tabla 4: *Tabla de contingencia de la frecuencia de aparición de palabras según su número de letras en cuatro trozos de 'Tirant lo Blanc' y en 'Història de Josef'*

Letras por palabra	Tirant_1	Tirant_2	Tirant_3	Tirant_4	Hist. Josef	Total
1 letra	1374 1307,78 3,354	1820 1838,81 0,192	1688 1409,57 54,998	1347 1347,69 0	1003 1328,16 79,603	7232
2 letras	2761 2703,62 1,218	3928 3801,44 4,214	2757 2914,06 8,465	2917 2786,14 6,146	2588 2745,75 9,063	14951
3 letras	2432 2417,18 0,089	3588 3398,69 10,545	2628 2605,33 0,195	2683 2490,96 14,781	2036 2454,85 71,510	13367
4 letras	1326 1288,06 1,117	1811 1811,09 0	1358 1388,32 0,662	1342 1327,38 0,161	1286 1308,14 0,375	7123
…			…			…
10 letras	203 198,01 0,126	221 278,41 11,84	292 213,42 28,93	117 204,05 37,14	262 201,1 18,445	1095
11 letras	102 81,01 5,437	86 113,91 6,838	84 87,32 0,126	60 83,49 6,607	116 82,28 13,824	448
Más de 11	46 60,76 3,585	99 85,43 2,155	72 65,49 0,647	36 62,61 11,312	83 61,71 7,348	336
Total	12449	17504	13418	12829	12643	68843

La distancia χ^2 es igual a la suma de las aportaciones de cada celda, es a decir:

$$\text{Distància } \chi^2 = \frac{(1374-1307{,}78)^2}{1307{,}78} + \frac{(1820-1838{,}81)^2}{1838{,}81} + \cdots + \frac{(83-61{,}71)^2}{61{,}71} =$$

$$= 3{,}354 + 0{,}192 + \cdots + 7{,}348 = 729{,}867$$

Cuanto mayor sea esta distancia, más discrepancia habrá entre los valores observados y los esperados y más evidencia tendremos de que la hipótesis nula es incorrecta. La contribución relativa de cada celda a la distancia χ^2 indica las principales celdas responsables de esta discrepancia. Es fácil observar que las contribuciones de las celdas de *Història de Josef* a la distancia χ^2 son mayores que las de *Tirant lo Blanc* y utilizando las técnicas adecuadas se puede constatar una evidencia en contra la hipótesis nula. Así pues, la distribución de palabras según su longitud en los cuatro trozos del *Tirant* es distinta que en la *Història de Josef*.

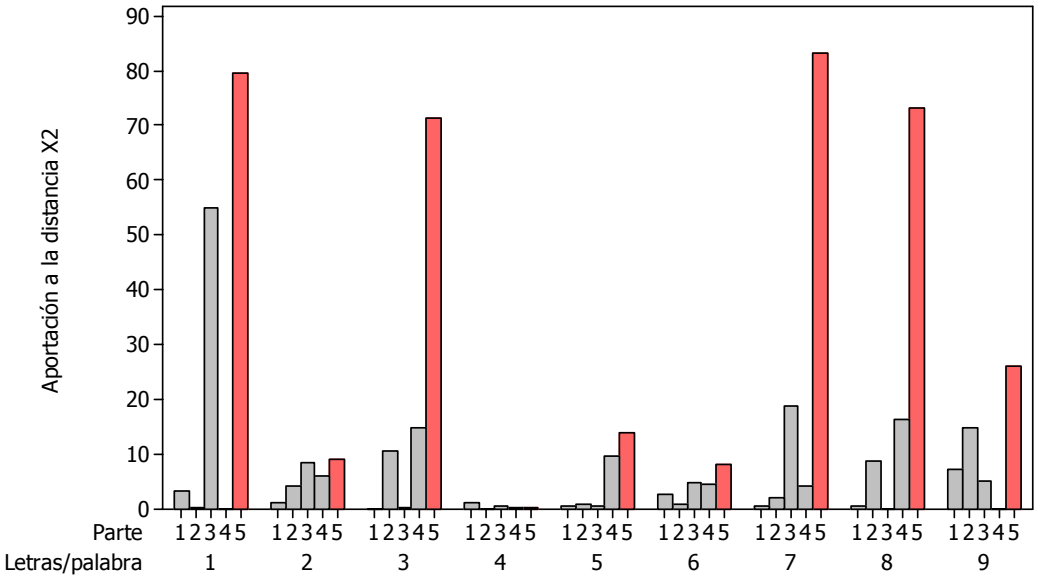


Figura 4: *Aportación de las partes estudiadas a la distancia χ^2, considerando las distancias de las palabras hasta 9 letras. La parte 5 indica la aportación de la 'Història de Josef'*

El problema de la autoría del *Tirant lo Blanc*

De las muchas hipótesis que han circulado sobre la autoría del *Tirant lo Blanc*, la que se ha decantado como más verosímil es la que afirma que el grueso de la obra fue escrito por Joanot Martorell, entre en 1460 y su muerte, mientras que Martí Joan de Galba se encargó de la edición de la obra en 1490, después de reelaborar algunos pasajes y quizá acabar la obra.

Por otra parte, desde muy antiguo se han detectado en el *Tirant lo Blanc* plagios de trozos de obras de Joan Roís de Corella. Basándose en las obras de Corella, y en toda una serie de detalles biográficos de Joanot Martorell, de Galba y de Corella, el profesor Josep Guia plantea la hipótesis de que el autor de *Tirant lo Blanc* (y de gran parte de la prosa catalana de la segunda mitad del siglo XV) es Corella. Según esta teoría, Corella se habría escondido detrás de dos amigos muertos, de los cuales no se conoce ninguna otra obra literaria, como medida de prudencia ante la Inquisición. Hilando más fino, el profesor Josep Guía defiende que un Martorell en las postrimerías de su vida podría haber proporcionado al joven Corella el manuscrito *Guillem de Varoic*, cartas de batalla y otros materiales sobre el mundo de caballerías del cual había formado parte toda su vida, y que Corella los aprovechó para escribir hasta el capítulo 154 del *Tirant lo Blanc*. Según Guia, hasta aquí llegaría el *Tirant lo Blanc* de 1468, cuando muere Martorell y Corella continúa la redacción en solitario.

De Martorell sólo nos han llegado cartas de batalla que no son comparables con un libro de caballerías completo. Por lo tanto, nos encontramos con que en la cuestión de la autoría de *Tirant lo Blanc* se aúnan ingredientes que hacen que el problema esté a caballo entre la segunda y la tercera familia de problemas descritos anteriormente. Primero, se nos plantea el problema de comparar el estilo del *Tirant lo Blanc* con el estilo de obras "comparables" de Corella. En segundo lugar, sería muy útil documentar la existencia de alguna frontera estilística en el *Tirant lo Blanc,* aunque debemos tener en cuenta que debido a la longitud de la obra, una frontera interior podría indicar tanto la existencia de dos autores como de dos etapas de escritura diferenciadas.

Joan Roís de Corella y el *Tirant lo Blanc*

Como no se dispone de textos de Corella más "comparables" a un libro de caballerías, ni ediciones de *Història de Josef* más modernas ni hechas por Martí de Riquer, habrá que ser muy prudentes al interpretar las diferencias de estilo que se puedan encontrar, ya que podrían ser debidas a la intervención de los editores o a que pertenecen a géneros literarios muy diferentes. Al estudiar la homogeneidad de estilo de *Tirant Io Blanc* estos peligros de confundir el efecto autor con los efectos género y/o editor desaparecerán.

De todos los aspectos cuantificables comentados, se han escogido como variables explicativas la distribución de palabras según su número de letras, la frecuencia de uso de 44 palabras herramienta y la distribución de las frases según su número de palabras.

Distribución de las palabras según sus longitudes

Ya hemos visto en la Tabla 3 cómo las celdas que corresponden a la *Història de Josef* tienen una aportación mayor a la distancia χ^2. También la Figura 2 indica que la proporción de palabras largas es mayor en la *Història de Josef* que en las partes analizadas del *Tirant Io Blanc*.

Frecuencia de uso de palabras herramienta

Se ha analizado la frecuencia de uso de las 44 palabras herramienta seleccionadas para las partes de *Tirant lo Blanc* y para el texto de Corella. Entre la lista de palabras hay conjunciones, preposiciones, artículos y algunos elementos de locuciones prepositivas o conjuntivas. Las celdas que tienen una contribución más alta a la distancia χ^2 vuelven a ser las de *Història de Josef* y alguna celda de la columna Tirant_3. Lo que parece

distinguir *Història de Josef* del resto es que aparecen menos "e", "como " y "mucho", y más "cual" y "a los" que en *Tirant lo Blanc*. En el último trozo del *Tirant* aparecen más "e" y "mucho" que en las otras partes analizadas del mismo libro.

Distribución de las frases según su número de palabras

La Tabla 5, muestra la distribución de las frases según su número de palabras, categorizadas en ocho grupos. Observamos cómo las contribuciones de las celdas de la columna de la *Història de Josef* a la distancia χ^2 también vuelven a ser las mayores. Esta coincidencia de los resultados obtenidos a partir de las tres variables es bastante sintomática.

Tabla 5: *Distribución de las frases según su número de palabras. El primer valor de cada celda es el número de frases y el segundo es la contribución de esa celda a la distancia χ^2.*

Palabras por frase	Tirant_1	Tirant_2	Tirant_3	Tirant_4	Hist. Josef	TOTAL
de 1 a 12	46 (0,451)	74 (0,444)	28 (6,063)	86 (15,064)	11 (15,274)	245
de 13 a 20	58 (1,524)	125 (1,486)	77 (0,256)	115 (5,258)	26 (15,39)	401
de 21 a 25	41 (0,012)	71 (0,341)	41 (0,089)	57 (0,08)	27 (0,956)	237
de 26 a 30	30 (0,006)	49 (0,008)	28 (0,36)	42 (0,092)	25 (0,062)	174
de 31 a 40	47 (0,316)	93 (0,996)	62 (1,068)	54 (3,457)	44 (0,184)	300
de 41 a 50	36 (0,214)	47 (1,107)	39 (0,339)	38 (1,208)	36 (3,043)	196
de 51 a 70	43 (0,032)	58 (1,685)	43 (0,057)	44 (2,961)	58 (17,29)	246
más de 70	34 (1,179)	32 (4,47)	38 (2,08)	19 (9,844)	43 (17,843)	166
TOTAL	335	549	356	455	270	1965

Homogeneidad de estilo en el *Tirant lo Blanc*

Algunos estudiosos plantean la existencia de diferencias de estilo en el *Tirant lo Blanc*. Estas diferencias podrían ser debidas a la evolución en el estilo del autor o a que lo escribieron autores diferentes.

Para explorar la existencia de fronteras estilísticas, se ha analizado cómo se agrupan las distribuciones de palabras según su número de letras y la frecuencia de uso de las 44 palabras herramienta en los 10 bloques (1a, 1b, 2a, 2b, 3a, 3b, 4a, 4b, 5a y 5b)

descritos en la Tabla 1. Calculando las distancias χ^2 que resultan de comparar el perfil de la suma de las m primeras columnas y el perfil de la suma de las $(10 - m)$ últimas, tanto para letras por palabra como para la frecuencia de uso de palabras herramienta, las distancias mayores aparecen cuando comparamos los 7 primeros bloques con los 3 últimos. Por tanto, si hay una frontera de estilo, ésta podría estar entre los bloques 4a y 4b, que corresponde a una frontera entre los capítulos 325 y 350, tal como también plantean algunos expertos.

Para confirmar la evidencia a favor de esta frontera se han agrupado los 10 bloques del *Tirant* en dos grupos de 7 y de 3 bloques de las 120 maneras posibles, y se ha calculado la distancia χ^2 entre cada una de las 120 parejas de perfiles columna, obtenidas sumando las columnas de los grupos de 7 y las de los grupos de 3. El objetivo es hacer lo que se llama un test de permutaciones, contando cuántas combinaciones en grupos de 7 y de 3 columnas tienen perfiles más diferentes que entre los 7 primeros bloques y los 3 últimos. La Figura 5 muestra la distribución de estas 120 distancias χ^2.

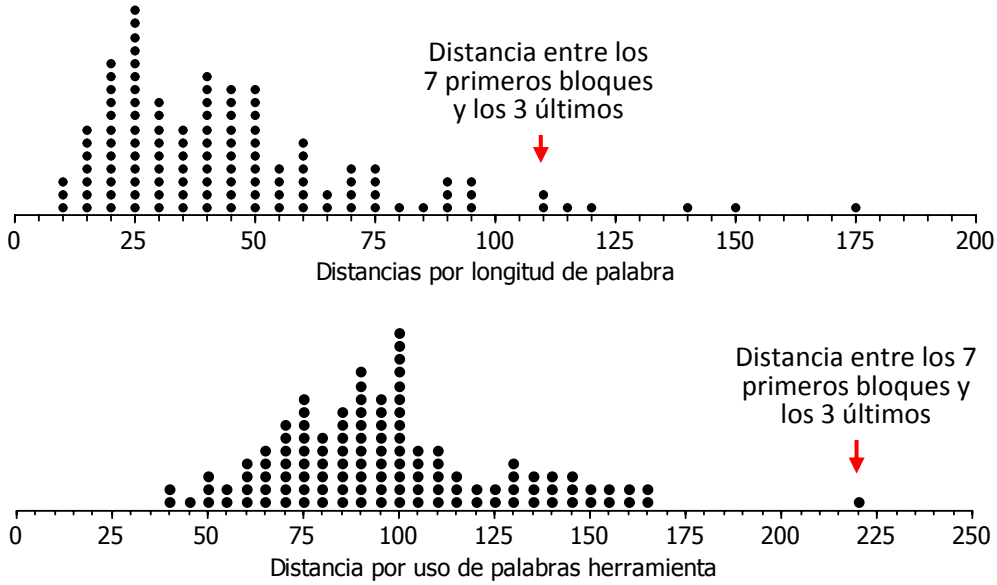


Figura 5: *Tests de permutaciones. Diagrama de puntos de las distancias χ^2 entre parejas de perfiles columna, obtenidas agrupando los 10 bloques del 'Tirant lo Blanc' en grupos de 7 y 3 bloques de las 120 formas posibles, por longitud de palabra y por uso de palabras herramienta.*

Para la frecuencia de uso de palabras herramienta, la combinación que coge los 7 primeros bloques, por un lado, y los tres últimos, por el otro, da la mayor distancia χ^2 de todas las combinaciones posibles. Para la distribución de longitudes de palabra, hay cinco combinaciones de las 120 posibles con distancia χ^2 mayor que cuando se comparan los siete primeros bloques con los tres últimos, pero estas cinco

combinaciones contienen los bloques 4b y 5b en el grupo de tres columnas. Todo esto indica que la evidencia a favor de una frontera de estilo en algún lugar entre los capítulos 325 y 350 es muy grande.

Conclusiones

Los análisis realizados no avalan la autoría de Joan Roís de Corella, ya que las diferencias de estilo entre el *Tirant lo Blanc* y la obra disponible de Corella son claras, pero estas diferencias podrían ser debidas a que los textos comparados corresponden a diferentes generaciones, o también a aspectos ligados a la edición.

Sí que se ha encontrado una clara frontera en el estilo en torno a los capítulos 325-350. A la estadística se le puede pedir más precisión al definir en qué punto puede estar esa frontera, y también en la cuantificación de la magnitud del cambio, pero son los expertos del ramo los que tendrán que decir las razones más verosímiles sobre las causas de este cambio de estilo.

(Proyecto de la Diplomatura de Estadística, presentado en febrero de 1997 con el título "Anàlisi estadística de l'estil literari: aproximació a l'autoria del Tirant lo Blanc*").*

Hemos escogido este proyecto porque fue el primero de una línea de investigación que ha dado excelentes resultados, y también porque muestra un campo donde, seguramente para sorpresa de muchos, la estadística es muy útil. Un resumen completo, incluyendo otras técnicas que se utilizaron, se encuentra en: Ginebra, J. i Cabos S.: "Anàlisi Estadística de l'estil literari: Aproximació a l'autoria del Tirant lo Blanc*". Afers, Nº. 29, 1998.*

Desde entonces se ha trabajado mucho sobre este tema. Alexandre Riba realizó su tesis doctoral "Homogeneïtat d'estil al Tirant lo Blanc*", dirigida por Josep Ginebra, que puede ser consultada en: http://www.tesisenxarxa.net/TDX-0411105-115931/index.html. También se han publicado artículos sobre este tema en revistes estadísticas de primera línea.*

Susanna Cabos cursó la Diplomatura de Estadística (tercer premio nacional final de carrera) y orientó su actividad profesional al campo de la calidad, dentro del sector industrial. Profundizó en estos temas cursando el Máster en Gestión de la Calidad (Fundación UPC), y también el *European Master's Program in Total Quality Management*, organizado por la *European Foundation for Quality Management* (EFQM). Su actividad profesional ha transcurrido siempre en el sector de la automoción. Actualmente es Responsable de Calidad y Medio Ambiente para España y Francia de una importante empresa del sector eléctrico.

2

Análisis estadístico de datos musicales:
Estudio interpretativo de la obra "TRÄUMEREI" (R. Schumann)

Proyecto realizado por**: Josué Almansa Ortiz**
Dirigido por**: Pedro Delicado Useros**

Como ocurre con otras disciplinas artísticas, a primera vista se diría que la percepción de una obra musical (tanto de su composición como de su interpretación) es algo subjetivo, que depende del espectador, y que es difícilmente cuantificable. Pero hay formas de realizar análisis objetivos de diferentes aspectos de una obra musical basándose en métodos de las matemáticas, la estadística y la física.

Este estudio objetivo de la música puede desarrollarse sobre tres grandes líneas de investigación: la composición, la interpretación y el sonido. En este trabajo nos centramos en el análisis de la interpretación musical desde el punto de vista estadístico.

Concretamente se consideran datos del "tempo" (velocidad de interpretación) en 28 interpretaciones de la obra "Träumerei" de Robert Schumann y se analizan mediante una técnica denominada Análisis de Componentes Principales Funcionales. Este estudio, como todo estudio interpretativo de una partitura, tiene por objetivo fundamental el describir de qué forma diferentes músicos interpretan una misma pieza musical, encontrar sus rasgos comunes (lo que en musicología se conoce como comunalidades) y también sus peculiaridades (diversidad).

El autor: Robert Schumann (1810-1856)

Robert Schumann sus estudios musicales empezó a los ocho años, aunque no se dedicó por completo a la música hasta los 20 años, cuando abandonó la carrera de Derecho. Una lesión le llevó a abandonar el estudio del instrumento (el piano) y a centrarse en la composición musical. En el año 1840 se casó con Clara Wieck, hija de su maestro, que se oponía al enlace por la inestable posición económica de R. Schumann. En 1843 fue profesor del Conservatorio de Leipzig, que había sido fundado por Mendelssohn. En 1844 se estableció en Dresde y en 1850 fue nombrado director de música en Dusseldorf. Una enfermedad mental fue afectando a su salud hasta su muerte en 1856. Su mujer Clara Wieck fue una de las mejores pianistas del s. XIX, y una gran difusora de la obra de R. Schumann.

R. Schumann es un claro exponente del romanticismo, donde no destaca tanto por su virtuosismo, como por su expresividad musical, su subjetividad y los matices psicológicos de un estado de ánimo. La obra que se analiza en este proyecto, Träumerei, es un claro ejemplo de ello.

La obra: *Träumerei*

Träumerei, que traducido al castellano significa “ensueño”, es la séptima de trece piezas cortas que constituyen el álbum *Kinderszenen* (escenas de la infancia). Fue compuesta por R. Schumann en 1838, cuando estaba secretamente prometido con Clara Wieck. Estas piezas fueron seleccionadas de entre unas 30 que fueron compuestas para Clara.

Träumerei está formada por tres frases de ocho compases cada una, con estructura

A A B A'

Según indica la partitura, la primera de las frases (A) se debe repetir y la tercera frase (A') no es más que la frase A del principio, pero con variaciones al final que le dan el carácter conclusivo. Cada frase se subdivide en dos periodos de cuatro compases cada uno. Todos los periodos tienen una estructura rítmica casi idéntica (sólo en el último compás de los periodos hay algunas diferencias), y básicamente sólo se diferencian en las notas y la armonía. Como se puede observar, es una obra de forma muy regular.

La música se crea mediante relaciones de tensión y distensión. En esta obra el momento de máxima tensión se crea en la frase B (más concretamente en la nota *si blanca* del compás 14) y posteriormente se relaja en A' para concluir la obra.

La partitura de la obra se muestra en la Figura 1 (se ha marcado sobre ella la situación de las frases).

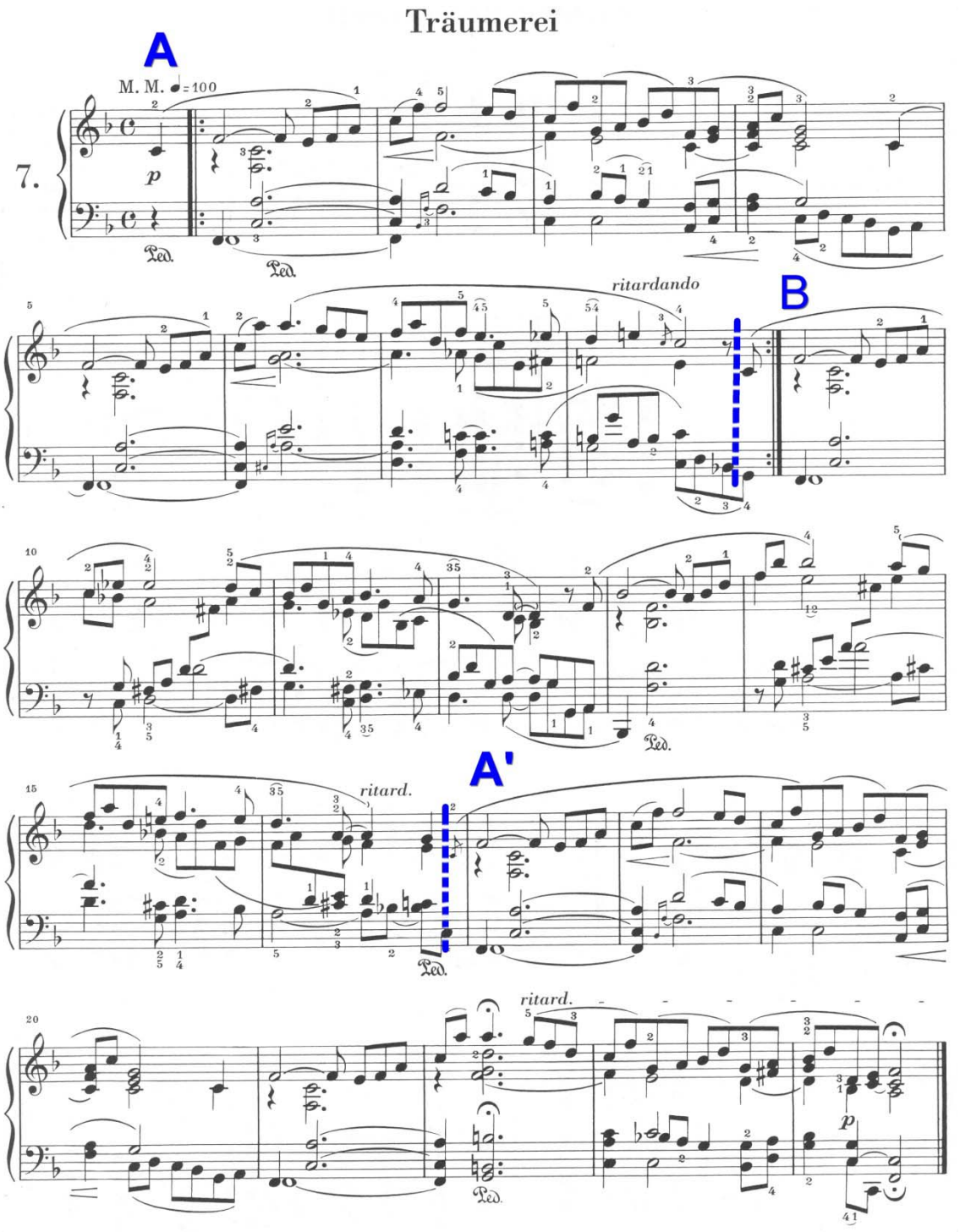


Figura 1: *"Träumerei" de Schumann, op. 15/7. Reproducido con permiso de G. Henle Publishers, http://www.henle.de.*

En cuanto a la interpretación rítmica de la obra, cabe destacar lo siguiente:

1. No hay ninguna indicación de que la repetición de la frase A se deba tocar de manera diferente a la primera vez, con lo cual a nivel rítmico estas dos frases (la A y su repetición) deben interpretarse de modo similar.

2. Al final de cada frase está indicado hacer un *ritardando* (una moderación en el tempo de las interpretaciones), y se puede ver que el *ritardando* del final de la obra es mucho más largo. También hay que tener en cuenta el contexto de cada *ritardando*, ya que el del final de frase A (la primera vez que aparece) no es el mismo que el del final de frase B (justo después del momento de máxima tensión).

3. En el compás 22 aparece un *calderón* sobre una blanca que indica que aquella nota debe alargarse a discreción.

4. Aparte de las indicaciones explícitas sobre cómo se debe interpretar una obra (aquéllas que están escritas en la partitura), siempre hay otras "indicaciones implícitas" que se deben entender en función del estilo y de la época en que fue compuesta. En esta obra romántica, los pianistas acostumbran a jugar bastante con la interpretación rítmica de la obra. Se suelen usar "herramientas interpretativas" como el *rubato*. Un *rubato* consiste en una ligera variación del ritmo de unas notas mediante breves aceleramientos o retrasos con el objetivo de aumentar la expresión. Un punto donde está ampliamente aceptado que haya un *rubato* es en las corcheas del primer compás de cada periodo, hasta llegar a la blanca del segundo compás (Figura 2).

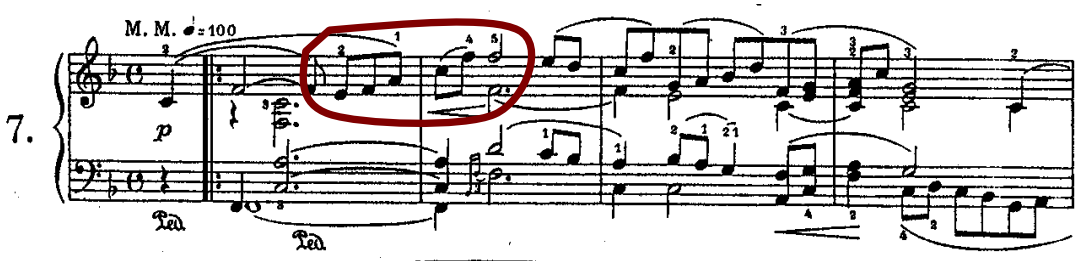


Figura 2: *Ejemplo de caso en que está aceptada la presencia de un "rubato"*

A partir de estos principios generales cada intérprete pone características propias de su estilo.

La estructura formal de la obra en sí es muy regular. En el romanticismo lo importante no es la forma (la estructura de la partitura), sino que el interés de la obra radica más en la expresión, en reflejar sentimientos. Por eso, es de esperar que esta obra no sea

interpretada por los músicos como un simple ejercicio de técnica instrumental, donde todas las notas son tocadas exactamente tal como la partitura indica, siendo todas las corcheas de igual duración, las negras de duración doble que las corcheas, etc. El carácter interpretativo tan expresivo de esta obra se conseguirá, por una parte, a través de constantes matizaciones y cambios en la intensidad del sonido, así como en continuas variaciones de la velocidad de las notas con respeto a lo que está "literalmente" escrito en la partitura, mediante ligeras aceleraciones y desaceleraciones de las notas (pero, obviamente, sin llegar a ser tan exageradas que parezca que lo qué se está tocando no tiene nada a ver con la partitura).

Así pues, lo que a priori se espera encontrar como más característico son los *ritardandos* y *rubatos*, y el carácter de cada frase (la frase B muy diferente a A y A', y la frase A' difiriendo de A en su final).

Los datos

Los datos con que ha sido realizado este trabajo han sido facilitados por Bruno H. Repp, psicólogo de formación e investigador en Haskins Laboratories (EEUU) en temas de psicoacústica y psicología musical. Estos datos provienen de 28 interpretaciones diferentes de la obra *Träumerei*, llevadas a cabos por 24 destacados pianistas, y se obtuvieron a partir de grabaciones sobre diferentes soportes (cassettes, CDs, etc.). Algunas son grabaciones en directo y otras fueron grabadas en un estudio. Algunas de ellas fueron ejecutadas aisladamente, y otras dentro de la colección "Kinderszenen" (Escenas de la infancia).

De entre los 24 pianistas destaca Fanny Davies (DAV) que fue alumna de Clara Schumann (Clara Wieck) con una grabación del año 1929 sobre un soporte que en el momento de la recogida de datos se encontraba bastante deteriorado (en palabras del propio Bruno H. Repp: "from a very scratchy original").

Dos de estos artistas (Cortot i Horowitz) están representados por tres grabaciones diferentes de cada uno de ellos. En la Tabla 1 se muestra la relación de las 28 grabaciones con los nombres de los pianistas y un código de tres letras que los identifica de ahora en adelante.

Tabla 1: *Relación de interpretaciones incluidas en los datos a analizar*

Código	Intérprete	Año de grabación
ARG	Martha Argerich	< 1983
ARR	Claudio Arrau	1974
ASH	Vladimir Ashkenazy	1987
BRE	Alfred Brendel	< 1980
BUN	Stanislav Bunin	1988
CAP	Sylvia Capova	< 1987
CO1	Alfred Cortot	1935
CO2	Alfred Cortot	1947
CO3	Alfred Cortot	1953
CUR	Clifford Curzon	1955
DAV	Fanny Davies	1929
DEM	Jörg Demus	1960
ESC	Christoph Eschenbach	< 1966
GIA	Reine Gianoli	1974
HO1	Vladimir Horowitz	1947
HO2	Vladimir Horowitz	<1963
HO3	Vladimir Horowitz	1965
KAT	Cyprien Katsaris	1980
KLI	Walter Klien	?
KRU	André Krust	1960
KUB	Antonin Kubalek	1968
MOI	Benno Moiseiwitsch	1950
NEY	Elly Ney	1935
NOV	Guiomar Novaes	< 1954
ORT	Cristina Ortiz	< 1988
SCH	Artur Schnabel	1947
SHE	Howard Shelley	< 1990
ZAK	Yakov Zak	1960

Los datos recogen el tiempo, en milisegundos, que dura cada nota de la melodía. Las notas de duración mayor que la corchea se dividieron en corcheas de igual duración cada una de ellas, de esta forma tenemos cada interpretación discretizada uniformemente a la corchea. La última nota no está medida, ya que el proceso de medición se calculaba mediante la diferencia en el tiempo de inicio de dos notas

consecutivas y por esta razón sólo hay medidas hasta la penúltima nota. Así pues, para cada interpretación tendremos medidas de la duración de 253 corcheas. La Figura 3 muestra los valores de estas duraciones para cada una de las 28 interpretaciones y en la Figura 4 tenemos la duración acumulada.

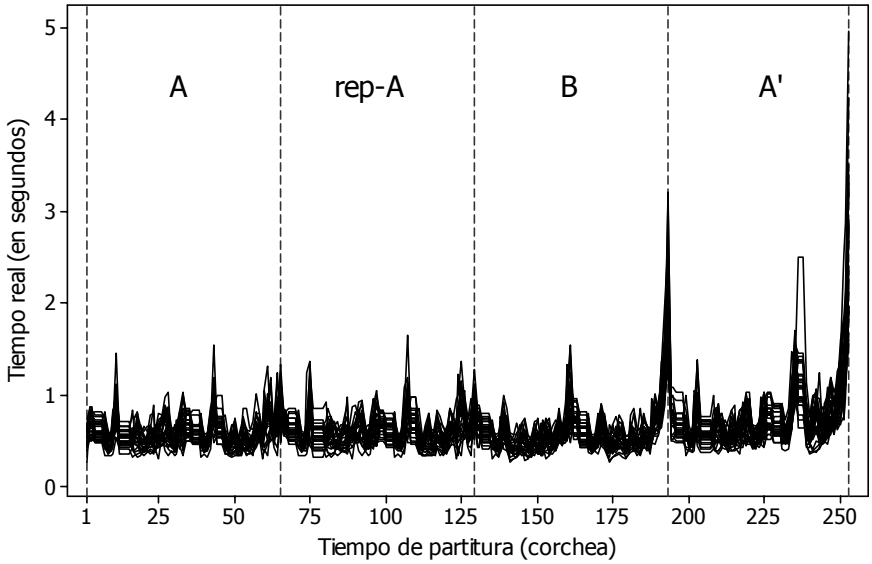


Figura 3: *Duración de cada nota en las 28 interpretaciones*

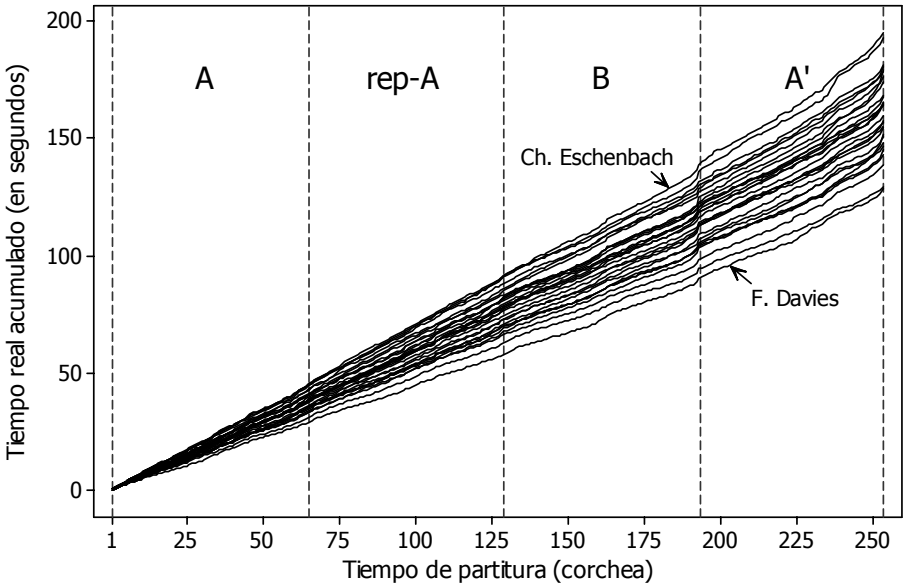


Figura 4: *Duración acumulada para cada nota en cada una de las 28 interpretaciones. Se indica también la línea correspondiente al pianista más rápido (F. Davies) i al más lento (Ch. Eschenbach)*

El tiempo acumulado en la corchea 253 da el tiempo total de la interpretación. La Figura 5 muestra un diagrama de puntos de estos tiempos totales con la identificación de los pianistas que tienen los valores extremos, y también los valores centrales.

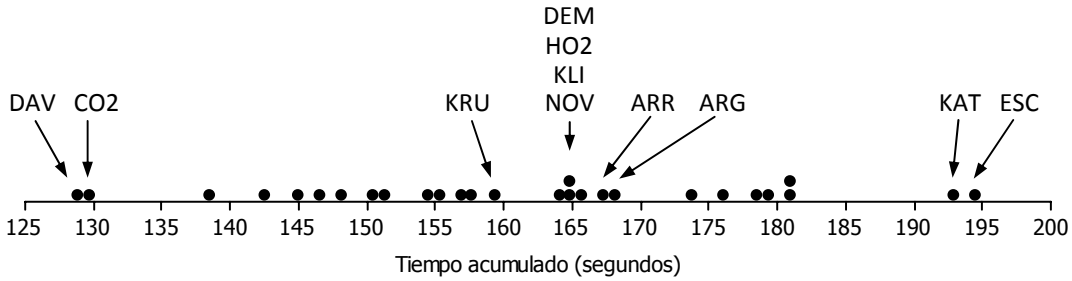


Figura 5: *Tiempo total de ejecución, con indicación de los pianistas que tienen los valores extremos y los centrales.*

Hay que decir que el proceso de obtención de estos datos por parte de Bruno H. Repp fue bastante artesanal y requirió mucho tiempo de trabajo manual, con lo que se asume que los datos se han medido con algo de error. Un programa de digitalización del sonido mostraba las ondas en la pantalla de un ordenador con una resolución de 2 milisegundos. Se situaba un cursor al inicio de cada nota y se etiquetaban esos puntos. La diferencia entre tiempos de dos etiquetas sucesivas forman los datos en que se basa este proyecto.

Expresión de los datos en forma de funciones

Los datos originales, que forman una matriz de dimensión 253x28, no tienen forma funcional sino que simplemente indican la duración de cada nota en la interpretación de cada pianista. La forma en que estos datos se transforman en datos funcionales es la siguiente: para cada interpretación se calcula una función que mide el tiempo transcurrido t (en milisegundos) desde el principio hasta cualquier posición dada s de la partitura (medida en notas, o en corcheas para ser más exactos). La estimación de esta función se realiza acumulando las duraciones de las notas precedentes (las representadas en la Figura 4).

Para cualquier número real s en [1, 253] es posible estimar (por interpolación lineal, por ejemplo) el tiempo acumulado t correspondiente. Podemos definir así, una función $t(s)$ para s en [1, 253]. Estas funciones son muy similares en todas las interpretaciones y difícilmente encontraremos en ellas peculiaridades de cada pianista.

Las inversas de las funciones $t(s)$ son funciones de posición $s(t)$. Dado un tiempo t, $s(t)$ calcula la posición en la partitura, s, de la nota que está interpretando el pianista en el instante t. La aplicación de las nociones básicas de la física permiten calcular la velocidad $v(t)$ y la aceleración $a(t)$ como primeras y segunda derivadas de la función de posición: $v(t) = s'(t)$, $a(t) = v'(t) = s''(t)$. Estas nuevas funciones (velocidad y aceleración) son más informativas que la posición, y pueden ayudar a diferenciar mejor unas interpretaciones de otras.

La opción que se consideró más adecuada fue la de trabajar con las funciones de tiempo acumulado $t(s)$ como datos funcionales de partida. La principal ventaja de trabajar con $t(s)$ es que de esta manera todas las funciones observadas $t_i(s)$ tienen un soporte común ($[1, 253]$). Éste no es el caso cuando se utilizan las funciones de posición $s(t)$ debido a que diferentes interpretaciones tienen duraciones diferentes. Tener un soporte común es muy conveniente para comparar las diferentes funciones y para relacionar sus valores directamente con posiciones en la partitura.

Se denomina *tiempo-transcurrido* a la función $t(s)$; *lentitud* a su primera derivada $w(s) = t'(s)$ y *desaceleración* a su segunda derivada $d(s) = w'(s) = t''(s)$.

Aspectos comunes a todas las interpretaciones

La Figura 6 muestra la función promedio de las 28 funciones de *lentitud* estimadas. Esta función resume las características comunes a las 28 interpretaciones, lo que en el estudio de la interpretación musical se conoce como *comunalidades*. En este caso es posible observar que al final de cada frase hay un *ritardando*, una ralentización en el tempo de las interpretaciones (máximos locales de la función lentitud), siendo esto más evidente al final de la frase B y al final de la pieza.

A una distancia de dos desviaciones estándar (calculadas para cada posición de la partitura) se han colocado bandas que dan idea de la variabilidad de la función lentitud a lo largo de la partitura. Al final de frase B, el *calderón* y el *ritardando* final son las fases de la obra donde la función de lentitud tiene más variabilidad.

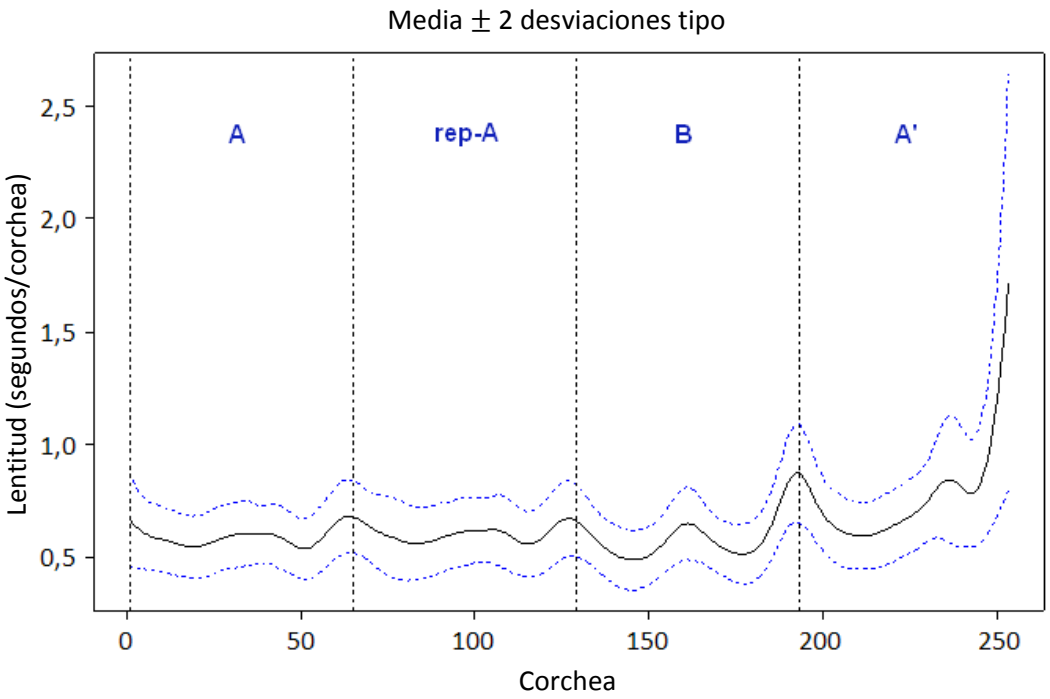


Figura 6: *Media de las 28 funciones lentitud, con bandas a dos desviaciones estándar*

Diversidad en las interpretaciones

Se han representado cada una de las 28 interpretaciones mediante una función $t_i(s)$, con $i = 1, 2, \ldots, 28$. La función promedio, $\bar{t}(s) = \frac{1}{n} t_i(s)$, recoge lo que tienen en común las interpretaciones estudiadas. Por lo tanto, las funciones centradas en la media, $t_i(s) - \bar{t}(s)$, representan las particularidades de cada interpretación: lo que se conoce como *diversidad*.

El *Análisis de Componentes Principales Funcionales* (ACPF) es una técnica estadística que analiza estas diversidades, las funciones $t_i(s) - \bar{t}(s)$, y las resume mostrando lo que hay de común en la forma en que los individuos son diferentes.

El ACPF proporciona un conjunto de funciones (llamadas *funciones principales*) que explican la variabilidad de los datos. En nuestro caso, realizado este análisis a las funciones lentitud $t_i(s)$ se observa que las dos primeras funciones principales destacan sobre las otras y llegan a explicar el 80% de la variabilidad. Estas dos primeras funciones son la velocidad global en la interpretación y la rapidez de la interpretación (*ritardando*). La primera explica el 60% de la variabilidad de las interpretaciones en torno a la interpretación media y la segunda explica el 20%. Las 5 primeras llegan a explicar el 92% del total de la variabilidad.

La representación gráfica de los coeficientes permite ordenar a los individuos analizados según los valores que en ellos toman las diferentes funciones principales. La Figura 7 muestra la nube de puntos formada por los coeficientes correspondientes a las dos primeras funciones principales. Las interpretaciones con valores en la primera función principal más negativos (más lentas) son ESC, KAT y BUN. Las interpretaciones con primera componente principal más positiva (las más rápidas) son CO2 y DAV. Evidentemente, estos resultados coinciden con que ya habíamos visto en la Figura 5.

Fijándonos también en la segunda componente (eje vertical) se observa que BUN, además de ser lento, enfatiza el *ritardando* final (más lento) con respeto al resto de la obra. Por su parte, ORT tiene una velocidad global próxima a la media y enfatiza el *ritardando* final con respeto al resto de la obra. ARG tiene una velocidad global próxima a la media y no pone un énfasis especial en el *ritardando* final (es una interpretación con velocidad homogénea).

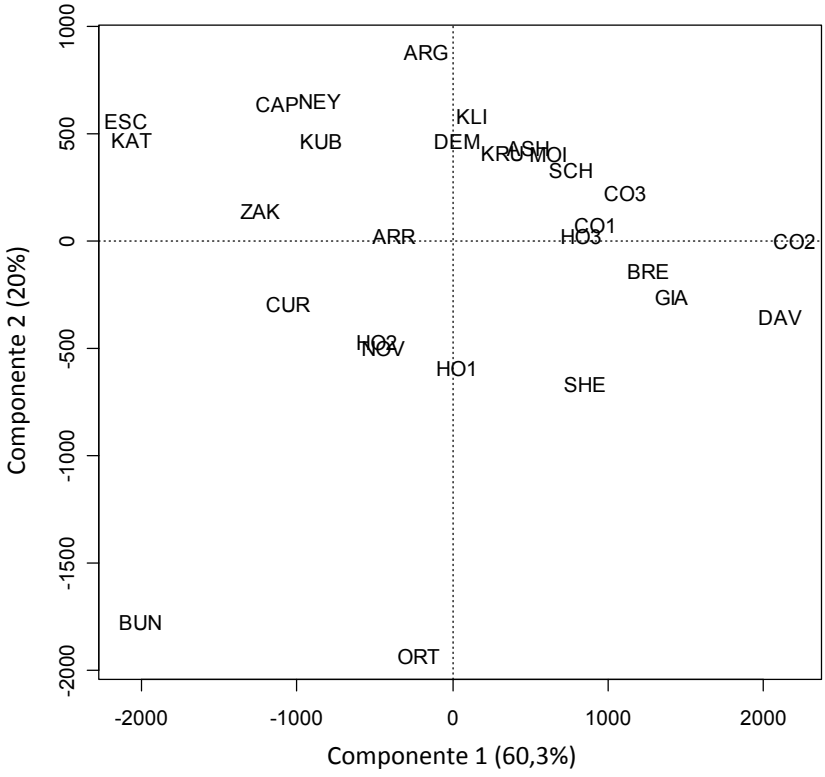


Figura 7: *Coeficientes de las funciones lentitud en las 2 primeras funciones principales*

Conclusiones

Este proyecto se basa en unos datos recogidos laboriosamente por Bruno H. Repp y muestra que estos datos, tratados como datos funcionales, ayudan a revelar unas estructuras que de otra manera pasarían desapercibidas.

En este trabajo se muestra como el Análisis de Componentes Principales Funcionales (ACPF) permite clasificar a los intérpretes en función de dos factores que explican el 80% de la variabilidad de los datos. Estos factores son la velocidad global en la interpretación –explica el 60% de la variabilidad total– y la rapidez de la interpretación al final de la obra (*ritardando*), que explica el 20%. Estos análisis también han permitido describir diferentes patrones de interpretación en la estructura rítmica de cada frase de la pieza musical, y diversas maneras de interpretar el *rubato* (Figura 2), que son muy difíciles de detectar simplemente con el oído.

(Proyecto de la Licenciatura en Ciencias y Técnicas Estadísticas presentado en Junio de 2005 con el título "Análisis de datos musicales mediante componentes principales funcionales: Estudio interpretativo de la obra Träumerei")

El objetivo fundamental de este proyecto es mostrar las posibilidades de una nueva forma de analizar los datos, poniendo de manifiesto sus ventajas frente otras técnicas más habituales. No es un tema fácil de explicar pero lo hemos escogido porque muestra la aplicación de la estadística en un entorno que, al igual que la literatura, puede parecer muy lejano.

Josué Almansa cursó la Diplomatura de Estadística y también la Licenciatura en la UPC. Mientras estaba terminando la licenciatura se incorporó como becario al Instituto Municipal de Investigación Médica (IMIM-Hospital del Mar) donde participó en diferentes proyectos de investigación. Posteriormente, obtuvo una beca de formación de investigadores e inició los estudios de doctorado en la Universidad de Barcelona. Actualmente está realizando una estancia en la Universidad de Tilburg con su codirector de tesis, realizando diversos estudios. También obtuvo el título de profesor de flauta travesera en el Conservatorio del Liceo y ha participado en el conjunto instrumental de la Universidad Pompeu Fabra, pero nos cuenta que ha ido dejando de lado la música y que piensa ganarse la vida con la estadística.

3

Aplicación de técnicas estadísticas al estudio del arte pictórico de los siglos XV al XIX

Proyecto realizado por: **Miquel Romero Obón**
Dirigido por: **Lídia Montero Mercadé**

Cualquier forma artística presenta claramente un vínculo con la parte emocional humana y no es de extrañar que las valoraciones que puedan hacerse sobre una obra se basen en factores subjetivos. Sin embargo ¿qué hay en una pintura que nos hace pensar en un autor o en un movimiento artístico concreto? Hay mecanismos de clasificación basados en el entrenamiento visual en los cuales se busca, conscientemente o no, ciertos patrones propios de la época o del pintor.

Por otra parte, las técnicas actuales de carácter científico que dan apoyo a las aserciones de los estudiosos del arte resultan caras y agresivas debido al carácter destructivo de la toma de muestras y su posterior análisis químico. Por tanto, la búsqueda de mecanismos de identificación y clasificación aplicando técnicas estadísticas resulta bastante interesante, ya que en este caso las técnicas destructivas pueden aplicarse solo con carácter confirmativo.

El alcance de este estudio se centra en las pinturas al óleo de entre los siglos XV y XIX, que comprende las épocas desde el renacimiento hasta el fauvismo.

Evolución de la técnica y de los materiales usados para pintar

La primera parte de este proyecto ha consistido en el estudio de los materiales y de las técnicas utilizadas por los pintores a lo largo del periodo estudiado, que va desde el siglo XV hasta los inicios del siglo XX en el territorio europeo.

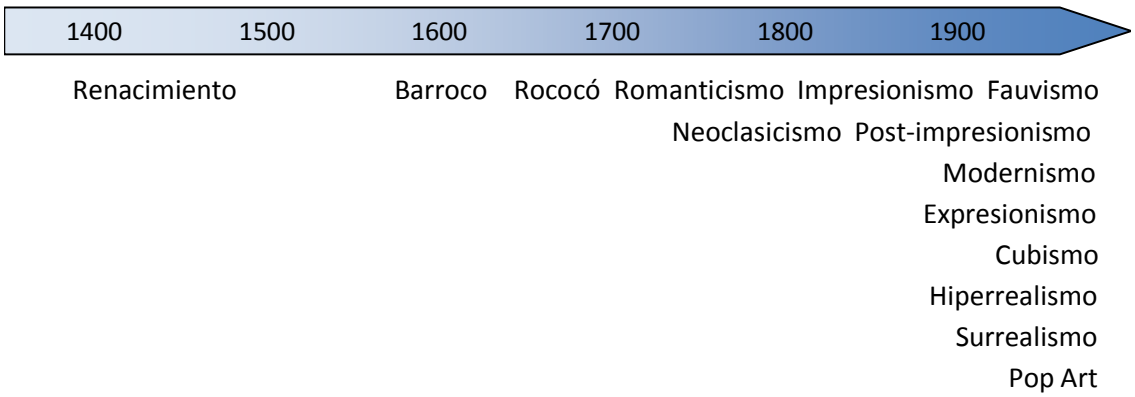


Figura 1: *Movimientos artísticos a lo largo del tiempo*

De este estudio se desprende que los hechos diferenciales más importantes durante el periodo estudiado, y que pueden ser útiles para identificar el estilo y la época en que ha sido creada una obra son los siguientes:

- *Uso de pigmentos*

  A lo largo del tiempo se han utilizado diferentes pigmentos para salvar las incompatibilidades químicas. Por ejemplo, la incompatibilidad entre el blanco de plomo y el rojo de sulfuro de mercurio impidió la existencia de cierta gama de rosas hasta que se descubrió el blanco de Cinc.

- *Forma de interpretar la luz*

  Mientras que los movimientos más antiguos representan la luz sobre un cuerpo añadiendo blanco al pigmento local y la oscuridad añadiendo negro, los movimientos más modernos juegan con el color complementario del local, llegando a eliminar el color negro de la paleta de los pintores impresionistas.

- *Gama cromática*

  Los mismos colores y las combinaciones son propias del momento y del autor, de forma que se encuentran preferencias por el uso monocromo, las tríadas de colores básicos, las tríadas de colores secundarios, etc.

- *Perspectiva*

  Partiendo de la práctica inexistencia de la perspectiva, hay una evolución hacia el uso lineal y más adelante hacia las perspectivas atmosférica y cromática.

- *Temas representados*

  Con respecto a la tipología de temas representados, mientras la pintura es considerada un oficio y no un arte, y los pigmentos no están al alcance de la economía de los pintores, la representación habitual es la determinada por el cliente principal, la Iglesia. Más adelante, el pintor se libra del tema impuesto y el tema representado se basa principalmente en la influencia del pensamiento del momento. Posteriormente, la invención del tubo de estaño hace "portátil" la pintura al óleo y se produce un incremento de las representaciones al aire libre.

- *Recursos compositivos*

  El recurso compositivo aplicado para llevar la mirada del observador hacia el centro de interés no es siempre el mismo. La influencia de los maestros y de las escuelas artísticas marcan una evolución en el tiempo de forma que la composición es uno de los elementos útiles en la identificación y clasificación de las obras.

Selección de variables y comprobación de su validez

Una vez identificados los hechos diferenciales hacía falta encontrar la forma de medirlos a partir de copias de las imágenes. Estas copias se obtuvieron de dos fuentes:

- Bibliografía de editoriales especializadas en gran formato, de las cuales se escanearon las obras seleccionadas.

- Páginas web de los museos de mayor prestigio del mundo y de enciclopedias electrónicas especializadas en arte.

Para cada hecho diferencial se estudiaron una o más variables y se sometieron a prueba modificando factores como el origen de la imagen, su tamaño, la proporción entre la imagen real y la tratada, y la resolución y el tiempo de captura cuando se utilizaba escáner.

Finalmente, las variables que superaron esta prueba, demostrando ser poco sensibles al origen y el formato de la imagen, fueron las que se utilizaron para los estudios posteriores. Las variables son las siguientes:

***Grado de mezcla**:* Variable continua que expresa la relación porcentual del número de colores diferentes que contiene la imagen después de saturarla al 100%, respecto al número de colores de la imagen original. Da idea del nivel de mezcla de los pigmentos o de la aplicación directa de los colores sin mezcla previa. Esta variable es medible con el software comercial *Image Analizer*.

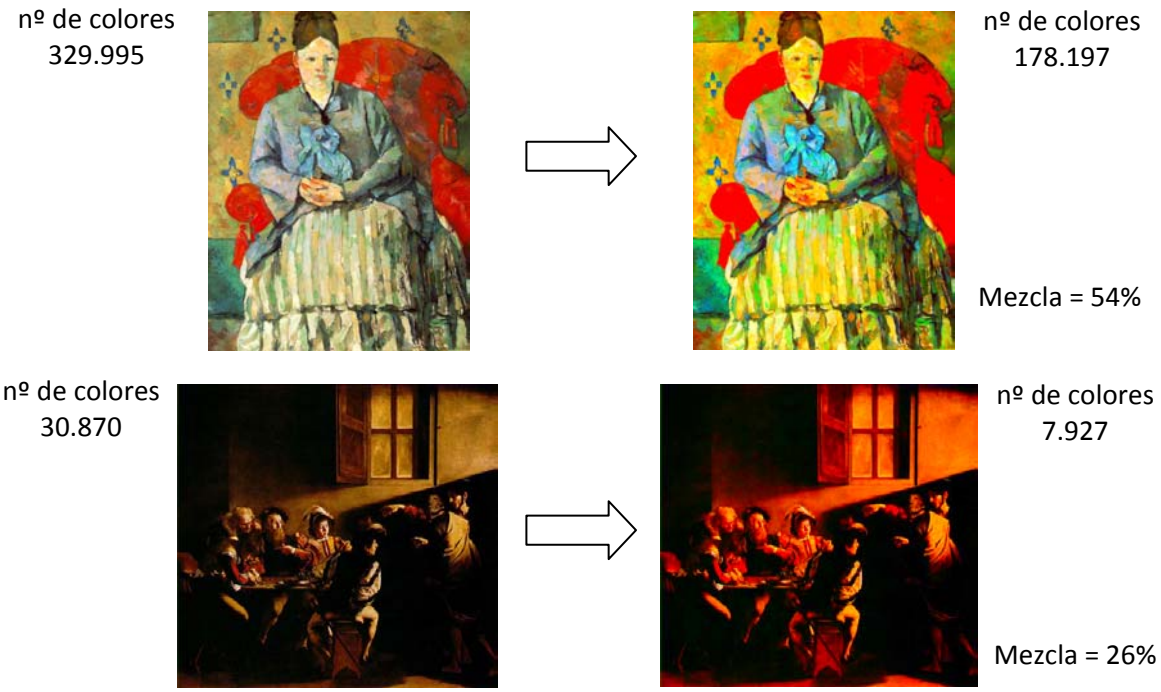


Figura 2: *Ejemplos de uso colorista y valorista, determinados con el método de la saturación, y valores de la variable mezcla*

Luminosidad: Variable continua que corresponde a la media de la intensidad de luz de los pixeles de la imagen. Es medible con el software comercial *Adobe Photoshop*.

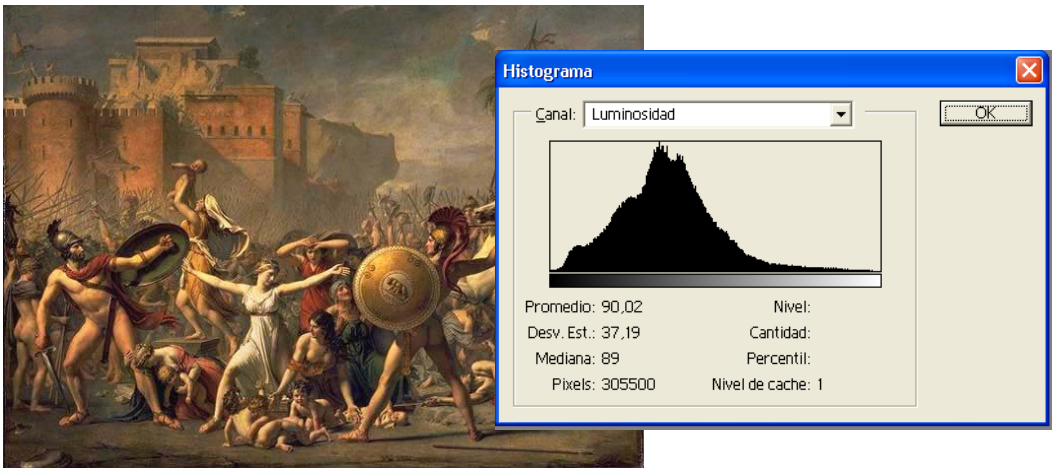


Figura 3: *Las Sabinas de David y su histograma de luminosidad*

Contraste: Corresponde a la desviación estándar (una medida de la variabilidad) de la luminosidad e identifica la propensión al uso de contraposiciones valoristas (luz-sombra). Es medible con *Adobe Photoshop*.

Tonos pastel: Mide el porcentaje de píxeles que tienen tonos formados por cualquier mezcla de un color saturado y blanco. Esta variable se determina a través de la medida de los componentes S (saturación) e I (intensidad o brillo) al obtener las imágenes en el sistema HSI. Es considerado tono pastel aquél que tiene un grado de saturación entre 20-80% y un brillo entre 80-100% simultáneamente. La medida se hace a través del muestreador de colores de *Adobe Photoshop.*

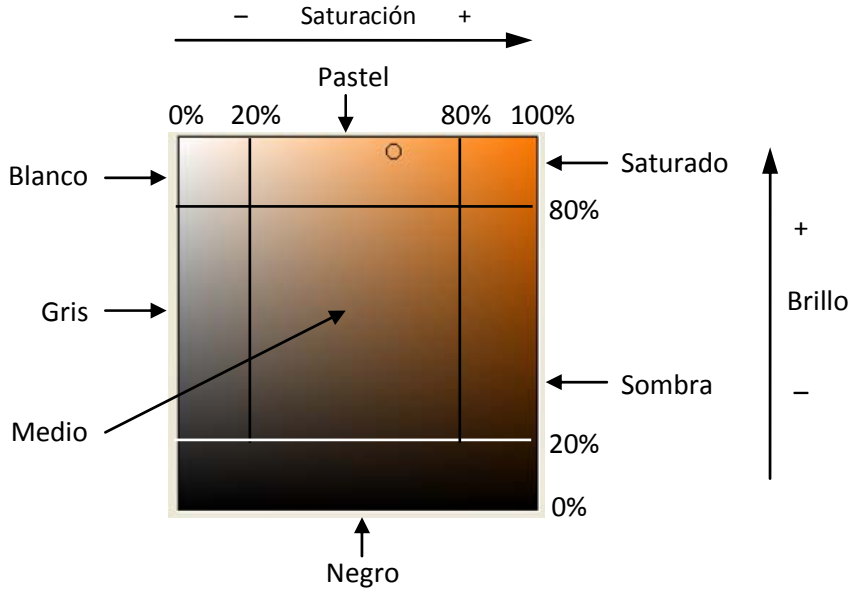


Figura 4: *Cuadro de tonalidades para un matiz constante*

Tonos sombra: Mide el porcentaje de píxeles que tienen tonos formados por cualquier mezcla de un color saturado y negro. Esta variable se determina también a través de la medida de los componentes S (saturación) e I (intensidad o brillo) al obtener las imágenes en el sistema HSI. Es considerado tono sombra aquél que tiene un grado de saturación entre 80-100% y brillo entre 20-80% simultáneamente.

Tonos medios: Mide el porcentaje de píxeles que presentan colores cumpliendo una saturación 20-80% e intensidad 20-80% simultáneamente.

Tonos saturados: Mide el porcentaje de píxeles que tienen tonos de color puro sin adición de blanco, negro ni gris que le reste saturación. Quedan clasificados como saturados los tonos que tengan un valor de los parámetros S e I entre el 80 y el 100% (cuadrado del extremo superior derecho de la Figura 4).

Matiz de la máxima saturación: Debido a que la máxima saturación que se puede obtener de un pigmento proviene de una sustancia en la que no hay mezcla, para cada obra se ha evaluado cuáles son los matices de máxima saturación ya que pueden ser los promotores de la gama completa de tonalidades utilizadas. Esta evaluación se ha realizado a través del valor máximo del canal S al obtener las imágenes en el sistema HSI. De esta forma se ha extraído tanto el valor del matiz máximo saturado (H), como el propio valor de saturación (S). Esta medición también se ha realizado a través del muestreador de colores de *Adobe Photoshop.*

Gradación: Ésta es una variable cualitativa que identifica diferentes formas de graduar los colores para representar la luz y la sombra. Las categorías que se han definido para esta variable son:

MSN: Medio-Sombra-Negro

SSN: Saturado-Sombra-Negro

PMN: Pastel-Medio-Negro

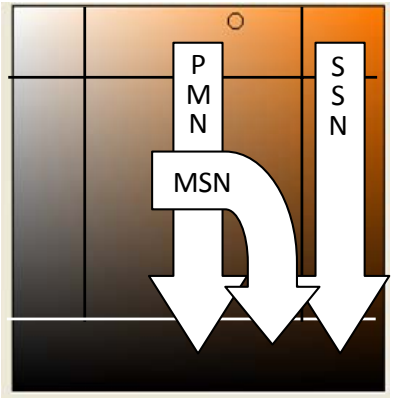


Figura 5: *Gradaciones valoristas principales*

Gamut cromático: En este trabajo se creó un gráfico específico para identificar la paleta del pintor dentro de una de las siguientes clases: monocroma, bicolor, tríada básica, tríada secundaria o policromía. Este gráfico ha sido denominado gamut por analogía con los gráficos que representan el espacio visual reproducible por un determinado dispositivo electrónico (monitores de ordenador, pantallas de televisión, etc.). Se trata de un gráfico de coordenadas polares que contiene la información del matiz de color (en grados) y del nivel de saturación (lineal sobre el radio).

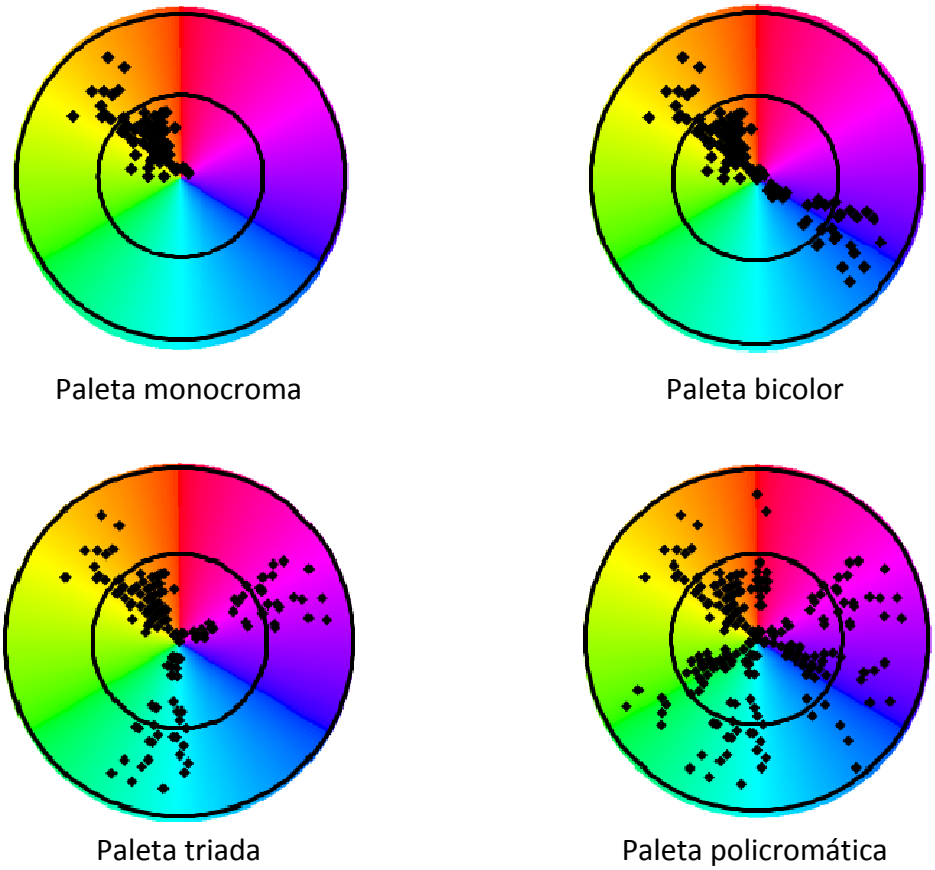


Figura 6: *Gamuts para los diferentes tipos de paleta*

Tipo de contorno de los objetos: Toda la pintura de tipo valorista tiene en común un tipo de gradación que engaña nuestro cerebro con el fin de simular aquello que se quiere representar. Sin embargo, la pintura de estilos próximos al impresionismo, postimpresionismo y fauvismo tiene en común un tipo de gradación y de pincelada distinguible del anterior por el hecho de generar los colores de forma retiniana (dentro del ojo, sin mezcla previa sobre el paleta del pintor). Basado en algoritmos aplicados al campo de la visión artificial que permiten a los robots distinguir a los objetos de sus sombras, se realizó un programa en MatLab que, a partir del canal de luminosidad de una imagen digital, la clasifica en uno de los grupos llamados contorno figurativo, contorno no figurativo y contorno artificial.

Figura 7: *Mujer reclinada de Boucher original y después de aplicar la segunda derivada donde puede apreciarse que el resultado es un dibujo de los contornos claramente figurativo*

Perspectiva: Variable que identifica el mecanismo utilizado por el autor para dar sensación de tridimensionalidad. La perspectiva puede ser lineal (basada en líneas rectas que convergen hacia uno o dos puntos focales sobre el horizonte), aérea (fundamentada en el uso de contrastes altos en primeros planos y leves en la lejanía, teniendo en cuenta la existencia de gradaciones de luz valoristas) o cromática (basada en relaciones caliente-frío para simbolizar cerca-lejos).

Figura 8: *Ejemplos de perspectiva lineal, aérea (valorista) y cromática*

Resultados del estudio de datación de obras

Mediante la aplicación de la regresión lineal múltiple se ha obtenido un modelo estadístico que permite conocer el año de realización de una obra con un error inferior a 50 años para un nivel de confianza del 95% (en el 95% de los casos, el valor real es el previsto por el modelo con un error de ±50 años).

Este modelo ha sido probado con un juego de datos externo, que no se ha utilizado para la construcción del modelo, y que está constituido por el 25% del total de obras muestreadas. El resultado es satisfactorio y muestra que el modelo funciona correctamente con cualquier pintura que corresponda al periodo estudiado, aunque "no la hayamos visto nunca antes".

Los requisitos sobre los que se apoya el modelo lineal se cumplen satisfactoriamente (linealidad, normalidad, variancia constante e independencia). Su capacidad de explicación (medida a través del coeficiente de determinación) es del 96%.

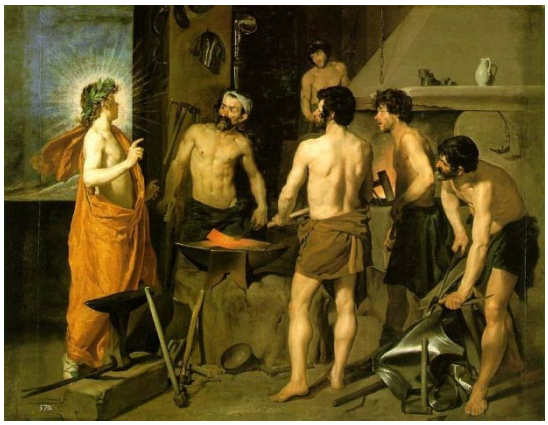
Concretamente el modelo resultante tiene como respuesta el año de realización de la obra y como variables explicativas el grado de mezcla, el tipo de perspectiva, el tipo de gradación y el tipo de contorno.

$$\text{Año} = 1694 + 0{,}982\ \text{Mezcla} - 45{,}5\ \text{CF} + 27{,}5\ \text{CFA} + 125\ \text{PC} + 51{,}2\ \text{PA} - 125\ \text{MON} - 69{,}9\ \text{SON} + 44{,}8\ \text{PMN}$$

Donde CF: contorno de tipo figurativo
CFA: contorno de tipo artificial
PC: perspectiva cromática
PA: perspectiva aérea
MON: gradación medio-sombra-negro
SON: gradación saturado-sombra-negro
PMN: gradación pastel-medio-negro

La variable mezcla es continua y su valor se puede obtener mediante el software *Image Analyzer*. Las otras son variables binarias que toman el valor 1 (= Sí) o 0 (= No). La variable CFA se evalúa a través de un programa de MATLAB desarrollado expresamente para este trabajo. Los valores de la perspectiva se obtienen a partir de la observación del cuadro y según la definición de cada tipo. Finalmente, si la gradación utilizada por el pintor para representar la luz y la sombra corresponde a alguno de los tres tipos que forman parte del modelo se valora a través de medidas que se pueden realizar con *Adobe Photoshop*.

La Figura 9 muestra tres ejemplos de aplicación del modelo para estimar el año en que fue realizada una obra.



La Farga de Vulcano. Velázquez, 1630

Mezcla:	28
Contorno tipo figurativo:	Sí (=1)
Contorno tipo artificial:	No (=0)
Perspectiva cromática:	No (=0)
Perspectiva aérea:	Sí (=1)
Gradación Medio-Sombra-Negro:	No (=0)
Gradación Satur.-Sombra-Negro:	Sí (=1)
Gradación Pastel-Medio-Negro:	No (=0)
Año previsto por el modelo:	1657
Error de previsión:	+ 27 años

Los fusilamientos del 3 de mayo. Goya, 1814

Mezcla:	39
Contorno tipo figurativo:	No (=0)
Contorno tipo artificial:	No (=0)
Perspectiva cromática:	No (=0)
Perspectiva aérea:	Sí (=1)
Gradación Medio-Sombra-Negro:	No (=0)
Gradación Satur.-Sombra-Negro:	No (=0)
Gradación Pastel-Medio-Negro:	Sí (=1)
Año previsto por el modelo:	1828
Error de previsión:	+ 14 años

Interior con estuche de violín. Matisse. 1919

Mezcla:	71
Contorno tipo figurativo:	No (=0)
Contorno tipo artificial:	Sí (=1)
Perspectiva cromática:	Sí (=1)
Perspectiva aérea:	No (=0)
Gradación Medio-Sombra-Negro:	No (=0)
Gradación Satur.-Sombra-Negro:	No (=0)
Gradación Pastel-Medio-Negro:	No (=0)
Año previsto por el modelo:	1916
Error de previsión:	–3 años

Figura 9: *Ejemplos de aplicación del modelo a tres cuadros de diferentes épocas*

Utilidad del modelo para la datación de obras

Este modelo permite fechar con un margen de error conocido cualquier pintura al óleo realizada entre los siglos XV y XIX. Pero esta no es su única utilidad, ya que también permite identificar qué autores se adelantaron a su época; es decir, qué artistas fueron innovadores frente a los que aplicaron tratamientos más conservadores.

Se han identificado autores para los que el modelo prevé sistemáticamente un año inferior al real de sus obras. Entre éstos tenemos a Dominique Ingres y a Jacques-Louis David, hecho esperable teniendo en cuenta que pertenecen al movimiento neoclásico en el cual predomina el dibujo sobre el color con una clara tendencia a reencontrar estilos de épocas clásicas.

Por otra parte, también se han identificado autores para los que el modelo prevé sistemáticamente años posteriores a los reales. Esto ocurre cuando los autores aplican tratamientos adelantados a su tiempo, y los podemos considerar como innovadores que han marcado alguno de los puntos de inflexión en la evolución de la técnica pictórica. Por ejemplo, Leonardo es un aplicador precoz de la perspectiva aérea, Tiziano fue un innovador en el uso del frotis (extensión de la pintura en una capa muy fina) que es detectado como transición entre el valorismo y la mezcla retiniana para la variable contornos, Turner presenta claros síntomas preimpresionistas y Monet es el precursor de las gradaciones con uso predominante de tonalidades pastel e inexistencia del negro.

Esta misma técnica aplicada a la obra pictórica de un mismo autor nos da una buena idea de sus diferentes fases evolutivas en el tiempo: fase de aprendizaje, académica o imitativa; fase de busca de un estilo propio; cambios posteriores en el estilo por influencia de otros coetáneos, etc.

Un buen ejemplo para estudiar la evolución técnica de un pintor se ha encontrado en las obras de Monet. Este pintor muestra una fase académica donde aplica los recursos y técnicas que le enseñaron en la escuela de aquel tiempo, posteriormente realiza un salto con notables avances tanto en la extensión cromática de la paleta como en el uso de tonos pastel, supresión de tonos sombra y del negro, cambios de mecanismo en la perspectiva (menos lineal y aérea, más cromática) y también pérdida de los contornos figurativos. Finalmente, Monet aplica técnicas de cariz fauvista cuando este movimiento todavía no se ha puesto de manifiesto en los autores que lo liderarán (Matisse, Signac, Derain). Es de resaltar que los resultados de nuestro trabajo respecto al cambio de estilo de Monet muestran una estrecha relación con las conclusiones del artículo científico de Ralf Dahm "*Painting the world with different eyes*" (Max-Planck-

Research, Núm. 3/2002), donde se da como explicación del repentino cambio de estilo de este pintor el problema de cataratas que sufrió en aquella época.

Estudio de los movimientos artísticos

Siguiendo con la modelización estadística y utilizando la regresión logística ordinal de respuesta politómica, se ha obtenido un modelo que permite clasificar las obras de arte en una de las siguientes corrientes artísticas: Renacimiento, Barroco, Rococó, Neoclasicismo, Romanticismo, Impresionismo, Postimpresionismo y Fauvismo.

Este modelo tiene como variables explicativas: el grado de mezcla, la paleta (gamut) y la composición. Produce clasificaciones correctas en el 89% de las obras y tiene un funcionamiento correcto con un juego de datos externo (no utilizados en la construcción del modelo) que suponen un 25% del total de las consideradas en este trabajo.

También se ha estudiado un segundo modelo, considerado más innovador desde el punto de vista estadístico, utilizando la regresión logística ordinal sobre las componentes obtenidas por mínimos cuadrados parciales, a partir de las variables continúas: grado de mezcla, tonos pastel, tonos saturados, tonos medio y luminosidad. Este segundo modelo también ha superado una validación con datos externos y tiene una capacidad de clasificar correctamente ligeramente superior al anterior, un 93%.

Resumen y conclusiones

A través del tratamiento informático de las imágenes digitalizadas ha sido posible descubrir algunos mecanismos que marcan el carácter de un pintor y los rasgos que diferencian unos movimientos de otros. Estos mecanismos tienen medidas objetivas válidas que permiten ser utilizadas para construir modelos estadísticos.

Este trabajo, desarrollado durante aproximadamente 2 años sobre una pinacoteca de más de 200 óleos pintados entre los siglos XV y XIX, muestra la existencia de mecanismos identificables objetivamente posibilitando fechar obras con un error de 50 años dentro del periodo estudiado de 5 siglos, potenciar los estudios de autoría dudosa (identificación de falsificaciones) y objetivar los rasgos diferenciales de un movimiento artístico o de un autor específico. Además, abre también nuevas vías de investigación ampliando conocimientos sobre los autores, especialmente a través de los cromogramas y el análisis de la pincelada. También abre camino en el campo de la

reconstrucción de la imagen original de un cuadro que ha quedado sometido a procesos de degradación que han causado el viraje de determinados colores.

(Proyecto de la Licenciatura de Estadística, presentado en junio de 2006 con el título "Aplicació de tècniques estadístiques a l'estudi de l'art pictòric dels segles XV-XIX –del Renaixement al Fauvisme-")

Este proyecto cierra la trilogía "Literatura, música, pintura" y estadística. Por lo que sabemos, es un proyecto pionero en la creación de un modelo estadístico para fechar pinturas. Probablemente, los avances en el campo de la cuantificación del aspecto visual harán que estas técnicas se puedan aplicar cada vez con más precisión. El camino está abierto.

Miquel Romero estudió ingeniería técnica química en la UPC porque empezó a trabajar en la industria farmacéutica y le gustaba ese mundo de las moléculas. Después realizó la Licenciatura en Estadística, también en la UPC, porque se dio cuenta de que le era imprescindible tanto en la investigación de nuevos productos, como en la mejora de los procesos ya desarrollados. Actualmente es el responsable del aseguramiento de la calidad de los medicamentos en los Laboratorios Almirall, y desarrolla también una actividad docente como profesor de Estadística en Tecnología Farmacéutica en la Facultad de Farmacia de la Universidad de Barcelona.

Su pasión por la pintura le viene de muy pequeño. Nos explica que empezó a pintar al óleo a los 7 años y ha hecho diversas exposiciones. "La pasión por la pintura y por el tratamiento de datos ha hecho que dedicara tiempo a encontrar vínculos... ¡y los hay"!

Conocimiento y protección del medio ambiente

4

Influencia de las condiciones climáticas en el crecimiento del pino silvestre en Cataluña

Proyecto realizado por**: Natàlia Adell Calvet**
Dirigido por: **Llorenç Badiella Busquets**
Tutelado per**: Josep Anton Sánchez Espigares**

El cambio climático y sus posibles consecuencias es un tema importante que despierta muchas discusiones y controversias, y en el cual el uso de la estadística es necesario para entender lo que realmente está ocurriendo.

Entre muchos otros, el cambio climático podría tener efectos sobre el crecimiento de los árboles. Hay factores, como el aumento de la sequía, que juegan en contra y otros, como el aumento del dióxido de carbono (CO_2), que parece que juegan a favor a corto plazo, aunque sus efectos a largo plazo no están claramente definidos.

Este estudio, realizado por el Servicio de Estadística de la Universidad Autónoma de Barcelona, por iniciativa del Centro de Investigación Ecológica y Aplicaciones Forestales de la misma Universidad, y en el que ha participado activamente la autora del proyecto, analiza el crecimiento de una muestra de árboles de pino silvestre (Pinus sylvestris) en Cataluña y estudia la relación que tiene su crecimiento con variables ambientales como la temperatura, las precipitaciones y la concentración de CO_2.

Planteamiento general y objetivos del estudio

Valorar de forma científica el impacto de los cambios climáticos en la naturaleza en general, y en el crecimiento de los árboles en particular, es una tarea importante para conocer las posibles implicaciones que puede tener una actitud descuidada frente al mundo que nos rodea.

Los principales factores que se cree que pueden modificar el crecimiento de los árboles son el llamado cambio climático y la cantidad de dióxido de carbono presente en el ambiente (factores que seguramente están relacionados). Por una parte, se sabe que el cambio climático provoca una mayor sequedad en el ambiente, hecho que podría tener un efecto negativo en el crecimiento de los árboles. Por otra parte, debido a que los árboles absorben CO_2 para producir celulosa y madera, también podría ser que el aumento de CO_2 en la atmósfera fuera beneficioso para su crecimiento.

Para profundizar en estos temas con datos de nuestro entorno, el Centro de Investigación Ecológica y Aplicaciones Forestales (CREAF) planteó este estudio con el apoyo del Servicio de Estadística de la Universidad Autónoma de Barcelona. La autora del proyecto ha participado directamente en el planteamiento y en la realización de los análisis estadísticos. Los objetivos que se plantean son:

- Detectar si el patrón de crecimiento de los árboles ha cambiado a lo largo del periodo en que se tienen datos.
- Detectar si el efecto del cambio climático ha provocado una variación en la tasa de crecimiento del pino silvestre en Cataluña.
- Observar si hay diferencias en el crecimiento de los árboles en las diversas parcelas (diferentes regiones de Cataluña) en que se han recogido los datos.

Para alcanzar los objetivos planteados se parte de una base de datos de pino silvestre de Cataluña (que se comentará más adelante), y a partir de ella se construyen modelos estadísticos para explicar la evolución del crecimiento en función de determinadas variables climatológicas.

¿Por qué el pino silvestre?

El pino silvestre (*pinus sylvestris*) es una de las especias de árboles más extensamente distribuidas de la Tierra. Aunque las mayores poblaciones de esta especie se

encuentran en regiones boreales, el pino silvestre también ocupa grandes zonas en regiones relativamente secas de la cuenca mediterránea, desde la Península Ibérica hasta Turquía. Las poblaciones españolas de este tipo de pino constituyen el límite sur del rango de la especie, por lo cual son especialmente interesantes para estudiar los impactos del cambio climático.

Recogida de datos

Para estudiar el crecimiento de los árboles, una de las magnitudes más útiles es la distancia entre los anillos que se observan al hacer un corte transversal (Figura 1), ya que estas distancias contienen información ecológica e histórica del periodo de vida del árbol.

La formación de estos anillos se produce de una forma rítmica, ya que los árboles crecen en un momento determinado del año, hacia la primavera, en la época en que empieza la sequía. Los anillos recogen información de los estímulos ambientales que afectan a las funciones fisiológicas del árbol (como la temperatura, las precipitaciones...), tanto del año en curso como del anterior. Por este motivo esta medida resulta muy interesante e informativa del crecimiento.

Figura 1: *Sección transversal de un tronco donde se observan los anillos*

Origen de los datos

La colección de anillos utilizada para este estudio fue obtenida entre 1988 y 1998 por el CREAF como parte del Inventario Ecológico y Forestal de Cataluña. Este inventario incluye un total de 10.664 parcelas distribuidas aleatoriamente a lo largo de toda el área forestal catalana (1.214.408 ha de bosque). El pino silvestre está presente en 3.219 de estas parcelas (un 30,2%), y es la especie dominante en 1.962 (18,4% de las

parcelas muestreadas). El área total forestal estimada de pino silvestre en Cataluña es de 219.754 ha, representando la segunda especie más abundante de árbol en la región, después del pino carrasco (*Pinus Halepensis*).

Proceso de medida

Para medir las distancias entre anillos se utiliza una herramienta denominada barrena de Pressler (Figura 2 izquierda). Con esta herramienta se realiza un agujero en el árbol, a unos 1,3 m del suelo, y se extrae un testigo de madera. El testigo consiste en una pequeña muestra del tronco del árbol en la cual se pueden observar sus anillos. Esta muestra se guarda en una protección rígida, también de madera, con el fin de mantener su forma recta y así poder realizar las medidas correctamente, tal como se muestra en la Figura 2 (fotografía de la derecha).

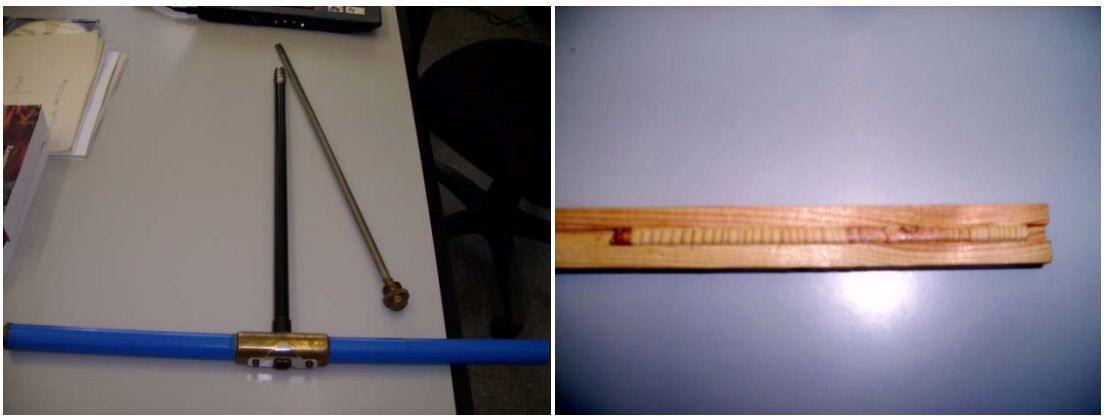


Figura 2: *Barrena de Pressler (izquierda) y testigo de madera (derecha)*

Sobre este testigo de madera se realiza la medición de la distancia entre anillos, que corresponde al aumento de grosor del árbol cada año. Desde el año en que se toma la muestra se va retrocediendo, anillo por anillo, obteniendo así la medida del grueso del anillo para cada año de vida del árbol. En la Figura 3 se puede observar el procedimiento de obtención de la medida "grueso de crecimiento anual" de una forma esquemática y más visual.

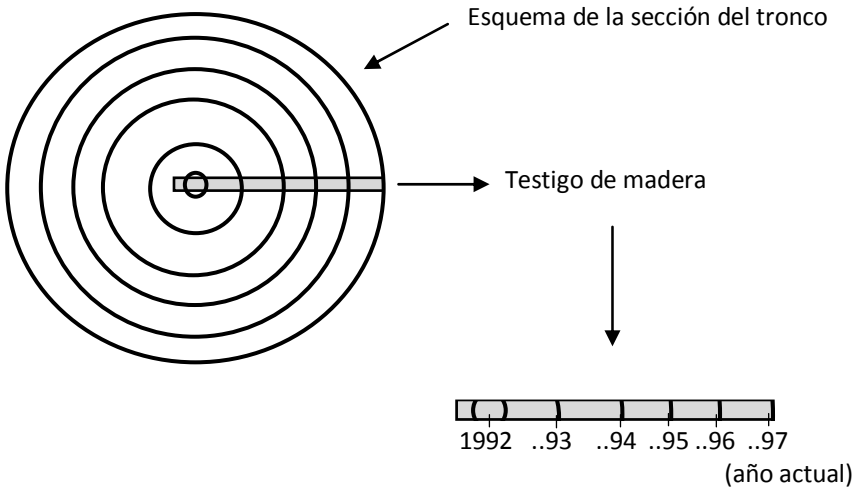


Figura 3: *Esquema de obtención de los testigos de madera y de la medición de los anillos*

Dificultades en la recogida de los datos

El hecho de que el tronco de los árboles no sea una circunferencia perfecta introduce un cierto error en la medida, ya que al recoger la muestra es posible que se realice por la parte más estrecha o la más gruesa del árbol (Figura 4, izquierda). La ventaja del estudio del pino silvestre es que su tronco tiene una forma bastante redonda pero, aun así, hay que comprobar que la distancia hasta el centro es aproximadamente la mitad de su diámetro.

Otro inconveniente que se ha encontrado cuando se recoge la muestra es que resulta difícil apuntar al centro del árbol, sobre todo en los árboles viejos ya que éstos tienen un diámetro superior a los jóvenes. Esta dificultad hace que exista la posibilidad de que no se obtenga la información relativa a los anillos correspondientes a los primeros años de vida del árbol (Figura 4, derecha) y no se pueda utilizar la información de este árbol. Esto obliga a ser muy cuidadoso en el proceso de toma de muestras.

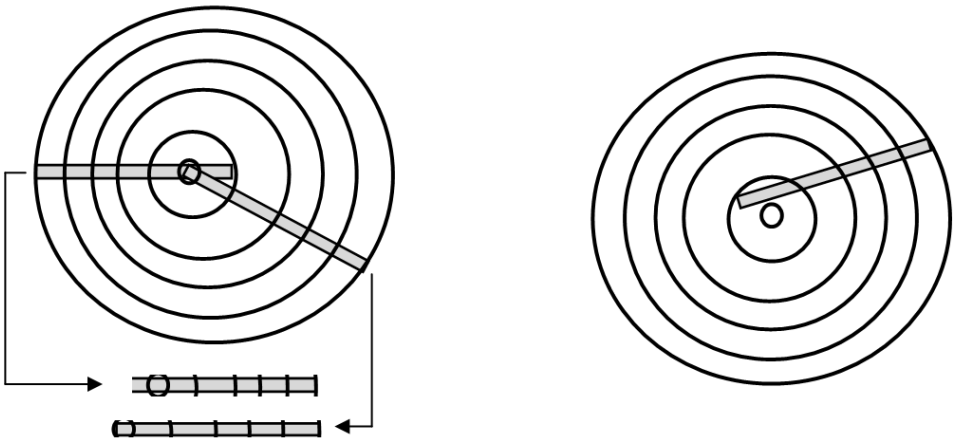


Figura 4: *Posibles errores debidos a la estructura del árbol y a la dificultad de apuntar al centro*

Datos obtenidos

De la base de datos original se seleccionaron aquellas parcelas en las que el pino silvestre era la especie dominante y se habían muestreado al menos 4 ejemplares. Después de validar y depurar los datos se dispone de un total de 548 árboles distribuidos en 135 parcelas.

Para cada árbol se tiene una serie temporal con el incremento anual del radio de los anillos. Además, también se dispone de su altura y diámetro en el momento de ser incluido en la muestra. A partir de las medidas de los anillos se calculó, para cada año, la llamada área basal (área entre anillos) y su diámetro correspondiente.

Por otra parte, se dispone de información climática de toda Cataluña durante diversos años, y por parcelas.

De forma resumida, las variables que se contemplan en este análisis se pueden clasificar en los siguientes tipos:

- Variables identificadoras de las observaciones: información referente al árbol en cuestión, como el número identificador, la parcela a la cual pertenece, el año en que se tomó la medida o la edad del árbol.
- Variables dendocronológicas para árboles: variables relativas a las características físicas de los árboles tales como altura, tamaño de los anillos y variables derivadas de estas medidas (área basal y diámetro de cada año de vida del árbol).
- Variables climatológicas generales por años: valores medios anuales de las temperaturas y las precipitaciones, evapotranspiración potencial anual, índice de sequía anual y contenido de dióxido de carbono en el hemisferio norte.
- Variables climatológicas y físicas por parcelas: temperaturas y precipitaciones medias, índice de sequía, pendiente de la vertiente y litología (estado del suelo).
- Variables dendocronológicas por parcelas: información en lo referente a si hay evidencias de gestión humana (por ejemplo, talas de árboles).

Análisis descriptivo

Datos de los árboles

El análisis gráfico de los datos de los árboles confirma el comportamiento esperado: la altura tiene una relación lineal con el diámetro (Figura 5), pero sorprende el hecho de que los diámetros tiendan a agruparse en torno a valores concretos con incrementos

de 5 cm (12, 17, 22, 27, 32 cm) lo que hace pensar en alguna peculiaridad del proceso de medida, o a que se tendieron a escoger árboles de unos gruesos determinados.

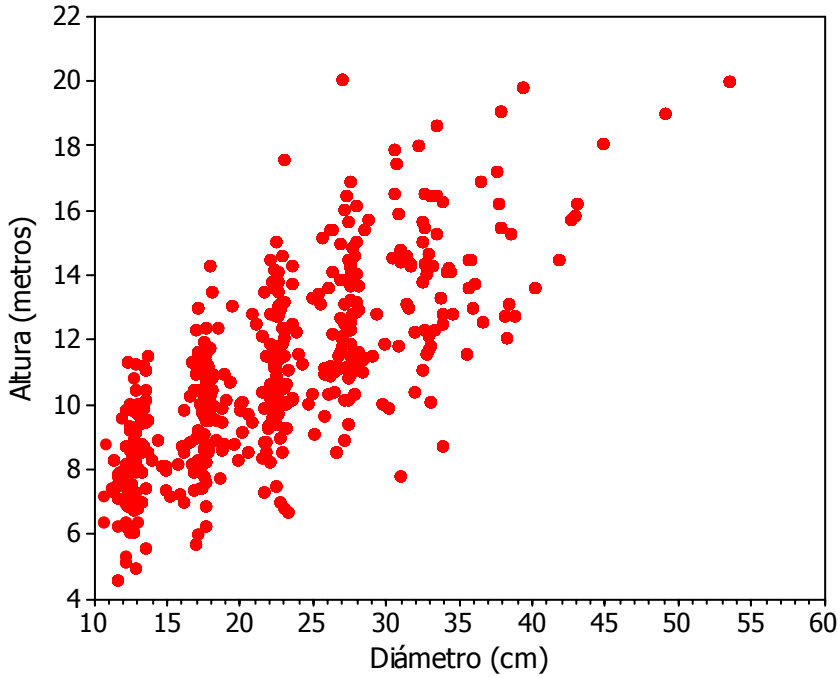


Figura 5: *Diagrama bivariante entre la altura y el diámetro de los arboles*

La relación entre altura y edad del árbol se puede ver en la Figura 6. Se observa una relación creciente hasta una cierta edad en que la altura se estabiliza. Podría parecer que la altura tiene un máximo en torno a los 50-60 años y después vuelve a bajar, pero esto, evidentemente, no es cierto. El número de árboles de más de 70 años es escaso, y también lo son los árboles de más de 15 metros de altura tal como se puede ver en los histogramas marginales de la misma figura. Por esta razón, no tenemos árboles viejos y altos en la muestra.

Se puede realizar también un estudio transversal con el fin de observar las características de los árboles en el momento en que tenían 5, 10, 15, 20 y 30 años de edad. Se observa que cuanto más viejos son los árboles, el grueso de los anillos disminuye. En cambio, respecto del diámetro, al ser acumulativo, cuanto más viejos son más diámetro tienen (Figura 7).

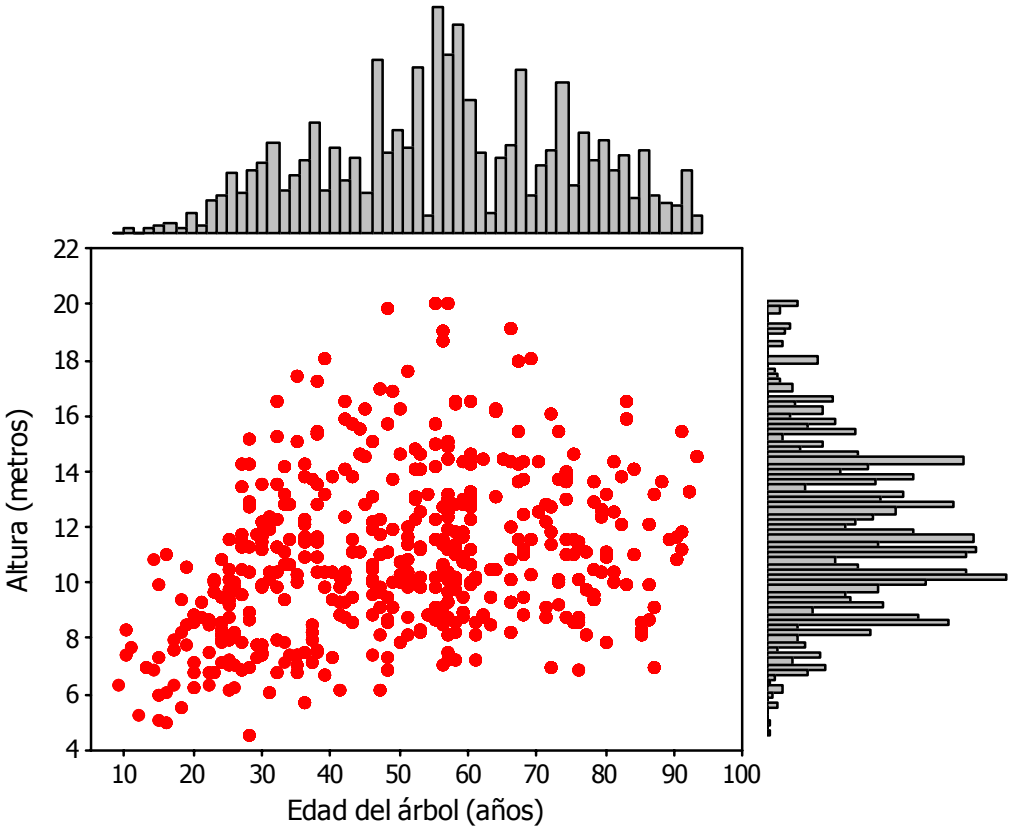


Figura 6: *Representación de la altura respecto a la edad de los árboles en el momento del muestreo*

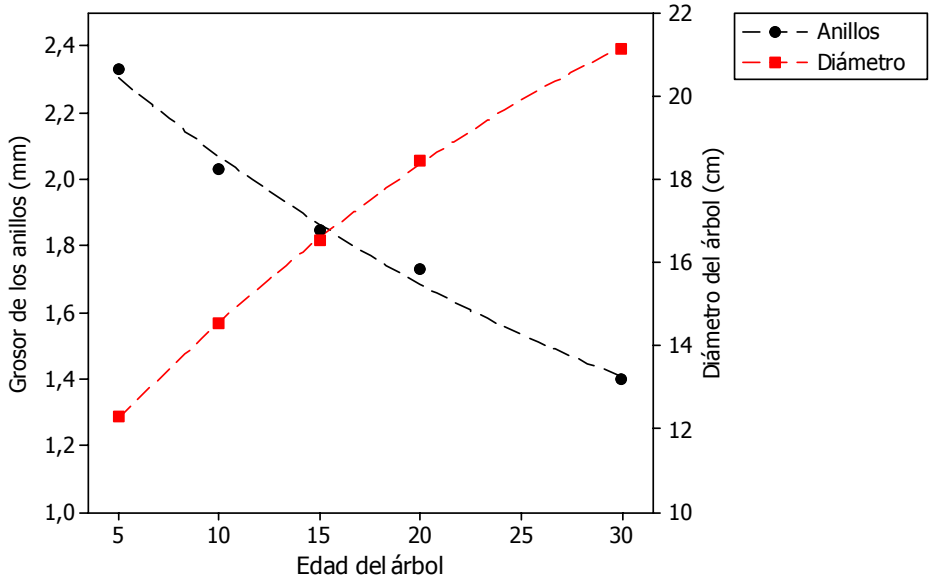


Figura 7: *Valores del grosor de los anillos y el diámetro medio de los árboles de la muestra a las edades de 5, 10, 15 20 y 30 años. Se han añadido también las curvas de regresión que mejor se ajustan a estos puntos*

La Figura 8 muestra la evolución del diámetro para un conjunto de árboles de la muestra (los que tienen un número de identificación múltiplo de 20) y se puede ver que el diámetro crece continuamente, a pesar de que no siempre al mismo ritmo.

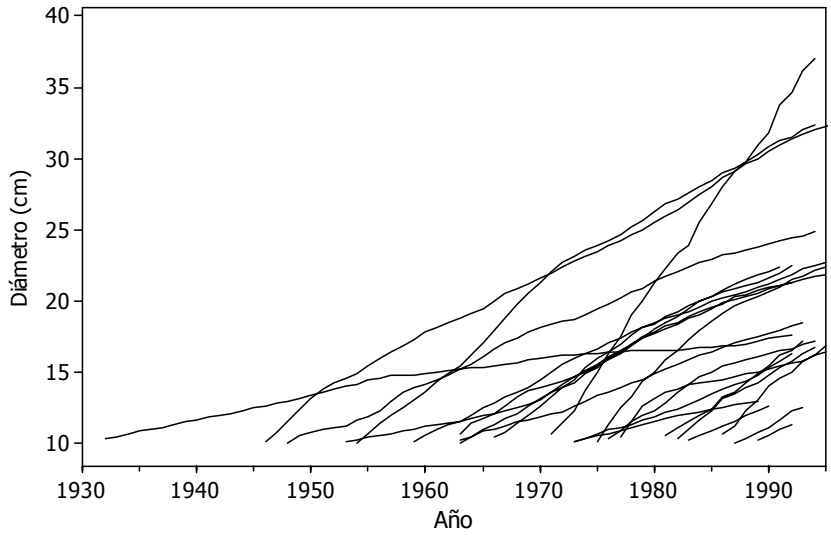


Figura 8: *Serie temporal de les variables longitudinales*

Datos climatológicos

Se dispone de los datos climatológicos de casi un siglo. De la medida de la concentración de CO_2 se ha conseguido tener los datos entre los años 1900 y 1997 y del resto de variables climatológicas se dispone de los datos entre 1901 y 1997 (ambos incluidos).

La temperatura media anual muestra una tendencia claramente creciente. Para analizar si hay diferencias importantes entre parcelas, éstas se han dividido en 3 grupos, aproximadamente del mismo tamaño, según sea su temperatura media: un grupo está formado por las parcelas que tienen una temperatura media por debajo de 8,35°C, otro con temperaturas medias de entre 8,35 y 8,73°C, y un tercero con las parcelas que han tenido una temperatura media por encima de 8,73°C. La Figura 9 muestra la evolución de las temperaturas en estos 3 grupos de parcelas. Las diferencias entre ellas no son relevantes, pero la tendencia creciente está clara.

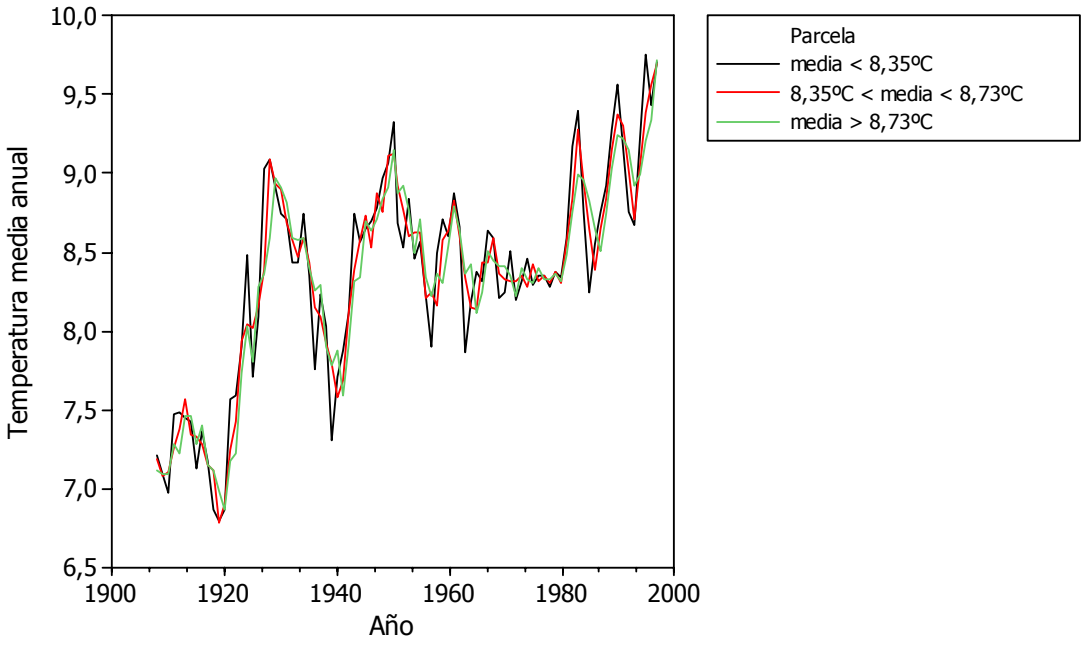


Figura 9: *Evolución de la temperatura media anual. Les parcelas se han clasificado en 3 grupos según su temperatura media. Cada línea corresponde a un grupo.*

Las precipitaciones y el índice de sequía no muestran ninguna tendencia, pero sí una importante inestabilidad, con años lluviosos seguidos de otros mucho más secos. La Figura 10 muestra la evolución de las precipitaciones en el periodo en que se tienen datos, la media de la precipitación anual ha sido de 960,73 m^3, y sus valores están comprendidos entre 698,78 m^3 y 1.323,4 m^3.

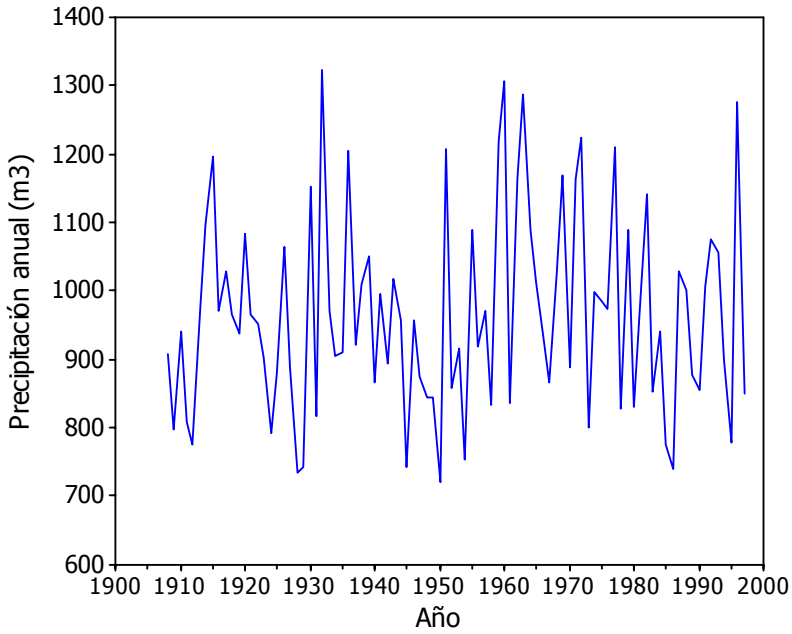


Figura 10: *Evolución de la precipitación anual a lo largo del tiempo*

Como era de esperar, la temperatura está inversamente correlacionada con la precipitación y con el índice de sequía. Es decir, a mayor temperatura menor precipitación y menor índice de sequía. Además, la precipitación está correlacionada, en menor grado, con la pendiente del terreno.

El gráfico de la evolución de la concentración de CO_2 en el hemisferio Norte (Figura 11, datos tomados del *Carbon Dioxide Information Analysis Center*, http://cdiac.ornl.gov) muestra una tendencia muy clara. Hasta el año 1935 hubo un ligero crecimiento, entre 1935 y en 1950 se mantuvo constante, y a partir de 1950 ha habido un aumento muy pronunciado.

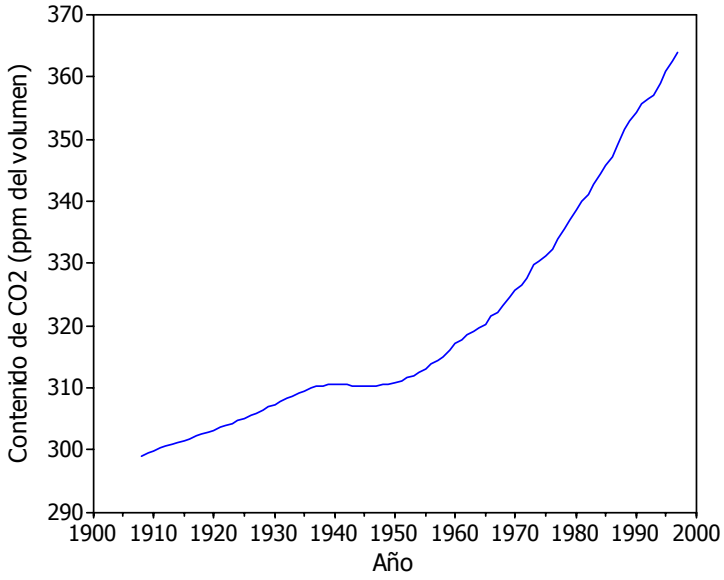


Figura 11: *Evolución de la cantidad de dióxido de carbono en el siglo XX en el hemisferio Norte*

Relación entre las variables ambientales y el crecimiento de los pinos

La creación de un modelo que explique el diámetro de los árboles en función de variables ambientales es una tarea que precisa del uso de técnicas estadísticas avanzadas. Se ha tenido en cuenta que la estructura de los datos es de tipo jerárquico, ya que se dividió Cataluña en 135 parcelas densas en contenido de árboles de pino silvestre y en cada una de ellas se midieron entre 4 y 11 árboles (Figura 12). Es decir, las observaciones entre los niveles (en este caso parcelas o series) son independientes, pero las observaciones dentro de cada parcela son dependientes, ya que provienen de la misma subpoblación. Por tanto, hay dos tipos de variabilidad: la variabilidad entre

las diferentes parcelas y la variabilidad entre los árboles de dentro de una misma parcela.

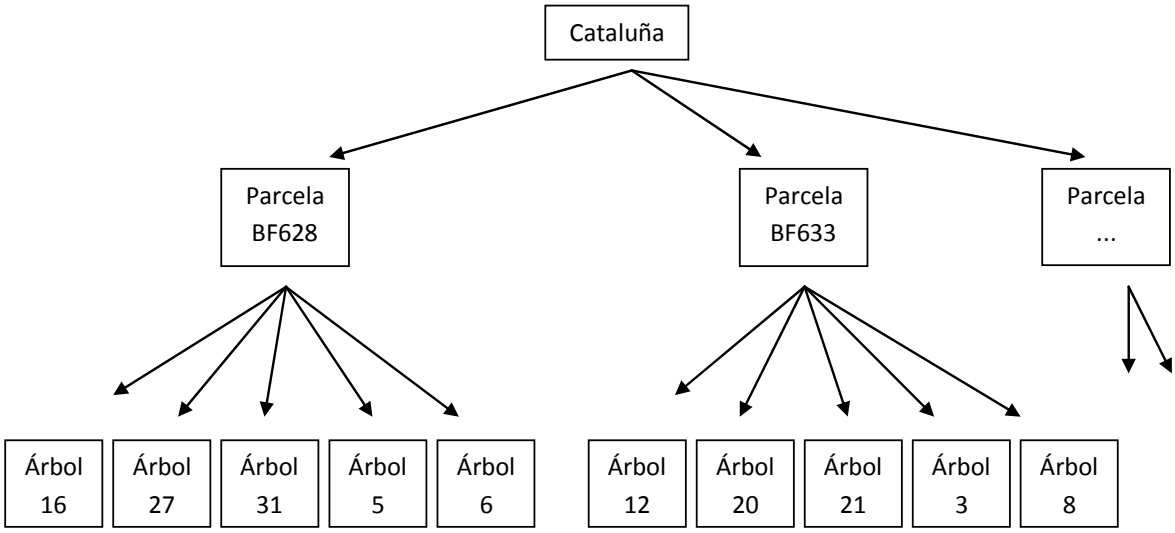


Figura 12: *Estructura jerárquica de los datos*

Con respecto al tipo de modelo, los llamados modelos de efectos mixtos son apropiados para analizar medidas repetidas, datos longitudinales o datos multinivel (es decir, cuando hay presencia de jerarquía) como en nuestro caso. Se han ajustado diferentes modelos, y uno de los más completos es un modelo mixto no lineal que tiene la expresión:

$$y_i = a + (b - a) \cdot e^{-t \cdot c} + \varepsilon_i$$

Donde:

y_i: Diámetro del árbol i
a, b y c: Expresiones que están en función de les variables climáticas
t: Tiempo, en años
ε_i: Parte aleatoria no explicada por el modelo

Analizando el modelo se observa que en condiciones donde hay más concentración de CO_2, más temperatura y más precipitación, los árboles crecen más.

Aunque a priori se podría pensar que los factores asociados al cambio climático, como por ejemplo el aumento de la temperatura y de la concentración de dióxido de carbono, pueden influir negativamente en el crecimiento del pino silvestre, nuestros resultados indican lo contrario.

Estudios complementarios ayudarían a entender mejor esta cuestión. De entre las diversas líneas de investigación que podrían continuar el trabajo presentado en este proyecto, se pueden mencionar las siguientes:

- Obtener los datos climatológicos anuales y por parcela de años anteriores a 1900 con el fin de disponer de más historial, hecho que podría ayudar a estudiar con mayor detalle la evolución que han seguido.
- Analizar el comportamiento de otras especies, como por ejemplo el de pino negro (*pinus uncinata*) en Cataluña para contrastar y comparar los resultados obtenidos.
- También sería interesante estudiar la evolución del crecimiento de estas especies durante el siglo pasado en otros lugares de la península Ibérica y de Europa, con el fin de poder obtener una visión más global de los efectos del cambio climático sobre los bosques.

Como conclusión, podemos decir que los datos analizados ponen de manifiesto la dependencia del crecimiento del pino silvestre de las condiciones climatológicas (temperatura, precipitaciones, concentración de CO_2) en el área estudiada de Cataluña.

(Proyecto del Máster en Estadística e Investigación Operativa, presentado en febrero de 2008 con el título "Models mixtos lineals i no lineals en l'anàlisi de creixement dels arbres de pi roig a Catalunya")

Evidentemente, la estadística es fundamental para poner de manifiesto los efectos del cambio climático. Este trabajo realizado por un grupo interdisciplinar en el que la autora del proyecto formó parte como estadística, ha sido aceptado para su publicación en la revista "Global Change Biology", una de las que tiene más impacto en el ámbito de la biología y el cambio climático.

Natàlia Adell estudió la Diplomatura de Estadística en la UPC y siguió con la Licenciatura, realizando uno de los cursos como estudiante Erasmus en la Universidad de Bath (Inglaterra). También ha realizado el Máster en Estadística e Investigación Operativa de la UPC. Actualmente trabaja como consultora estadística en Servicio de Estadística de la Universidad Autónoma de Barcelona.

5

Buscando productos más limpios y eficaces para luchar contra las plagas de los frutales

Proyecto realizado por**: Lourdes Rodero de Lamo**
Dirigido por**: Josep Ginebra i Molins**

La mosca mediterránea de la fruta se alimenta de multitud de frutos y puede estropear una buena parte de los cultivos frutales. Sus ataques son especialmente graves en los melocotones y las naranjas pero puede afectar también a muchas otras frutas cultivadas como manzanas, peras, albaricoques e higos.

Como solución al ataque de estas moscas y a los insecticidas tradicionales (poco respetuosos con el medio ambiente) se proponen métodos de trampeo. Estos métodos consisten en colocar en los árboles trampas con sustancias atrayentes de forma que las moscas se sientan atraídas y entren. La geometría de las trampas no permite que las moscas puedan salir una vez dentro.

La forma de las trampas ya está muy estudiada, pero no está claro cuál es el atrayente más efectivo. Para seleccionar el mejor entre diferentes alternativas hay que hacer pruebas, y estas pruebas tienen que estar bien pensadas y llevadas a cabo de una forma muy cuidadosa y bien planificada. Después, hace falta analizar los datos obtenidos y sacar conclusiones.

La mosca mediterránea de la fruta

La *ceratitis capitata* o mosca mediterránea de la fruta es originaria de la costa occidental de África y se ha extendido por la mayoría de las zonas cálidas y templadas de todo el mundo. En Cataluña está presente en el litoral sur, es muy habitual en la Ribera del Ebro y más esporádica en la llanura de Lérida y el resto de la franja costera.

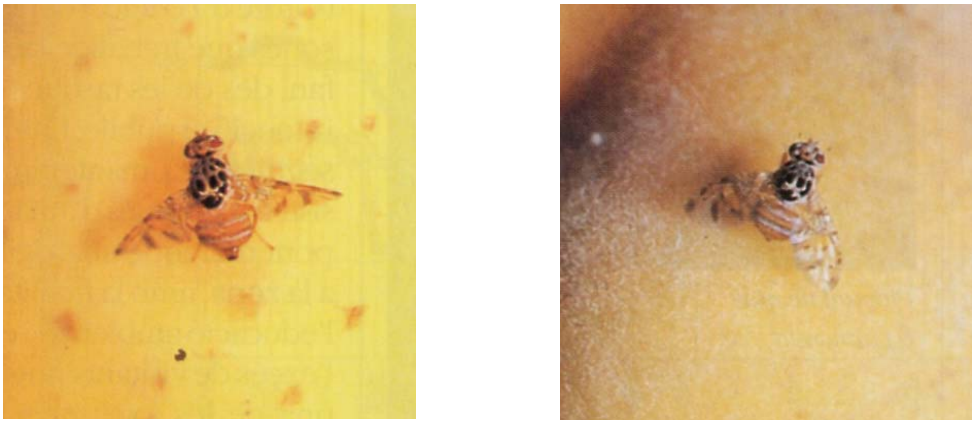


Figura 1: *Macho (izquierda) y hembra de la mosca de la fruta*

La mosca de la fruta es una especie extremadamente polífaga, es decir, que se alimenta de multitud de frutos. A finales de primavera se inicia la emergencia, o celo, de los adultos. Al cabo de tres días las hembras ya empiezan la puesta, depositan los huevos bajo la epidermis de los frutos que están madurando y están a punto de enverar, con una media de 8 huevos por picadura. La larva completa su desarrollo en el interior del fruto en 6 o 7 días, desde donde salta al suelo, se entierra y forma una pupa o larva que entre 8 y 10 días da lugar a un nuevo adulto. En la Figura 2 se puede observar una larva de mosca de la fruta.

Figura 2: *Larva de la mosca de la fruta*

El primer síntoma observable del ataque a los frutos es la picadura que efectúa la hembra al depositar los huevos, que rápidamente es rodeada por un círculo de color amarillo si la picadura es sobre naranjas o mandarinas y de color marrón si la picadura se realiza sobre melocotones o nectarinas. La picadura produce, en primer lugar, una vía de infección de hongos que favorecen el deterioro del fruto.

Figura 3: *Picadas de la mosca de la fruta en un melocotón y en una naranja*

Los daños en los frutos los originan las larvas que al alimentarse de la carne del fruto forman una bolsa con la pulpa totalmente destruida. Estos daños, que pueden pasar desapercibidos durante la cosecha y la manipulación, causan fuertes pérdidas durante la comercialización ya que los frutos afectados se convierten en totalmente inservibles.

Figura 4: *Muestra de los daños de la mosca de la fruta en melocotones*

El uso de trampas

Para evitar que estas moscas estropeen la fruta de los árboles se plantea el uso de unas trampas con sustancias atrayentes en su interior. Este método supone ahorros económicos para los agricultores y hace posible la lucha contra las plagas favoreciendo la preservación del entorno.

Los métodos de trampeo consisten en la colocación de unas trampas en los campos de frutales. Cada trampa contiene una sustancia atrayente que atrae a las moscas hacia ella. Además, las trampas están diseñadas de forma que las moscas no pueden salir y contienen un insecticida que las mata. En la Figura 5 se puede observar un modelo de trampa para capturar la mosca de la fruta.

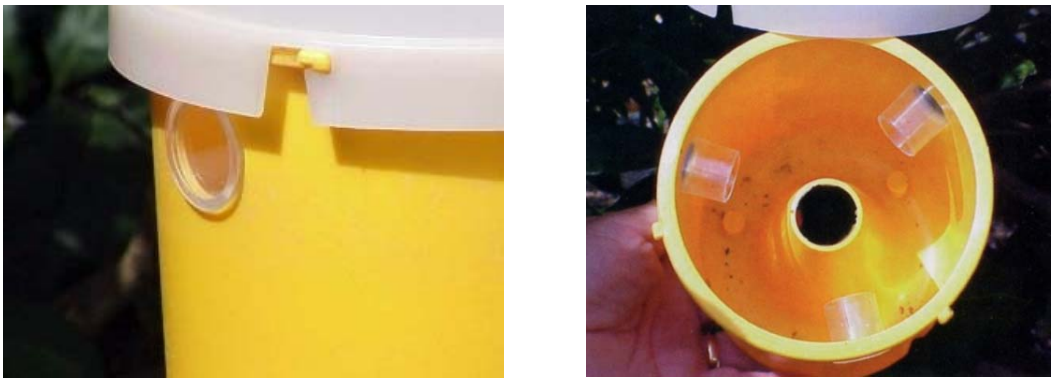


Figura 5: *Ejemplo de trampa para moscas de la fruta*

La geometría de las trampas está muy bien estudiada y existen modelos comerciales baratos y prácticos, pero hay diferentes productos que se pueden utilizar como atrayentes, algunos que se comercializan y otros de fabricación casera. El objetivo del estudio era comparar la eficacia de cuatro atrayentes diferentes que llamaremos A1, A2, A3 y A4. Tres de los atrayentes son experimentales y uno de ellos es comercial.

Diseño del experimento

Cuando se planifica la recogida de los datos hay que estar atento a todos los detalles. Por ejemplo, es posible que en una finca, con unas características específicas de tipo de suelo, humedad, corrientes de aire, etc., un atrayente sea más eficaz que los otros, pero que en otra finca los resultados sean distintos. Para estar seguros de que los resultados no dependen de la finca, el estudio se ha realizado simultáneamente en

cuatro fincas localizadas en el Montsià y en la Ribera d'Ebre (comarcas de Cataluña) que se denominan Barranc, Vergel, Fornòs y Balada. Se considera que estas cuatro fincas dedicadas al cultivo de árboles frutales son representativas de las que hay en la zona.

En cada finca se han colocado tres baterías o hileras de cuatro trampas cada una, una para cada tipo atrayente. Las baterías de trampas son independientes entre sí, es decir, dentro de cada batería, las trampas están separadas unos 40-50 metros para evitar la posible interferencia entre los atrayentes de cada trampa. El orden de colocación de las trampas para cada batería puede ser diferente.

En cada finca, pues, se han colocado doce trampas, que suman un total de cuarenta y ocho si contamos las de las cuatro fincas.

Figura 6: *Trampa para moscas de la fruta en una de las fincas del estudio*

Las trampas se han revisado unas dos veces por semana adaptando los días entre revisión al nivel de capturas de cada momento, de manera que se dejaban pasar más días si no había suficiente número de capturas. En cada revisión se ha contado el número de moscas de la fruta capturadas distinguiendo machos de hembras.

Con el fin de evitar la posible influencia de la ubicación de las trampas en la batería y, en consecuencia, dentro de la finca, cada día de revisión se ha realizado una rotación de todas las trampas; es decir, se han cambiado de posición las trampas de una misma batería. Cada cuatro revisiones, y una vez que todas las trampas han pasado por todas las posiciones disponibles, se cierra un ciclo, y se vuelve a empezar con otro. Se han

realizado un total de seis ciclos con duraciones que van de los nueve días el más corto hasta los dieciocho días el más largo.

Resultados obtenidos

La variable respuesta que se analizará es el número de capturas de moscas hembra. Sólo se tienen en cuenta las moscas hembra porque son éstas las que estropean la fruta al depositar los huevos. La tabla 1 contiene estos datos.

Tabla 1: *Número de moscas hembra atrapadas por ciclo, atrayente y finca*

Ciclo	*Atrayente*	*Finca*			
		Barranc	*Vergel*	*Fornos*	*Balada*
1. Del 15 de junio al 2 de julio (18 días)	*A1*	*215*	*152*	*62*	*612*
	A2	*130*	*105*	*14*	*344*
	A3	*162*	*116*	*23*	*363*
	A4	*247*	*215*	*56*	*747*
2. Del 4 de julio al 13 de julio (10 días)	*A1*	*138*	*237*	*33*	*266*
	A2	*108*	*133*	*16*	*192*
	A3	*72*	*113*	*15*	*131*
	A4	*120*	*310*	*43*	*293*
3. Del 17 de julio al 25 de julio (9 días)	*A1*	*136*	*73*	*60*	*97*
	A2	*86*	*91*	*36*	*46*
	A3	*102*	*34*	*23*	*51*
	A4	*202*	*138*	*65*	*95*
4. Del 27 de julio al 5 de agosto (11 días)	*A1*	*27*	*127*	*23*	*134*
	A2	*21*	*81*	*19*	*44*
	A3	*42*	*76*	*27*	*60*
	A4	*31*	*213*	*29*	*114*
5. Del 9 de agosto al 14 de agosto (16 días)	*A1*	*54*	*137*	*36*	*122*
	A2	*25*	*53*	*43*	*65*
	A3	*31*	*72*	*28*	*54*
	A4	*30*	*189*	*37*	*167*
6. Del 27 de agosto al 7 de sept. (12 días)	*A1*	*38*	*97*	*71*	*377*
	A2	*30*	*39*	*78*	*126*
	A3	*29*	*41*	*102*	*210*
	A4	*18*	*66*	*33*	*287*

Análisis de los resultados

Análisis gráfico

Siempre conviene empezar analizando los datos gráficamente. Las técnicas gráficas son sencillas, intuitivas y también muy útiles para identificar posibles valores anómalos y para sacar unas primeras conclusiones. Muchas veces estas conclusiones son las que después se ven confirmadas aplicando técnicas más sofisticadas.

En este caso, para tener una primera visión del comportamiento de los datos se utilizó un gráfico como el de la Figura 7, que permite comparar la evolución del número medio de capturas diarias a lo largo de los 6 ciclos para cada uno de los atrayentes.

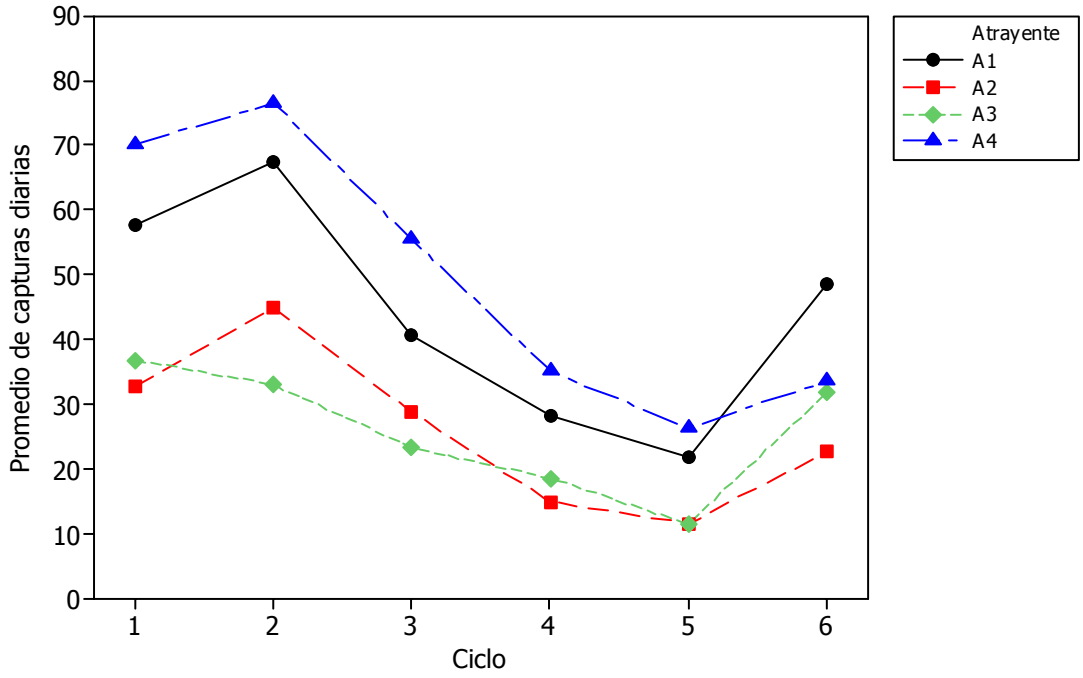


Figura 7: *Evolución del número medio de capturas diarias por ciclo y atrayente*

Se puede ver que el número medio de capturas diarias tiene un ligero crecimiento parecido para todos los atrayentes entre el primero y el segundo ciclo. A continuación va disminuyendo hasta que entre el quinto y el sexto ciclo se observa un remonte que no llega a colocar el número de capturas al mismo nivel que en los primeros ciclos. El comportamiento decreciente de todos los atrayentes en la mayoría de ciclos era de esperar, ya que los atrayentes colocados dentro de las trampas pierden eficacia conforme van pasando los días. De todas formas, el efecto del atrayente está confundido con la población de moscas (no podemos saber si el número de moscas

capturadas disminuye porque el atrayente pierde eficacia, o porque hay menos moscas). Los aumentos entre el primer y segundo ciclo y entre el quinto y el sexto pueden ser debidos a un incremento del número de moscas durante este periodo.

El hecho de que la evolución de los cuatro atrayentes sea prácticamente idéntica indica que la persistencia podría ser parecida para todos ellos. No existe, a simple vista, interacción entre atrayente y ciclo; es decir, la diferencia entre los efectos de los atrayentes sobre el número de capturas diarias parece que no depende del ciclo.

Se puede observar también que el atrayente que sistemáticamente captura más moscas es el A4 seguido del A1, A3 y A2. Estos dos últimos se comportan de manera muy similar. En el último ciclo se puede ver que el atrayente A4, que había sido el mejor, reduce el número de capturas.

En la segunda parte del análisis gráfico se quiso determinar si la evolución del número de capturas diarias era diferente en cada una de las fincas estudiadas. A la vista de los gráficos de la Figura 8 es razonable pensar que el tipo de finca influye en las capturas. Existen dos comportamientos diferenciados: el primero corresponde a las fincas Barranc y Vergel y el segundo a las Fornòs y Balada. En el primer caso se puede observar que el número medio de capturas diarias tiende a crecer entre los ciclos 1 y 3 (en algunos tipos de atrayente hay una breve bajada entre el ciclo 1 y el 2) y a partir del tercer ciclo se produce una reducción bastante drástica de las capturas hasta dejarla casi constante en los últimos ciclos. Por otra parte, encontramos que durante los primeros ciclos en las fincas Fornòs y Balada hay un comportamiento similar al de las fincas Barranc y Vergel, pero entre los ciclos quinto y sexto se observa un gran incremento del número medio de capturas diarias. La finca con mayor número medio de capturas es Balada y a continuación tenemos, por orden decreciente, Vergel, Barranc y Fornòs.

Hay que remarcar que la escala del número de capturas es diferente en los cuatro gráficos. Comparar gráficos con escalas diferentes puede provocar errores de interpretación (puede parecer que en las fincas haya un número similar de capturas cuando en realidad hay muchas más en la finca Balada que en la Fornòs) pero en este caso se ha hecho así porque si se mantienen las escalas, en las fincas donde ha habido menos capturas prácticamente no se apreciarían las diferencias entre atrayentes. Una consecuencia de estas diferencias entre fincas es que el comportamiento general de las cuatro fincas juntas estará muy influenciado por las que tienen un mayor nivel de capturas.

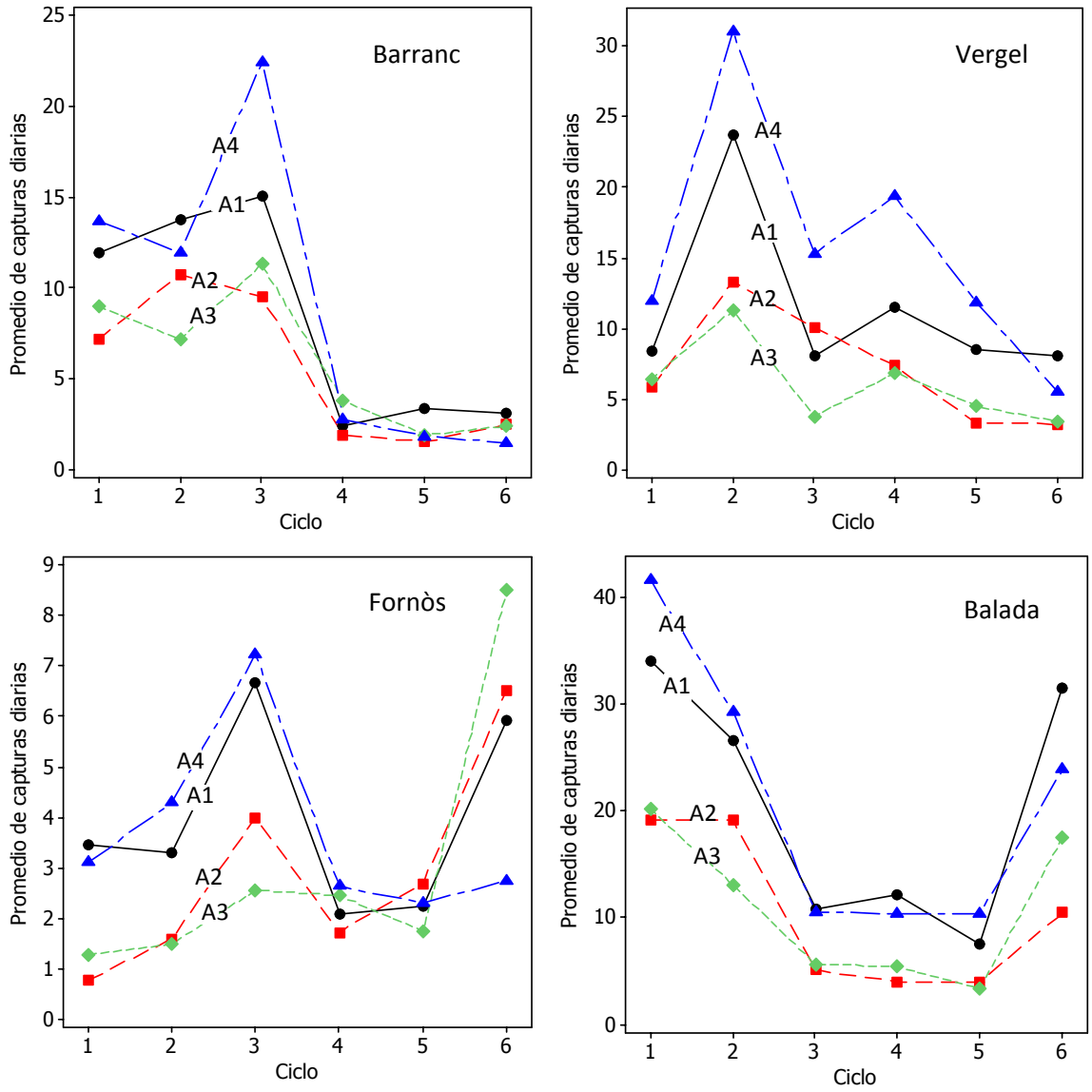


Figura 8: *Evolución del número medio de capturas diarias por ciclo y atrayente en cada finca*

Con respecto a la eficacia de los atrayentes, de este análisis gráfico se puede concluir que el que más capturas diarias realiza en todas las fincas es el A4, aunque, como ya se ha comentado anteriormente, sufre un desgaste notable entre el quinto y el sexto ciclo. El segundo atrayente en número de capturas es el A1.

Modelización de los datos. Pruebas de hipótesis

El análisis cuantitativo de los datos se realiza a través del ajuste de modelos estadísticos intentando averiguar si existen diferencias significativas entre atrayentes (y si existen, cuál es el mejor) eliminando el efecto de la ubicación de la trampa (finca). Cada modelo se ha ajustado seis veces, una vez por cada ciclo. A partir de los modelos

ajustados y validados se puede responder a las preguntas sobre qué atrayentes son más eficaces y sobre si tienen o no el mismo comportamiento en las diferentes fincas. Para responder a estas preguntas se utilizan las pruebas de hipótesis.

El objetivo fundamental de este proyecto era comparar diferentes técnicas y modelos para el análisis de los datos recogidos, valorando la adecuación de cada modelo a los datos que se tienen, y comparando los resultados obtenidos. Entre otros, se utilizaron:

- **Análisis de la varianza con un factor para la respuesta "porcentaje de capturas"**: Compara los cuatro atrayentes con los datos agrupados, sin distinguir la existencia de diversas fincas.

- **Análisis de la varianza con un factor para el arcoseno del porcentaje**: Igual que el anterior pero haciendo un cambio de variable en la respuesta para adaptar mejor su comportamiento a las hipótesis del modelo (requisitos que deben cumplir los datos para el método de análisis sea válido).

- **Análisis de la varianza con un factor para el logit del porcentaje**: Igual en el primero pero ahora con una transformación de la respuesta de tipo logarítmico.

- **Análisis de la varianza con dos factores y efectos fijos para el logaritmo del número de capturas**: En este modelo, además de la influencia del atrayente, también se considera la influencia de la finca. Que el modelo sea de efectos fijos quiere decir que consideramos que en la población sólo tenemos los 4 tipos de atrayentes que consideramos y las cuatro fincas utilizadas.

- **Análisis de la variancia con dos factores y efectos mixtos para el logaritmo del número de capturas**: Similar en el anterior pero en este caso se considera que las fincas son una muestra del conjunto de fincas que pueden existir en la zona estudiada. Por lo tanto, es más realista.

- **Modelo log-lineal para el número de capturas**: Cuando se aplican técnicas de análisis de la variancia se supone que los datos obtenidos provienen de una distribución Normal. Aunque en nuestro caso se podría considerar una aproximación razonable, en rigor la respuesta son datos que provienen de un conteo y la distribución que mejor se adapta a este tipo de datos es la Poisson. La aplicación de este modelo implica análisis más complicados que todavía no están implementados en los paquetes de software estadístico más habituales.

Todos los modelos que tienen en cuenta la existencia de diferentes fincas ponen de manifiesto que en todos los ciclos se comportan de forma diferente con una probabilidad de error muy pequeña.

Las diferencias entre atrayentes son muy claras en los 3 primeros ciclos. En el cuarto y quinto ya no son tan claras, y en el sexto ciclo no hay diferencias entre atrayentes.

Conclusiones

Con respecto al comportamiento de las fincas, se obtiene que su efecto es significativo en los seis ciclos. Esto quiere decir que el nivel de capturas es diferente según la finca estudiada a lo largo de los seis ciclos.

Con respecto a los atrayentes, se concluye que en los primeros cinco ciclos el que captura más moscas es el A4, seguido por el A1. En el último ciclo, el sexto, no existen diferencias significativas entre los cuatro atrayentes y todos capturan moscas por igual, aunque se observa que en este ciclo el atrayente A4 pierde eficacia (persistencia) respecto de los ciclos anteriores.

A partir de estos resultados habría que determinar cuál de los dos atrayentes es más rentable económicamente, dado que el atrayente A1 consigue menos capturas pero su efectividad dura más tiempo, mientras que el A4 siempre empieza realizando más capturas pero su periodo de efectividad es menor.

Lo que sí se puede asegurar es que los agricultores pueden escoger sistemas más ecológicos para producir fruta de calidad, evitando, en la medida del posible, los ataques de la mosca de la fruta.

Algunas consideraciones

A partir de la experiencia obtenida en este estudio, se ha reflexionado sobre los puntos fuertes y también sobre los aspectos mejorables que se deberían tener en cuenta en futuros experimentos.

Como aspecto positivo destaca la colocación de las trampas en diferentes fincas con el fin de estudiar el comportamiento de un factor que afecta a la respuesta (la finca) y que a priori es conocido. También es un acierto separar adecuadamente las trampas con el fin de evitar la interferencia entre atrayentes y, sobre todo, el hecho de

permutar las trampas evitando que se pueda confundir el número de capturas según el atrayente con el número de capturas según la posición dentro de la finca.

Con respecto a los aspectos mejorables se puede mencionar que tal como se ha diseñado el experimento es imposible conocer la persistencia de un atrayente ya que el efecto de la intensidad del atrayente a lo largo del tiempo está confundido con la densidad de moscas en el entorno. O sea, no se puede discernir si una mayor o menor captura a lo largo del tiempo se debe a la persistencia del atrayente o al número de moscas presentes en los cultivos. Una posible solución podría haber sido medir la persistencia relativa de los 3 atrayentes experimentales respecto al que ya se está comercializando.

Otro aspecto a mejorar en el estudio es el hecho de tomar diferentes longitudes para los ciclos recogidos. Con el fin de tener datos más comparables entre ciclos habría que ser estrictos al contar los días entre revisiones y tener ciclos de igual tamaño. Por último, una mejora a tener en cuenta para futuras recogidas de datos sería colocar una trampa sin ningún atrayente para comparar también el efecto de tener atrayente en la trampa respecto a no tener.

(Proyecto de la Licenciatura de Estadística, presentado en junio de 2002 con el título "Comparativa de mètodes per l'anàlisi dels assajos de captures de 'Ceratitis Capitata'")

Este proyecto muestra perfectamente el esquema de un estudio estadístico. Primero se plantea una pregunta (¿qué atrayente es mejor?), para responderla se necesitan datos (no podemos saber cuál será mejor sólo analizando su composición química, por ejemplo), conseguir los datos no es fácil, y se tiene que hacer de una forma muy cuidadosa para que reflejen correctamente aquello que se quiere medir (hay que realizar correctamente lo que llamamos "diseño del experimento"). Finalmente hay que analizar los datos obtenidos y sacar conclusiones.

En el proyecto original se utilizan diferentes técnicas de análisis para comparar las conclusiones que se extraen con cada una de ellas. Muchas veces el análisis exploratorio de los datos ya da pistas muy claras sobre cuáles serán las conclusiones finales.

Lourdes Rodero estudió la Diplomatura y la Licenciatura de Estadística en la UPC (dudaba entre matemáticas y estadística, y no se decidió hasta el último momento). Empezó su actividad profesional en una mutua de salud desde donde se incorporó a la UPC como a profesora de estadística. Imparte sus clases en la Escuela de Ingeniería Industrial de Barcelona y en la FME, y está terminando su tesis doctoral. Es la única persona que aparece en este libro como autora de un proyecto y directora de otro.

Estudios de mercado

6

El diseño emocional: cómo crear productos que sean atractivos (y un ejemplo sobre zumos de frutas)

Proyecto realizado por: **Ana Gómez Muñoz**
Elisabeth Peralta Coll
Dirigido por**: Lluís Marco Almagro**

Diariamente convivimos con infinidad de productos de muchas marcas diferentes. Pero aunque sean de marcas diferentes, todos se parecen mucho. Pensemos por ejemplo en los teléfonos móviles. La funcionalidad (las cosas que pueden hacer) es similar: se puede hablar, enviar y recibir mensajes, muchos tienen cámara, pueden reproducir música... Tampoco hay diferencias grandes en el precio. ¿Por qué entonces algunos modelos son mucho más populares que otros? Al menos una parte de la respuesta es que, simplemente, nos gustan más (aunque quizá nos resulta difícil explicar por qué nos gustan más).

El diseño emocional (o ingeniería kansei) intenta ayudar a diseñar productos (y servicios) que sean atractivos y que transmitan determinadas sensaciones a sus usuarios. Este aspecto del diseño ha tomado mucha importancia y la estadística juega un papel muy destacado.

Como nunca habíamos hecho un estudio de ingeniería kansei, pensamos en un producto sencillo y accesible (los zumos de frutas) para probarlo. Los pasos que se siguieron (y que son los típicos en estos estudios) son los que se describen en este proyecto.

Qué es la ingeniería kansei y por qué es importante

'Kansei' es una palabra japonesa difícil de traducir (como pasa con tantas palabras japonesas; ya se sabe, culturas diferentes, alfabetos diferentes...). Pero básicamente se refiere a la impresión subjetiva que cada uno de nosotros se lleva ante una determinada situación. Por ejemplo, visualicémonos caminando por una playa tropical, desértica, de arena blanca, con aguas cristalinas, palmeras... Oímos el sonido de unas olas suaves, notamos una brisa ligera... La sensación (el kansei) que posiblemente muchos de nosotros tendríamos es de tranquilidad, de paz, de gozo de vivir.

Los productos también nos provocan sensaciones. Un reproductor de música nos parece moderno, un coche potente o una lavadora cómoda de utilizar (incluso antes de haberla utilizado). La ingeniería kansei intenta relacionar las características técnicas de los productos con las sensaciones que provocan. A esta disciplina también se le denomina diseño emocional, ingeniería emocional, ingeniería afectiva y algún otro nombre. Desgraciadamente, tener tantos nombres para lo mismo es una dificultad para difundir su uso.

¿Y por qué es importante diseñar productos atractivos? En un mundo en el que cada vez hay menos diferencia en la funcionalidad y el precio de los productos, lo que puede marcar la diferencia entre los que se venden y los que no es la impresión que producen en los compradores. Pero no sólo se trata de una estrategia para vender más: ¡los productos atractivos también funcionan mejor! Existen estudios que muestran que se cometen menos errores y los usuarios van más rápido utilizando productos que valoran como bonitos, como atractivos.

Así pues, si diseñar productos atractivos es tan importante, ¿como podemos hacerlo? La ingeniería kansei pretende ayudar en esta tarea. Muchas empresas han utilizado ya esta técnica para diseñar coches, electrodomésticos, envases de alimentos... El producto con que se aplicó esta metodología fueron zumos de fruta presentados en vasos. Era un producto fácil de conseguir y familiar para mucha gente. Simon Schütte, un profesor de la Universidad de Linköping (Suecia), impartió un curso sobre ingeniería kansei en nuestra universidad, y propuso un modelo muy claro para hacer estos estudios, que es el que se utilizó.

El modelo para realizar estudios de ingeniería kansei

La primera fase de la metodología consiste en describir el producto a desarrollar, el grupo al cual se dirigirá y la situación actual del mercado.

La segunda etapa consiste en la definición del llamado espacio semántico. Se trata de elaborar una lista de palabras (llamadas palabras kansei) que intentan describir todas las sensaciones que puede provocar el producto en los usuarios. También se define el espacio de propiedades; es decir, las características técnicas del producto, a las cuales se asignan unos valores que el diseñador puede variar.

Posteriormente, hay que construir prototipos del producto y recoger datos sobre las sensaciones que provoca preguntando a un grupo de personas. En la etapa de síntesis se relacionan las palabras kansei con las propiedades (por ejemplo, si se está haciendo un estudio sobre relojes se podría llegar a la conclusión que un reloj con esfera redonda se percibe como más moderno que uno con esfera cuadrada).

Finalmente se validan las conclusiones obtenidas. Todo este proceso, que quizá ahora no dice demasiado, quedará mucho más claro –esperamos– con el ejemplo de los zumos.

La definición del espacio semántico en el caso de los zumos

El espacio semántico se describe mediante las llamadas palabras kansei. Las palabras kansei son palabras (a menudo adjetivos) que describen las sensaciones asociadas al producto, en este caso los zumos de fruta. Para hacer la recopilación de palabras hay que apoyarse en material que tenga relación con el producto a desarrollar. Nos podemos ayudar de revistas, manuales, literatura especializada, páginas web, o de otros posibles estudios kansei realizados previamente. Además de utilizar material escrito podemos hablar con personas expertas en la materia o consumidores que hayan utilizado el producto y estén dispuestos a compartir su experiencia.

En nuestro caso, para la busca de las palabras kansei se hizo uso de páginas web y de libros especializados en zumos naturales, batidos e infusiones. A partir de estas fuentes se seleccionaron un total de 125 adjetivos referentes a sensaciones o emociones capaces de representar los efectos que pueden provocar los zumos de fruta al ser observados. Estos 125 adjetivos son las palabras situadas en un cuadro de color amarillo en la Figura 2.

Es verdad, ¡125 palabras son muchas! (a pesar de que en algunos estudios kansei este primer paso supone listas de más de 300 o 400 palabras). Como después un grupo de personas tendrá que realizar una valoración de todas las palabras kansei para cada producto que se le presente, no se puede plantear la recogida de datos con tantas palabras, porque sería muy pesado, y los participantes perderían interés y se cansarían

(probablemente, la mayoría ya no aceptarían empezar con el estudio si se les dijera que tendrán que dedicar horas y horas valorando productos).

Por esta razón, es necesario reducir el espacio semántico resumiendo todas las palabras en unas pocas agrupadas según su significado. Eso se acostumbra a realizar con un diagrama de afinidad. Cada una de las palabras kansei se escribe en un post-it, y a continuación se van formando grupos de palabras con significado parecido. Para cada grupo se selecciona una de ellas, la que mejor lo representa, o se crea una nueva que tenga el significado global de todas ellas.

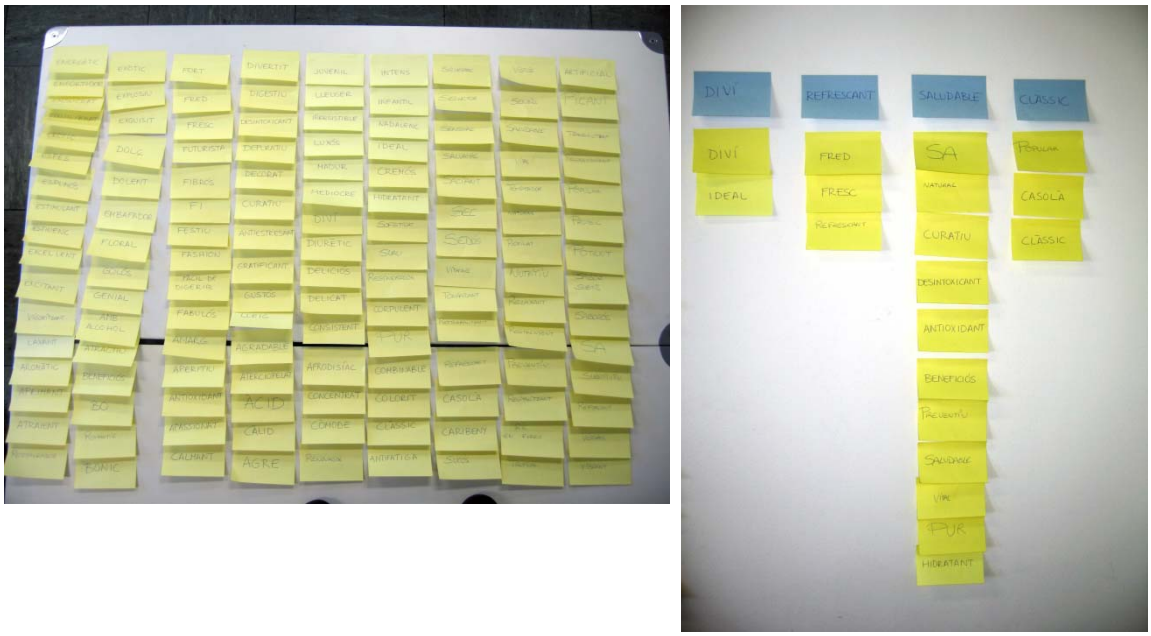


Figura 1: *Post-its con todas las palabras (izquierda). Ejemplo de agrupación de las palabras*

Como resultado de esta agrupación se consiguió resumir las 125 palabras kansei en 33. En un segundo paso, y con otro diagrama de afinidad, se resumieron las 33 palabras en solo 14 (Tabla 1).

Tabla 1: *Lista final de palabras kansei.*

Festivo	Artificial	Energético	Saludable
Ligero	Clásico	Refrescante	Gustoso
Aromático	Seductor	Exótico	Relajante
	Fuerte	Juvenil	

El diagrama de afinidad final se puede ver en la Figura 2. Las palabras de las dos primeras columnas corresponden a las 125 palabras iniciales ya agrupadas. Las palabras de la tercera columna son las 33 palabras que resumen las primeras y las 14 de la última columna son las palabras kansei finales.

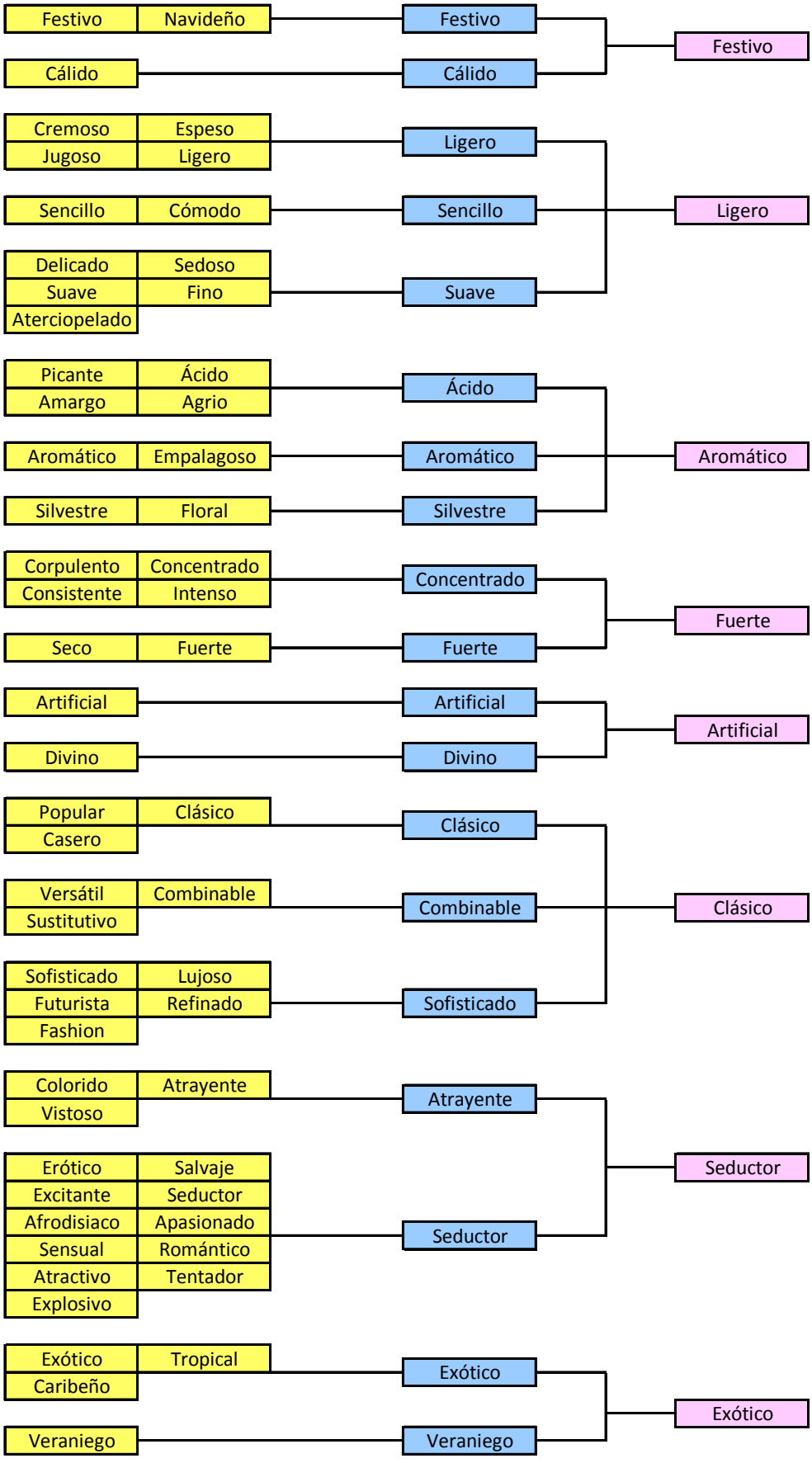


Figura 2, primera parte

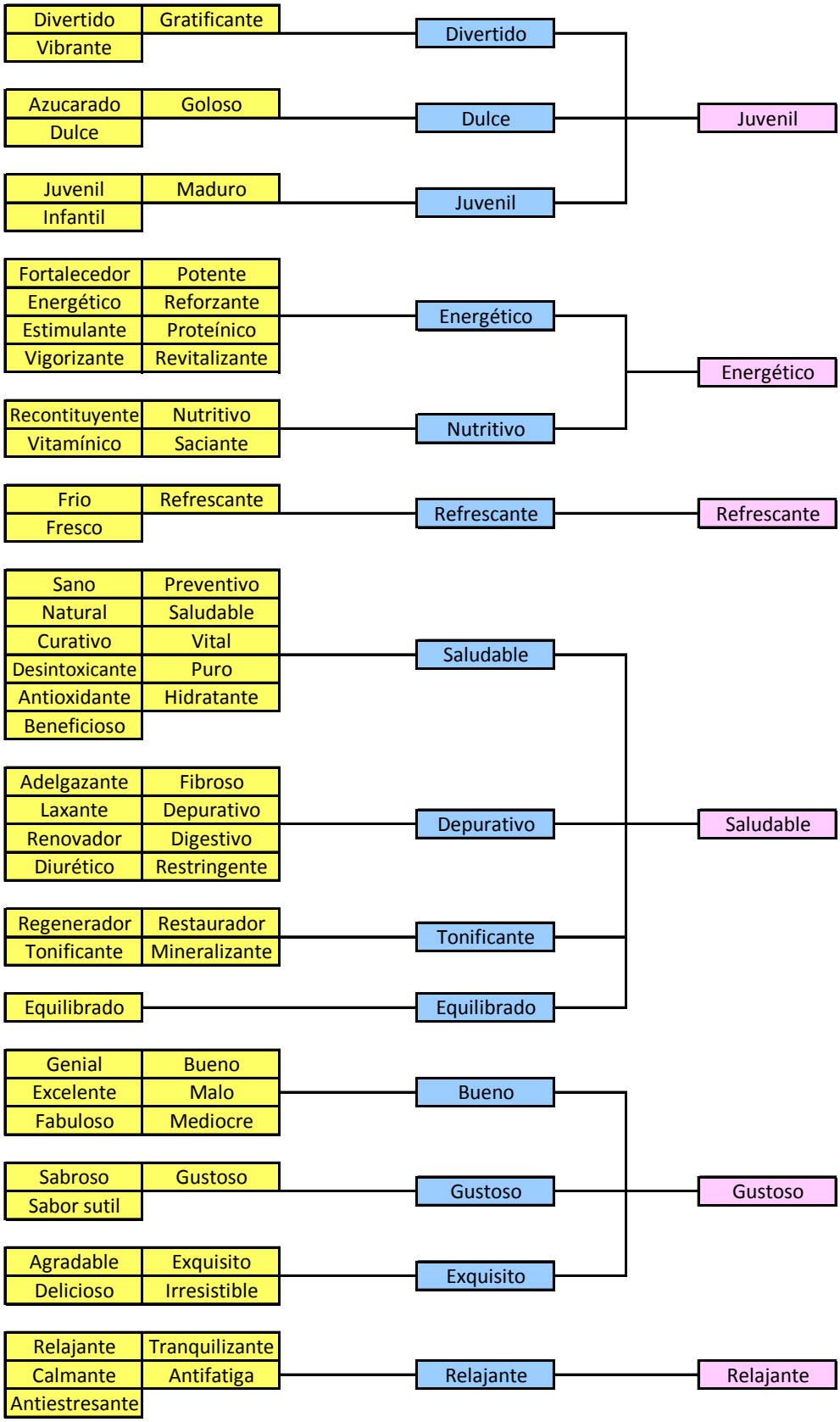


Figura 2: *Diagrama de afinidad de las palabras Kansei*

El diagrama de afinidad para reducir el espacio semántico se puede complementar con un análisis clúster. El análisis clúster es una técnica estadística que permite formar grupos de palabras similares. Es decir, cuando unos zumos siempre reciben puntuaciones altas para un conjunto de palabras, y otros reciben siempre puntuaciones bajas para esas mismas palabras, significa que este conjunto de palabras se percibe como lo mismo y por tanto se puede resumir en una sola.

El procedimiento seguido en este caso fue agrupar de nuevo las 33 palabras kansei obtenidas en el primer diagrama de afinidad y verificar así si la segunda agrupación era realmente la más adecuada. Este análisis se realizó a partir de las valoraciones de 6 personas sobre cuatro zumos completamente distintos (Figura 3), de diferente color, con diferentes decoraciones, unos con hielo y otros sin, en recipientes diferentes, etc. Las imágenes se extrajeron de uno de los libros usados para buscar palabras kansei.

Figura 3: *Zumos utilizados en la encuesta pera el análisis clúster*

Las seis personas que rellenaron los formularios valoraron las 33 palabras (ordenadas aleatoriamente) para cada uno de los cuatro zumos. Por ejemplo, ante la palabra "exótico", cada persona tenía que poner una nota en cada uno de los cuatro zumos, del 1 al 7 (un 1 quiere decir que el zumo es la cosa menos exótica del mundo, mientras que un 7 denota el zumo más exótico que uno se puede encontrar).

Algunas palabras (concretamente 3) cambiaron de grupo. Un grupo formado por dos palabras desapareció, ya que se mostró que estas sensaciones eran similares a las de otras palabras.

Aunque finalmente quedaron 13 grupos de palabras, consideramos que algunas de las sensaciones no eran de interés, y que todavía eran demasiadas palabras. Por eso, decidimos quedarnos únicamente con 7 palabras kansei. Una vez acabado todo el estudio nos dimos cuenta de que esta decisión quizá no había sido demasiado buena, y que la selección de estas 7 palabras no había sido la mejor, pero se nos tiene que perdonar: el estudio lo hacíamos para aprender...

Las 7 palabras finales eran artificial, seductor, refrescante, exótico, saludable, sabroso y relajante. Como todas las palabras excepto artificial tenían un sentido positivo, decidimos cambiar esta palabra por su antónimo: natural.

Las palabras kansei definitivas usadas en el estudio son las que figuran en la Tabla 2.

Tabla 2: *Palabras Kansei definitivas*

Natural	Seductor	Refrescante
Exótico	Saludable	Gustoso
	Relajante	

La definición del espacio de propiedades en el caso de los zumos

El espacio de propiedades determina los factores del producto que se someterán a estudio. Al definir el espacio de propiedades nos podemos encontrar con dos situaciones diferentes: que el estudio se base en prototipos de productos que se tienen que construir (de manera que se dispone de mucha más libertad para escoger los factores), o bien que se utilicen prototipos disponibles ya creados.

Por los zumos se elaboró una lista con factores que se podían controlar y modificar en las imágenes de los zumos, inspirándose en las mismas fuentes usadas para determinar el espacio semántico. Finalmente se decidió trabajar con las siguientes propiedades:

Tabla 3: *Propiedades escogidas y sus alternativas*

PROPIEDADES	ALTERNATIVAS
Color	*Amarillo* *Naranja*
Recipiente	*Vaso* *Copa*
Decoración	*Sí* *No*
Pajilla	*Sí* *No*
Hielo	*Sí* *No*

Si quisiéramos tener todas las combinaciones posibles de estos factores saldrían 32 imágenes diferentes. Pero no es necesario hacerlas todas para descubrir cómo afecta cada uno de estos factores a las palabras kansei seleccionadas. Con sólo 16 combinaciones (las que se muestran en la Tabla 4) es suficiente (esto es lo que corresponde a un diseño factorial fraccional 2^{5-1}). Las fotos de todos los zumos creados se muestran en la Figura 4.

Tabla 4: *Características de cada uno de los zumos*

ZUMO	PAJILLA	DECORACIÓN	HIELO	RECIPIENTE	COLOR
1	*No*	*No*	*No*	*Vaso*	*Naranja*
2	*Sí*	*No*	*No*	*Vaso*	*Amarillo*
3	*No*	*Sí*	*No*	*Vaso*	*Amarillo*
4	*Sí*	*Sí*	*No*	*Vaso*	*Naranja*
5	*No*	*No*	*Sí*	*Vaso*	*Amarillo*
6	*Sí*	*No*	*Sí*	*Vaso*	*Naranja*
7	*No*	*Sí*	*Sí*	*Vaso*	*Naranja*
8	*Sí*	*Sí*	*Sí*	*Vaso*	*Amarillo*
9	*No*	*No*	*No*	*Copa*	*Amarillo*
10	*Sí*	*No*	*No*	*Copa*	*Naranja*
11	*No*	*Sí*	*No*	*Copa*	*Naranja*
12	*Sí*	*Sí*	*No*	*Copa*	*Amarillo*
13	*No*	*No*	*Sí*	*Copa*	*Naranja*
14	*Sí*	*No*	*Sí*	*Copa*	*Amarillo*
15	*No*	*Sí*	*Sí*	*Copa*	*Amarillo*
16	*Sí*	*Sí*	*Sí*	*Copa*	*Naranja*

Figura 4: *Imágenes de los zumos*

Recogida de datos

La recogida de datos tuvo lugar en la Facultad de Matemáticas y Estadística el día 13 de abril de 2007. Para la recogida de datos se contó con 24 personas (familiares y amigos) que participaron voluntariamente (y realmente poniendo mucho interés, como dicen ellos "demostraron que nos aprecian..."). Las 24 personas se asignaron a 4 grupos de forma aleatoria. A cada uno de estos grupos se les sentó en una mesa, con un ordenador portátil que mostraría las imágenes de los zumos.

Inicialmente, se entregó a cada uno de los participantes una lista con las definiciones de cada una de las palabras kansei, dado que era muy importante que todo el mundo entendiera lo mismo al leer las palabras.

Cuando ya todos estaban preparados se empezó la recogida de datos. Las imágenes de los 16 zumos se fueron mostrando, una a una, en cada portátil. Para cada zumo, cada participante puntuaba las 7 palabras kansei en una escala de 1 a 7 (Figura 5). Las imágenes estaban en un orden diferente para cada grupo; así se pudo comprobar después que las valoraciones no se veían afectadas por el orden en qué aparecen las imágenes.

Se repitió una segunda vez todo el proceso de valoraciones para tener datos con los que poder comparar la primera y la segunda valoración de cada persona y poder valorar la consistencia de sus respuestas. Como las valoraciones se hacían el mismo día, se separaron por un aperitivo, de manera que los participantes se distrajeran un rato y no se vieran influenciado por las imágenes anteriores (y también para obsequiarlos con alguna cosa de comida, eso siempre gusta). Para la segunda ronda de valoraciones se volvió a variar el orden en que aparecían las imágenes y, de nuevo, para cada una de las combinaciones de zumos valoraron las siete palabras kansei.

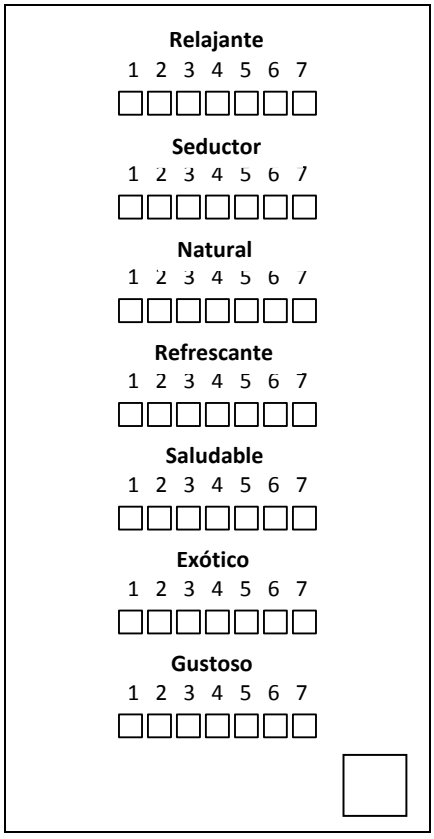


Figura 5: *Formulario utilizado para la evaluación de los zumos*

Figura 6: *Dos momentos de la reunió en que se recogieron los datos*

En definitiva, cada persona rellenó 16 formularios (uno para cada zumo) cada ronda. Como se hicieron dos rondas y se tenían 24 participantes, en total, se cumplimentaron 768 formularios. En la Figura 7 tenemos un ejemplo de las puntuaciones dadas. Corresponde a la primera y segunda ronda de las valoraciones para la palabra 'exótico'. Los colores de cada celda varían en una escala en la cual las puntuaciones más altas tienen color verde y las más bajas color rojo.

	Imma	Lydia	Mònica	David	Erli	Eva	Belén	Antonic	Héctor	Helena	Vicenç	Jose	Xavi	Raquel	Merce	Guillem	Marta	Pere	Núria	Esther	Sandra	Enrique	Sergi	Manolo
1	1	3	3	1	2	3	3	4	1	4	2	4	2	3	4	1	4	2	3	1	2	5	7	7
2	1	1	2	1	1	1	1	1	1	6	3	1	2	1	4	2	4	2	3	2	1	2	2	1
3	1	1	2	5	1	2	6	1	2	6	2	6	2	5	4	2	3	3	5	2	4	5	4	7
4	1	1	4	5	7	5	7	7	5	4	4	5	5	3	5	4	3	4	7	3	6	6	7	7
5	1	1	3	1	1	1	4	2	1	6	3	1	4	1	1	4	4	3	4	3	3	4	2	1
6	1	1	2	1	3	1	6	6	5	4	5	5	5	1	5	2	5	2	1	4	5	5	5	7
7	3	1	5	7	2	6	7	5	6	5	5	6	3	7	6	2	4	5	6	6	7	4	7	7
8	2	1	4	5	2	5	4	4	3	4	5	6	5	7	3	4	4	6	7	7	7	6	7	7
9	1	1	2	1	1	1	2	1	1	6	2	1	3	1	2	2	3	2	5	3	1	4	1	1
10	2	1	5	1	4	4	5	6	3	3	4	5	6	3	6	3	5	3	5	3	2	6	5	7
11	3	1	5	6	7	7	7	6	3	4	6	6	6	7	6	3	3	3	6	6	7	6	7	7
12	2	1	4	5	4	5	2	1	2	5	3	6	5	7	5	5	3	5	6	6	4	6	6	7
13	1	1	3	1	2	4	6	5	6	3	5	5	7	5	3	4	3	3	4	3	5	3	5	6
14	2	1	3	1	1	2	1	1	1	5	4	2	4	1	3	3	3	4	3	3	3	4	4	1
15	1	1	5	7	1	6	7	2	7	4	5	6	3	7	5	7	5	5	7	3	6	6	6	7
16	2	1	5	5	6	7	7	4	5	5	5	6	5	7	7	5	4	4	5	6	7	7	7	7

	Imma	Lydia	Mònica	David	Erli	Eva	Belén	Antonic	Héctor	Helena	Vicenç	Jose	Xavi	Raquel	Merce	Guillem	Marta	Pere	Núria	Esther	Sandra	Enrique	Sergi	Manolo
1	1	1	2	1	1	1	2	5	3	2	3	3	6	1	4	1	4	2	2	4	1	4	7	1
2	1	1	1	1	1	1	1	2	1	4	2	4	3	2	3	3	3	3	2	2	1	3	1	1
3	1	1	3	2	2	2	4	4	1	4	5	6	2	4	5	5	4	3	4	2	5	5	3	4
4	1	1	3	5	5	4	7	5	4	4	6	6	6	4	7	2	4	3	3	6	7	3	7	4
5	1	1	2	1	1	1	4	2	1	4	4	2	3	1	2	2	3	3	6	2	2	5	4	5
6	1	1	2	3	5	3	3	5	3	3	5	2	4	2	4	1	3	2	3	2	4	3	7	1
7	1	1	2	3	6	4	5	5	4	3	5	6	6	5	7	5	4	3	7	5	6	6	7	7
8	1	1	3	5	6	5	2	5	2	5	6	7	3	5	5	7	4	5	5	7	7	6	5	6
9	1	1	1	1	1	1	1	1	1	2	3	2	1	1	4	6	4	3	1	2	1	2	2	1
10	1	1	2	1	1	2	4	4	5	5	4	3	6	3	5	2	3	3	2	4	3	3	6	1
11	1	1	3	3	2	5	6	5	4	4	5	6	5	7	6	4	3	3	5	5	6	6	7	4
12	1	1	2	3	5	4	1	6	1	5	6	5	3	7	4	7	4	4	2	4	6	6	5	7
13	1	1	2	1	1	3	4	5	4	4	5	2	4	4	5	5	3	2	2	3	5	5	7	4
14	1	1	2	1	1	1	2	3	1	4	4	2	2	3	4	6	4	3	4	2	5	5	4	4
15	1	1	2	2	7	3	1	2	1	4	5	5	3	7	4	5	4	5	6	7	6	6	7	7
16	1	1	4	3	7	6	7	5	5	4	7	7	7	7	6	5	4	4	6	7	7	7	7	7

Figura 7: *Puntuaciones para la palabra "exótico"*

Mirando las puntuaciones solo para la palabra 'exótico' ya se ponen de manifiesto algunos hechos: hay personas que tienen tendencia a dar puntuaciones bajas y otras puntuaciones altas; y también hay personas con muy poca variabilidad entre zumos (Imma y Lydia, por ejemplo, prácticamente siempre puntúan con un 1 a todos los zumos), y personas con mucha más variabilidad (que tienen un rango de puntuaciones que va de 1 a 7). Volveremos a estos hechos en el apartado de conclusiones.

Síntesis

La síntesis es la etapa en que se relaciona el espacio semántico con el espacio de propiedades. Se trata de encontrar qué características de los zumos son importantes para cada una de las palabras kansei.

Hay muchas técnicas que se pueden utilizar en la síntesis. Una de las más populares en la literatura japonesa es la llamada QT1 (*Quantification Theory Type I*), que básicamente es una regresión lineal. Como la matriz de datos de los zumos corresponde a un diseño factorial fraccional, se pueden utilizar las técnicas habituales para hallar los efectos en un diseño factorial y ver qué propiedades son significativas en nuestro caso. Ninguna propiedad ha aparecido como significativa para las palabras 'saludable', 'natural' y 'relajante'. En cambio, se ve muy claro que los zumos que tienen hielo se valoran como más refrescantes. También descubrimos que los zumos de color naranja y con decoración se valoran como más exóticos y seductores. Los zumos de color naranja también se valoran como más sabrosos.

En la Figura 8 podemos ver los zumos representados en unos ejes donde las abscisas son las medias de las puntuaciones de los 24 participantes, y las ordenadas corresponden a la desviación tipo para la palabra refrescante. Vemos claramente la distinción entre los zumos que no llevan hielo (menos refrescantes, en la parte izquierda), y los que sí llevan hielo (más refrescantes, en la parte derecha).

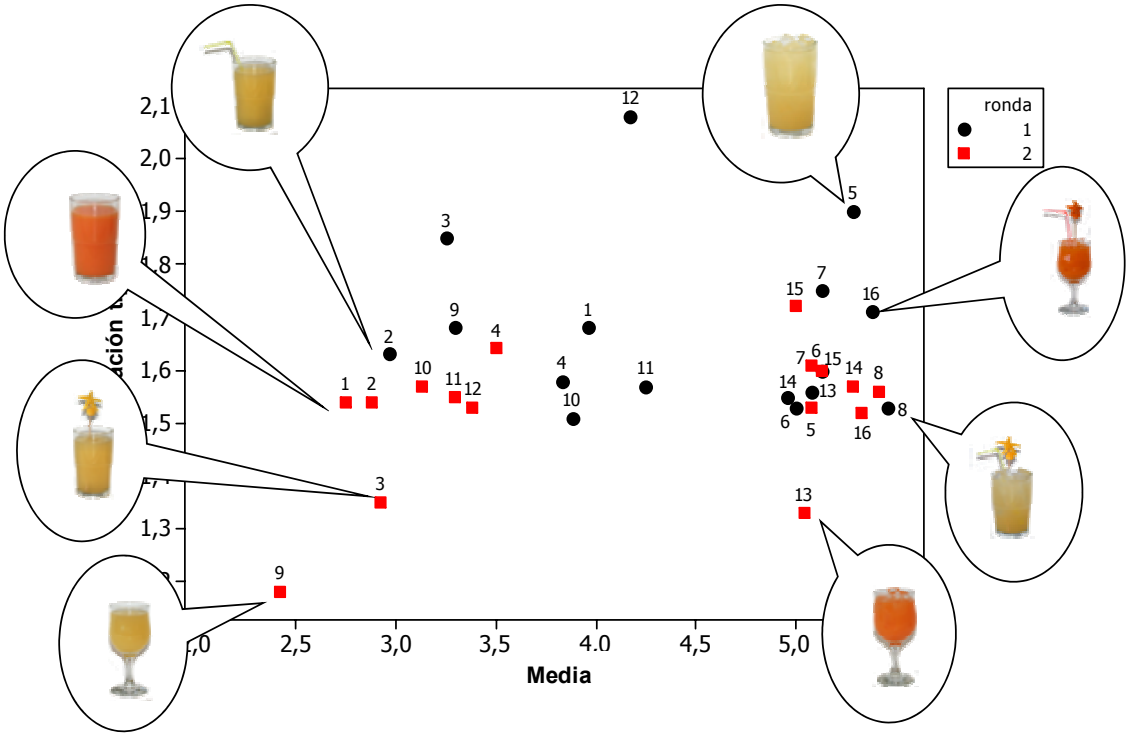


Figura 8: *Desviación tipo versus media de las puntuaciones para "refrescante"*

La misma gráfica se puede hacer para otras palabras. En la Figura 9 vemos como, para la palabra 'seductor', los zumos amarillos y sin decoración aparecen a la izquierda (poco seductores) y los naranjas y con decoración a la derecha (muy seductores). También observamos aquí cómo todo el mundo está muy de acuerdo en los zumos que son poco seductores (desviación tipo baja), pero hay mucha más diversidad de opiniones sobre los zumos que son muy seductores (desviación tipo alta).

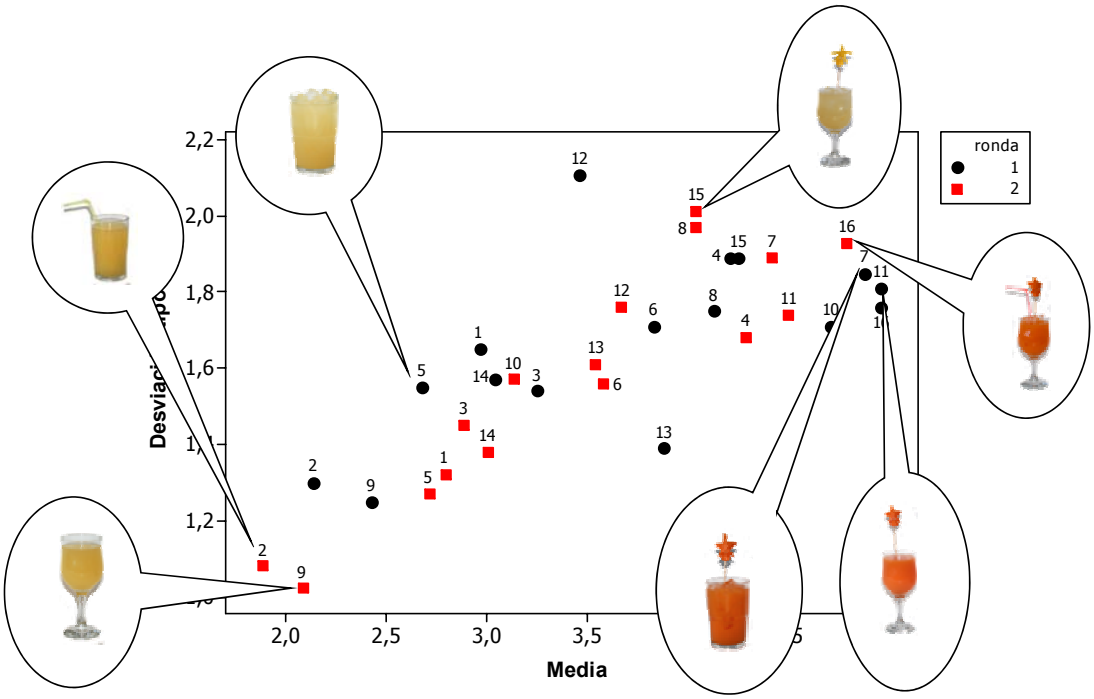


Figura 9: *Desviación tipo versus media de les puntuaciones para "seductor"*

Estudio de la similitud entre palabras kansei

Una forma visual y fácil de analizar los datos es mediante gráficos radar. Estas representaciones muestran, para cada una de las palabras kansei, la media de las valoraciones de cada zumo. Los números del 1 al 16 corresponden a cada uno de los zumos según la relación de la Tabla 4. Si estos gráficos presentan una forma redondeada (como pasa con las palabras 'relajante', 'saludable' y 'natural'), quiere decir que todos los zumos han recibido puntuaciones similares para esa palabra y que, por lo tanto, no hay ninguna propiedad que tenga un efecto significativo.

Cuando para una palabra vemos "entradas y salidas" en el gráfico radar, quiere decir que hay diferencias entre los zumos, y por lo tanto encontraremos propiedades con un efecto significativo sobre la palabra. Además, cuando dos palabras tienen un perfil en el gráfico radar parecido (cómo pasa con las palabras exótico y seductor), significa que las dos sensaciones se perciben como similares.

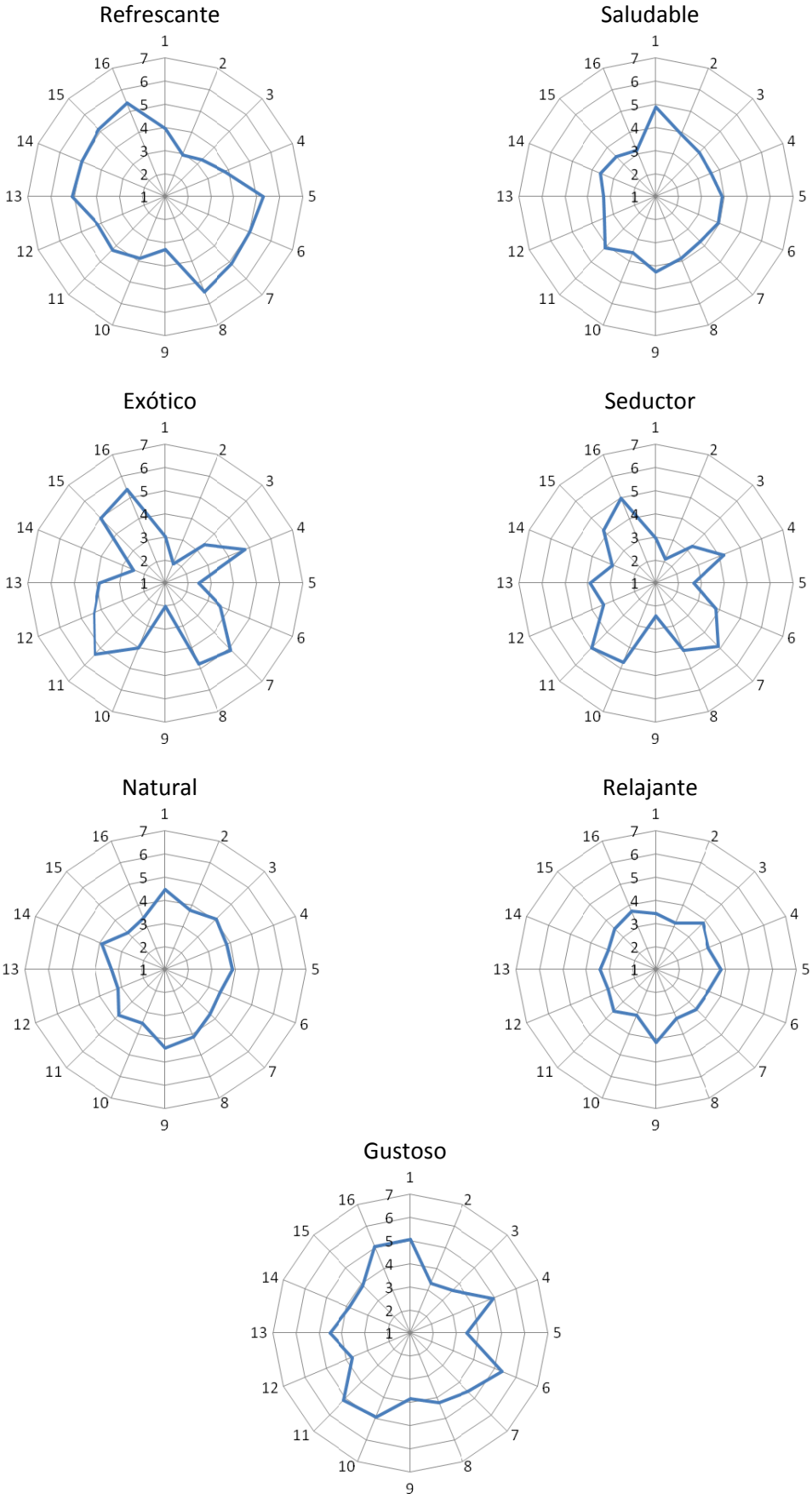


Figura 10: *Gráficos radar para cada una de las palabras kansei*

Hemos visto que las propiedades significativas para ‘exótico’ y ‘seductor’ eran las mismas, el perfil parecido en los gráficos radar confirma este hecho. Los gráficos radar para natural y saludable también tienen ciertas similitudes en su forma, aunque no tan marcadas como las que tienen seductor y exótico.

Se puede hacer también un análisis de componentes principales para ver qué palabras se perciben como semejantes. Esta es una técnica descriptiva en la cual se crean unas nuevas variables (llamadas componentes principales) que son combinaciones lineales de las variables originales. La particularidad es que la primera componente principal captura una gran parte de la variabilidad de los datos originales, la segunda un poco menos, etc. Con un poco de suerte, sólo con las dos o tres primeras componentes principales podemos capturar en torno al 70 u 80% de la variabilidad original (esto, claro está, depende en cada caso de los datos).

Pero la gracia de todo esto es que se pueden representar las palabras kansei en un diagrama bivariante de las dos primeras componentes (Figura 11). Fijémonos cómo los zumos que se perciben como seductores también se perciben como exóticos (y por eso las dos palabras están muy próximas). Lo mismo pasa con las palabras ‘saludable’ y ‘natural’. ¿Es bastante razonable, verdad? Es por eso que, al escoger las palabras para el espacio semántico, quizás habríamos podido escoger sólo una de cada pareja (por ejemplo, sólo preguntar por seductor, y no por exótico), y en cambio haber incluido algunas palabras que se descartaron.

‘Relajante’, en cambio, es bastante perpendicular a las otras, por lo tanto, no se parece a ninguna otra palabra kansei.

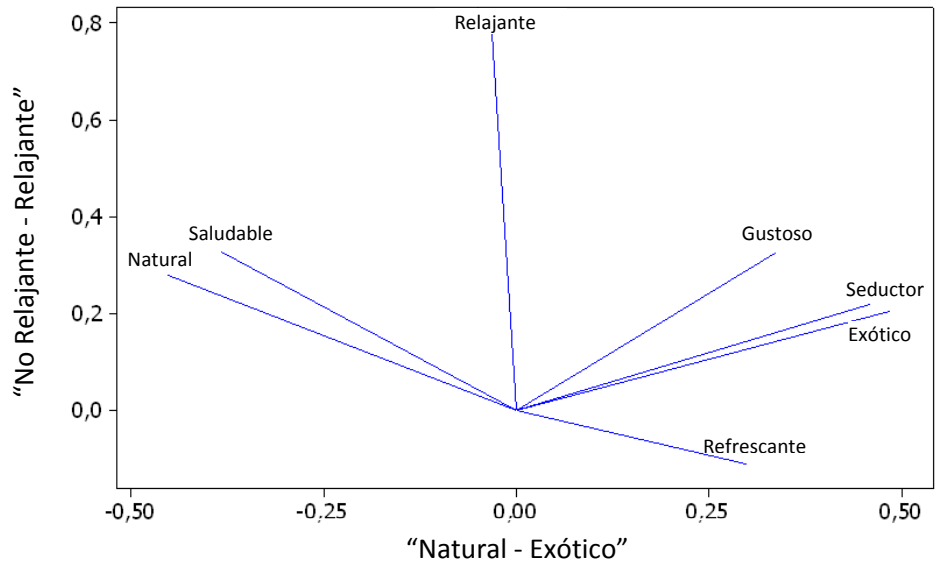


Figura 11: *Similitud de las palabras kansei*

Localización de los zumos en el espacio semántico

Se pueden utilizar los mismos ejes que antes para caracterizar cada una de las 16 combinaciones estudiadas. En la parte superior izquierda están los considerados como naturales. En la parte derecha se observan los zumos seductores y exóticos (básicamente, zumos con decoración).

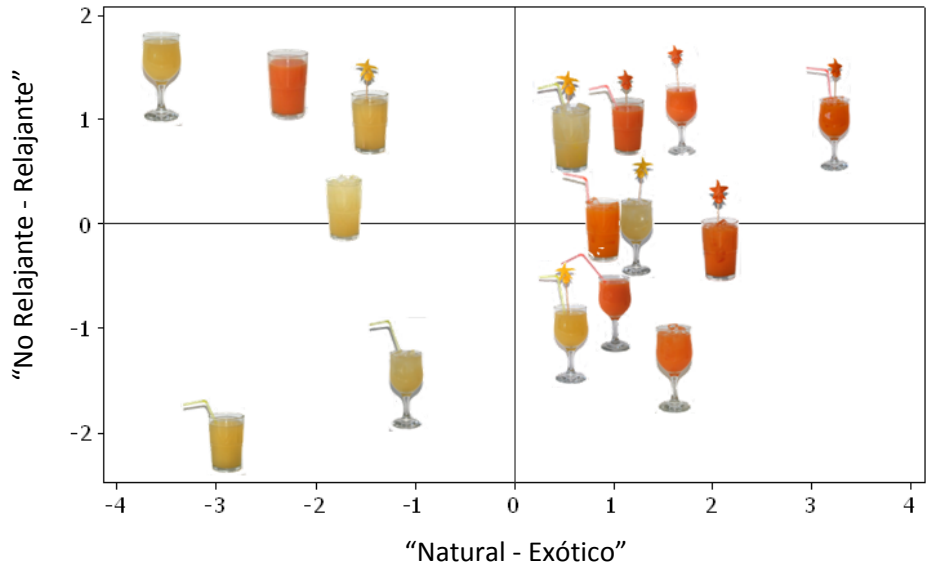


Figura 12: *Localización de los zumos en el espacio semántico*

Conclusiones

El proyecto realizado sirvió, sobre todo, para entender cómo hacer un estudio de ingeniería kansei. Como pasa muy a menudo, cuando más se aprende es cuando tienes que hacer un estudio de principio a fin. Recoger los datos, por ejemplo, tiene una dificultad en cualquier estudio estadístico y a menudo no se ve reflejado en los informes que después se elaboran analizando los resultados: pensar en cómo se organiza, destinar horas y horas a elaborar los formularios, no equivocarse en el momento de la recogida, pasar después todos los datos al ordenador ...

Los datos recogidos todavía pueden dar más de sí. Recordemos que hicimos dos rondas, es decir, para cada zumo y cada palabra teníamos las puntuaciones que cada persona dio en dos ocasiones. Comparando las dos rondas se ve cómo algunas personas son muy consistentes en las puntuaciones que dan, pero otros no lo son nada. Quizás sería interesante fiarse más de la gente más coherente... Hay que pensar.

Y tal como se puede ver en la Figura 8, la variabilidad entre las personas es muy grande (cosa bastante previsible, por otra parte). Trabajar siempre con la media de las puntuaciones de todas las personas es muy empobrecedor, también habría que pensar maneras de tener en cuenta esta variabilidad.

Bien, en definitiva, cada vez más los productos que se diseñan y que después utilizamos son productos "emocionales". Los estudios de ingeniería kansei, donde se utilizan multitud de técnicas estadísticas, irán tomando importancia en el futuro próximo. ¡Pero todavía queda tanto para hacer! ¿Y cómo se podría hacer todo esto para un servicio? Por ejemplo, ¿cómo tienen que ser las salas de espera de los hospitales para que transmitan sensación de tranquilidad, de profesionalidad...? La ingeniería kansei es, realmente, una puerta de entrada apasionante a un montón de técnicas estadísticas bien diversas.

(Proyecto de la Diplomatura de Estadística presentado en septiembre de 2007 con el título "Aplicació de la metodologia Kansei en la valoració de sucs de fruita")

Tema nuevo (y muy interesante,) pero sistemática de busca de conocimiento ya conocida: Se hacen preguntas (¿qué características de la presentación de un zumo transmiten "emociones" y de qué tipos?), planifican la recogida de los datos (que, naturalmente, tienen que ser relevantes para responder a las preguntas), los recogen (¡con cuidado!), los analizan y sacan conclusiones. Este es el paradigma de un estudio estadístico, perfectamente aplicado en este proyecto a un campo novedoso y, cómo ellas mismas dicen, con mucho futuro.

Ana Elisabeth

Ana Gómez y Elisabeth Peralta han realizado juntas el proyecto final de carrera de la Diplomatura de Estadística en la FME y las dos siguen estudiando pero siguiendo caminos diferentes. Ana ha empezado Administración y Dirección de Empresas en la UB y trabaja como auxiliar de dirección en DIR (cadena de gimnasios). A Elisabeth le gusta la informática y ahora está estudiando Ingeniería Técnica Informática en la UPC y trabaja como becaria en el departamento de Teoría de la Señal y Comunicaciones. Las dos están de acuerdo en que unos buenos conocimientos de estadística son muy útiles tanto en el mundo de la empresa como en el de la informática.

7

Formación de precios en el mercado eléctrico: mejores estrategias para compradores y vendedores

Proyecto realizado por: **Elisenda Vila Jofre**
Dirigido por: **F. Javier Heredia Cervera** y **Cristina Corchero García**

El sector de la energía eléctrica en España ha afrontado diversas reformas integrales durante los últimos 10 años. La última de ellas ha tenido lugar en julio de 2006 con la puesta en marcha del Mercado Ibérico de la Electricidad. En este nuevo marco las empresas productoras de energía se plantean un problema básico: ¿cuánta energía deben producir para tener el máximo beneficio?

Para poder dar una respuesta a esta pregunta se han utilizado técnicas propias del campo de la investigación operativa. La investigación operativa, que en inglés también se conoce con el nombre de "Management Science", es una disciplina que aplica métodos analíticos avanzados para ayudar a tomar las decisiones más acertadas. Usando técnicas como la modelización matemática, la investigación operativa es capaz de proporcionar a los ejecutivos la información necesaria para tomar las decisiones más adecuadas en cada caso.

A continuación se explica cómo se ha elaborado una estrategia que ayuda a tomar decisiones en el mercado de la energía eléctrica utilizando estas técnicas.

El mercado ibérico de la electricidad

El mercado ibérico de la electricidad (MIBEL) inicia su proceso de creación en el año 2001 con la firma de un acuerdo entre las administraciones española y portuguesa para alcanzar la convergencia de los sistemas eléctricos de los dos países. No es, sin embargo, hasta julio de 2006 cuando el proyecto se hace realidad y se convierte en plenamente operativo.

El MIBEL está formado por dos entidades principales: el OMEL y el OMIP. El OMEL es el polo español y es el responsable de la gestión del mercado diario e intradiario (se encarga del mercado de la electricidad para el día siguiente) mientras que el OMIP (o polo portugués) es el responsable de la gestión de los mercados de futuro (mercado de la electricidad a largo plazo).

Por el correcto funcionamiento del sistema se necesitan las figuras del operador de mercado, que se encarga de la gestión económica del mercado de producción, y del operador del sistema, que realiza la gestión técnica del sistema eléctrico.

Este mercado comprende el conjunto de mecanismos que permiten conciliar la libre competencia en la generación de la energía eléctrica con la exigencia de disponer de un suministro de electricidad que cumpla los criterios de seguridad y calidad exigidos.

En el mercado participan, por una parte, los agentes generadores de energía y por otra los agentes compradores, que son los distribuidores, los comercializadores y los consumidores cualificados. Las transacciones de energía que los agentes negocian dentro del mercado responden a sus previsiones de demanda, de capacidad de generación y de la capacidad de la red de transporte.

En la Figura 1 podemos ver representada la estructura del mercado ibérico de energía eléctrica.

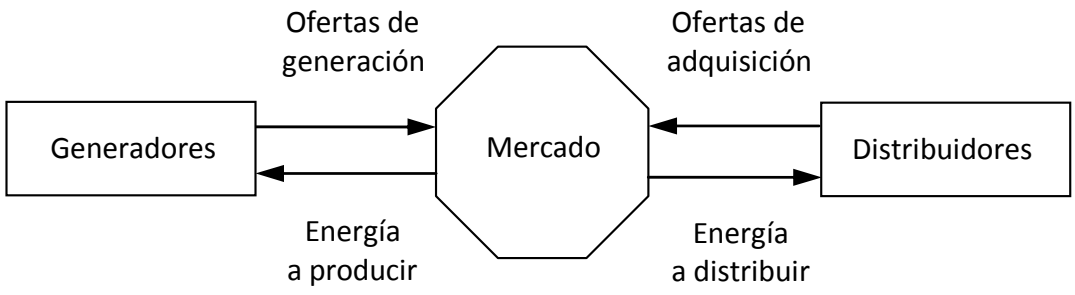


Figura 1: *Estructura del mercado ibérico de electricidad*

El mercado eléctrico incluye los mercados de futuro, el mercado diario y los mercados intradiarios. Tal como se ha comentado, el polo español controla el mercado diario e intradiario, mientras que el polo portugués controla el mercado de futuro. El mercado diario es la pieza clave dentro del mercado ibérico de la electricidad. Este mercado tiene como objetivo llevar a cabo las transacciones de energía eléctrica para el día siguiente mediante la presentación de ofertas de venta y de adquisición de energía eléctrica.

Cada día se divide en 24 periodos de programación que equivalen a las 24 horas. En cada uno de ellos se realiza una subasta de energía que fija el precio de venta y la cantidad de energía a producir. En la Figura 2 podemos ver representada la demanda de energía para un día en concreto y qué tipo de energía participó en su suministro.

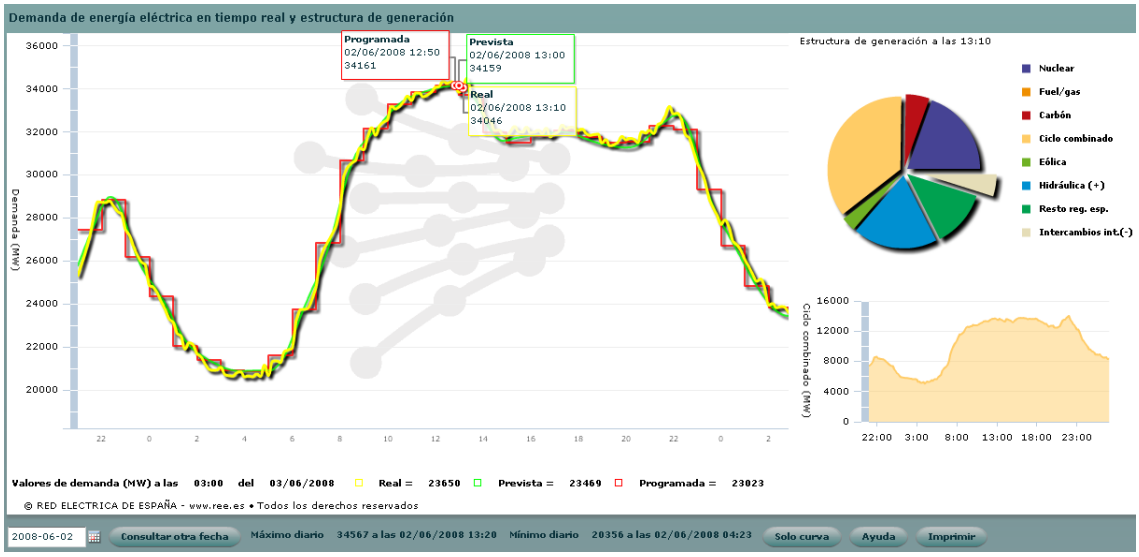


Figura 2: *Demanda de energía eléctrica del día 2 de junio de 2008 (Fuente: https://demanda.ree.es/demanda.html)*

El operador del mercado es quien realiza el proceso de subasta de la energía. El sistema básicamente consiste en recoger todas las ofertas de energía y todas las demandas con sus respectivos precios y llevar a cabo un proceso de casación. El precio en cada período horario corresponde al precio de la última oferta de producción que haya sido aceptada para poder satisfacer la demanda. Así, si el precio queda fijado en 6 c€/kWh, todas las ofertas de generación que se hayan efectuado a un precio más bajo serán aceptadas mientras que las que se hayan ofrecido a un precio más alto serán rechazadas. A través de la web de OMEL se pueden consultar todos los precios de mercado en tiempo real (www.omel.es).

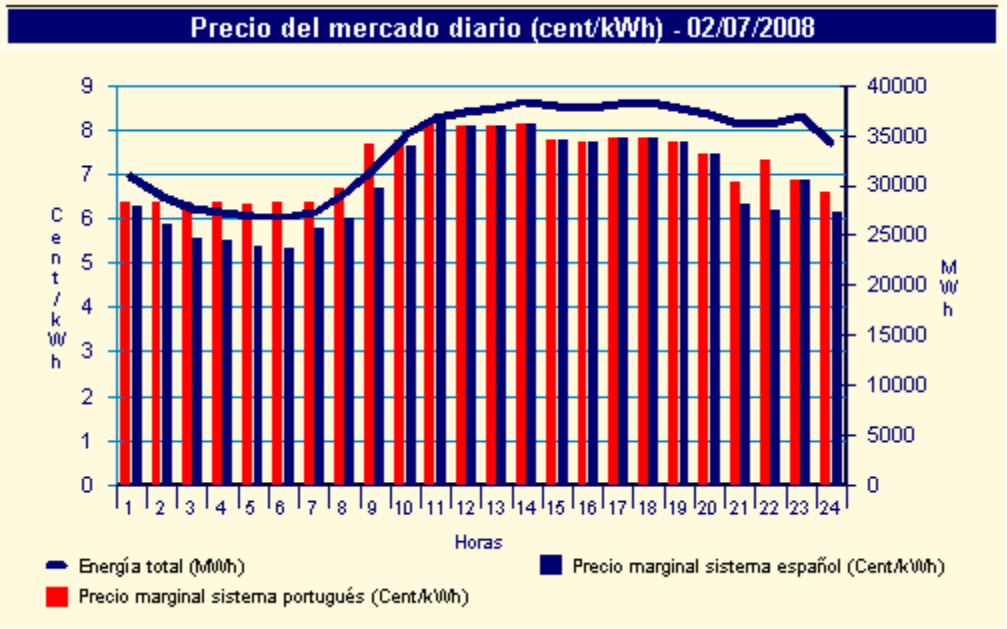


Figura 3: *Precio del mercado diario (Cent/kWh) del día 2 de julio de 2008. Fuente: www.omel.es*

Origen del problema

Dentro del mercado nos centraremos en una compañía concreta de generación de electricidad. En el entorno en que se mueve el sector eléctrico español de la energía desde su reestructuración, una de las principales tareas a desarrollar por parte de las compañías generadoras es volver a definir sus políticas de producción. Es necesario que calculen la mejor respuesta de una unidad generadora dentro de este mercado competitivo.

Para poder maximizar sus beneficios, deben decidir:

- Cuánta energía ofrecen al mercado
- En qué periodos del día
- A qué precio

Podemos escribir este problema de la manera más sencilla posible y maximizar una función que represente nuestros beneficios, que serán la diferencia entre lo que ganemos por cada unidad de energía vendida y los costes que nos supone producirla.

Podemos considerar que la función de costes de producción, $c(p)$, es una función de tipo lineal, como por ejemplo:

$$c(p) = a + bp$$

donde p representa la energía producida durante una hora (MWh) y a y b son los coeficientes de la función lineal de costes. Por otra parte, por cada unidad de electricidad vendida en el mercado ganamos λ (donde λ representa el precio de la energía en el mercado). No se puede producir lo que se quiera, ya que existen una serie de restricciones que hay que cumplir, por ejemplo, las restricciones técnicas de la unidad de producción.

El caso más sencillo podría ser que la producción tuviera que alcanzar un nivel mínimo y no pudiera superar un nivel máximo. Por ejemplo, la energía producida por una central no puede superar una cierta cantidad máxima $\bar{P}$, ni estar por debajo de una cierta cantidad mínima $\underline{P}$ (las centrales funcionan así: son máquinas muy grandes que si están puestas en marcha tienen que producir una cierta cantidad mínima y, además, necesitan unas cuantas horas para apagarse completamente).

Esta situación se puede expresar de la manera siguiente:

$$\underline{P} \leq p \leq \overline{P}$$

De forma que el problema de maximizar el beneficio queda como:

$$\max B(p) = \lambda p - c(p) = \lambda p - (a + bp) =$$

$$= (\lambda - b)p - a$$

$$\text{Sujeto a: } \underline{P} \leq p \leq \overline{P}$$

Si la realidad se correspondiera con este modelo sería muy fácil encontrar la solución óptima. Puede comprobarse que el valor máximo de los beneficios siempre se obtiene con un valor de la producción $p = \overline{P}$ si $(\lambda - b)$ es positivo y $p = \underline{P}$ si $(\lambda - b)$ es negativo.

Desgraciadamente la realidad es más difícil de describir. Uno de los principales problemas que se tienen es que los precios de mercado, que son los que al final marcarán los beneficios por unidad de energía vendida, se deciden una vez la compañía ha realizado las ofertas de venta. Así pues, cuando tienen que decidir la cantidad a producir el día siguiente no se conoce el precio que se pagará. Este precio no depende sólo de la oferta de esta compañía sino de la de todos los competidores, del precio de las ofertas de compra y de muchos otros factores. Por lo tanto, el precio de mercado es lo que se conoce como *variable aleatoria*.

Árbol de escenarios

Como ya hemos visto, las compañías eléctricas tienen la necesidad de incluir en sus modelos de optimización algún tipo de aproximación de la variable aleatoria precio de la electricidad, ya que no sabrán su valor real hasta que lo fije el Mercado Eléctrico. Hay que hacerlo, sin embargo, de forma que sea apropiada para su representación con un ordenador. Y una buena manera es mediante lo que se llama árbol de escenarios.

En general, un árbol es una forma de representar gráficamente una estructura jerárquica. Se llama árbol precisamente porque su estructura recuerda a la de un árbol, aunque se muestra hacia abajo en comparación con los árboles reales: la raíz se encuentra en la parte superior, mientras que las hojas están en la parte inferior.

Cada miembro del árbol se llama nodo. Todas las estructuras de árbol tienen un nodo que no tiene ningún nodo superior; este miembro se llama raíz (nodo A de la Figura 4). Podemos pensar pues en la raíz como el nodo inicio. En la Figura 4 se puede ver representada la estructura de un árbol.

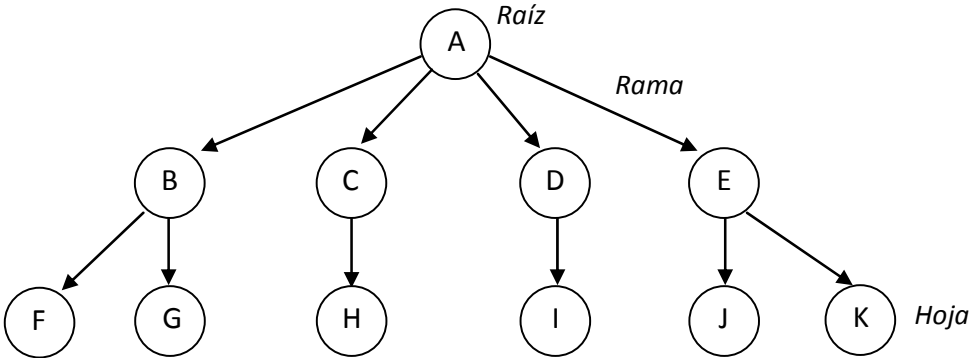


Figura 4*: Estructura de árbol*

Los nombres de las relaciones entre los nodos se llaman de la misma manera que las relaciones familiares. La terminología y propiedades básicas de un árbol son las siguientes:

- Cada nodo, excepto el nodo raíz, está conectado por una única arista (o rama) a un nodo padre, que se encuentra en un nivel superior de la jerarquía y más cerca del nodo raíz.
- La arista que conecta un nodo padre con su nodo hijo se llama rama.
- Un nodo hoja es un nodo que no tiene ningún hijo (nodo F, por ejemplo).

Un árbol de escenarios, sin embargo, tiene algunas características más que un árbol cualquiera. El nodo raíz en un árbol de escenarios representa el día de hoy y es una información conocida para nosotros. Los nodos siguientes representan los acontecimientos futuros, elementos desconocidos asociados a nuestra variable aleatoria. Los arcos que conectan los nodos representan posibles valores de la variable y su incertidumbre.

Volvemos a considerar el caso más sencillo: suponemos que tenemos la distribución del precio de mercado para una hora concreta H del día de mañana (es decir, el conjunto de valores que puede tomar la variable aleatoria y sus probabilidades). Entonces construir el árbol de escenarios para esta hora simplemente consiste en discretizar la variable continua que representa el precio de mercado para esa hora obteniendo una serie de valores concretos con una probabilidad correspondiente de forma que quede representado todo el rango de valores.

Suponemos que la distribución de los precios de mercado para la hora H sigue una distribución Normal con media 2 y desviación estándar 0,5 (Figura 5).

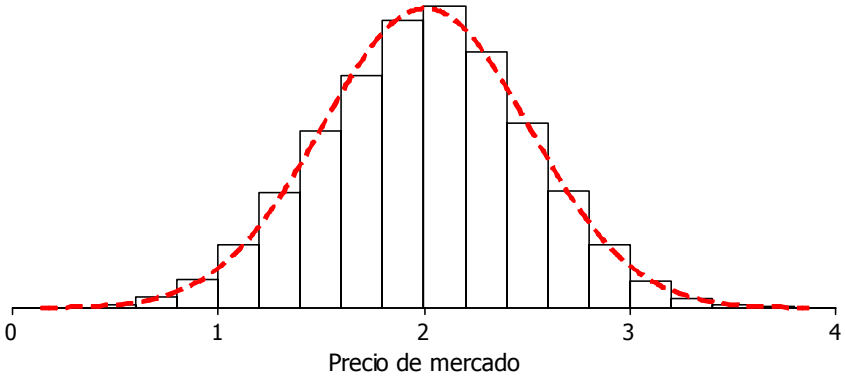


Figura 5: *Posible distribución de los precios de mercado [Normal(2; 0,5)]*

Si quisiéramos discretizar la distribución en 4 valores necesitaríamos dar esos 4 valores y sus probabilidades asociadas. De esta manera obtenemos la siguiente tabla:

Valor	Probabilidad
1.5	0.2
2	0.5
2.5	0.2
3	0.1

Así obtenemos el siguiente árbol de escenarios para la hora H:

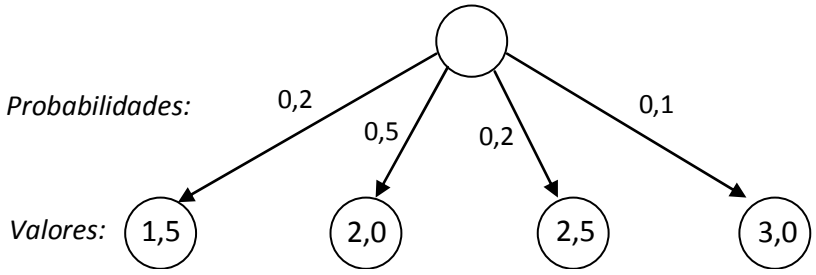


Figura 6: *Árbol de escenarios asociado a la hora H*

Para el caso estudiado en el proyecto se utilizó un árbol donde se representaban los posibles valores para "mañana" y el día siguiente para los 24 valores del precio, uno por cada hora del día. Por tanto, la estructura del árbol se complica un poco más y es necesario definir notación complementaria.

El conjunto de nodos y ramas que van desde la raíz hasta una de las hojas conforman uno de los posibles conjuntos de 24 valores del precio λ para mañana y de los 24 λ's para pasado mañana. Así pues se utilizará la palabra escenario para designar este conjunto de acontecimientos, tal como podemos ver representado en la Figura 7.

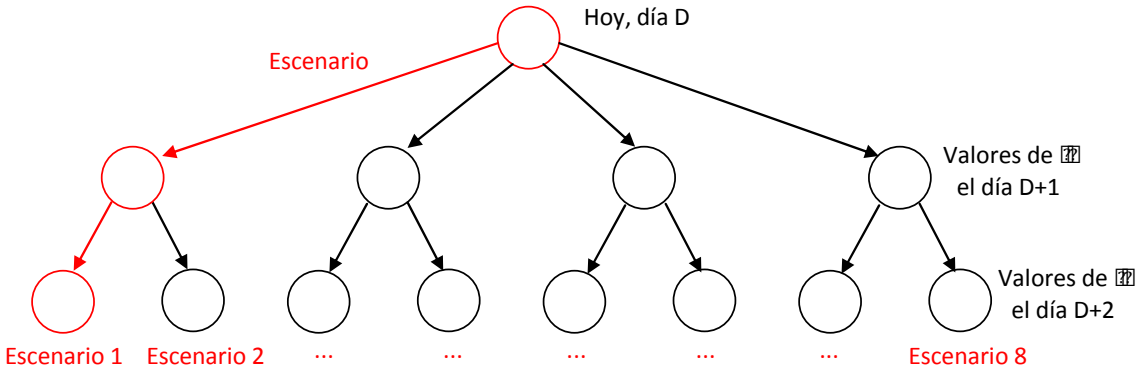


Figura 7: *Árbol de escenarios*

La situación ideal sería que todo el conjunto de posibles valores del precio de mercado quedaran representados por el conjunto de escenarios del árbol construido. Entonces, el árbol de escenarios tendría que incluir tanto los casos más optimistas como los más pesimistas.

Como ya se ha explicado, el mercado diario consta de 24 períodos horarios consecutivos. Cada nodo del árbol representa un día dividido en sus correspondientes periodos. Es decir, para cada nodo tenemos 24 valores del precio λ de mercado.

La profundidad de un nodo es la longitud del camino desde la raíz hasta el nodo. Todos los nodos que tengan la misma profundidad, es decir, que estén al mismo nivel, pertenecerán al mismo periodo o fase del algoritmo (en este ejemplo pertenecerán al mismo día).

Cada periodo está asociado a un punto del tiempo: la primera etapa contiene las realizaciones correspondientes al primer día (día D+1), la segunda etapa aquéllas que corresponden al segundo día (día D+2) y así sucesivamente.

Construcción de árboles

Hay diversas formas de enfocar la construcción de árboles de escenarios. En este caso se han usado técnicas de generación usando un método basado en la optimización. La optimización es la parte de las matemáticas que pretende encontrar la solución mínima o máxima a un problema dado. El caso más sencillo sería el de encontrar el mínimo de una función cuadrática.

En el modelo basado en la optimización es la persona encargada de tomar las decisiones quien especifica las propiedades estadísticas relevantes que se tienen que cumplir en la construcción del árbol. Por tanto, en este caso, hay que especificar las propiedades estadísticas de la distribución del precio de mercado. Estas propiedades han sido descritas a partir del tratamiento estadístico de los datos históricos disponibles. En el proceso de construcción de los árboles de escenarios se garantizan estas propiedades estadísticas haciendo que los valores del precio de mercado y las probabilidades de las ramas del árbol sean variables del problema. La función objetivo será minimizar la diferencia entre las propiedades estadísticas que se han especificado y las que tiene el árbol de escenarios que se ha construido.

Para construir el árbol, en vez de resolver un gran problema de optimización, se ha utilizado la optimización secuencial. Primero resolvemos el problema para el día de mañana (día D+1 de la Figura 7). A continuación, para cada uno de los nodos que hemos obtenido, resolvemos el problema para el día siguiente (día D+2). Así vamos encontrando todos los valores de para cada escenario de una manera progresiva.

Se puede escribir el proceso de construcción de un árbol de escenarios a través del esquema siguiente, que tiene la estructura propia de un algoritmo:

Repetir para todo el periodo

Repetir para todo nodo que pertenezca al periodo actual

Paso 1: Obtener mediante herramientas estadísticas los valores asociados al periodo, que serán los datos conocidos de nuestro problema.

Paso 2: Crear y solucionar el problema de optimización hallando los valores del precio de la energía para cada hora, de forma que se cumplan las especificaciones descritas en el paso anterior.

Paso 3: Guardar la información obtenida en las estructuras de datos correspondientes. Mediante estas estructuras de datos podremos guardar el árbol de escenarios.

Fin Repetir

Fin Repetir

Resultados

Para resolver esta serie de problemas fue necesario el uso de un lenguaje especial de programación y un programa específico para resolver este tipo de problemas de optimización. Finalmente, y después de ajustar correctamente el modelo a este caso particular, se obtuvo un árbol de escenarios que representa correctamente el conjunto de posibles valores del precio de la energía eléctrica (que como ya se ha comentado anteriormente es variable y, por lo tanto, se desconoce su valor real en un punto concreto del tiempo).

En la Figura 8 se puede ver la solución que se obtuvo al construir un árbol con 11 ramas para dos días (con 11 nodos en la primera fase que representan los posibles conjuntos de 24 precios para el primer día y de cada uno de ellos 11 posibles precios para el segundo día). Cada una de las líneas del gráfico representa uno de los 121 escenarios. En negro está representada la previsión del precio para estas 48 horas.

Con los árboles obtenidos se solucionó el problema de la oferta óptima en el mercado diario de la energía eléctrica, que era el problema que realmente nos interesaba resolver. Así pues, ya estábamos en condiciones de decirle a la compañía generadora la cantidad de energía que tenía que ofrecer en el mercado para el día siguiente. En la Figura 9 podemos ver gráficamente la energía que una central térmica vendería al mercado en cada una de las 48 horas siguientes para diferentes escenarios de precios.

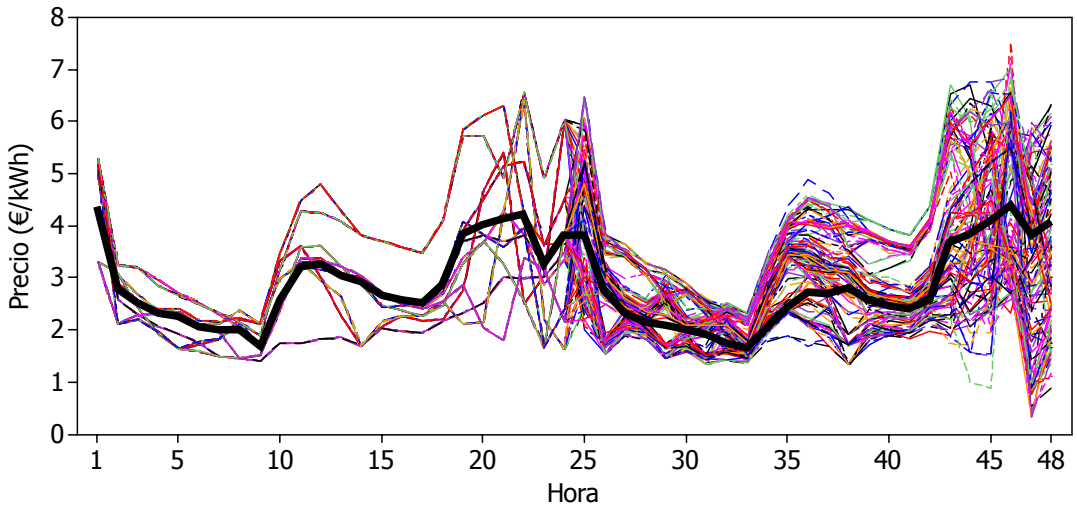


Figura 8: *Escenarios de un árbol con 11 ramas*

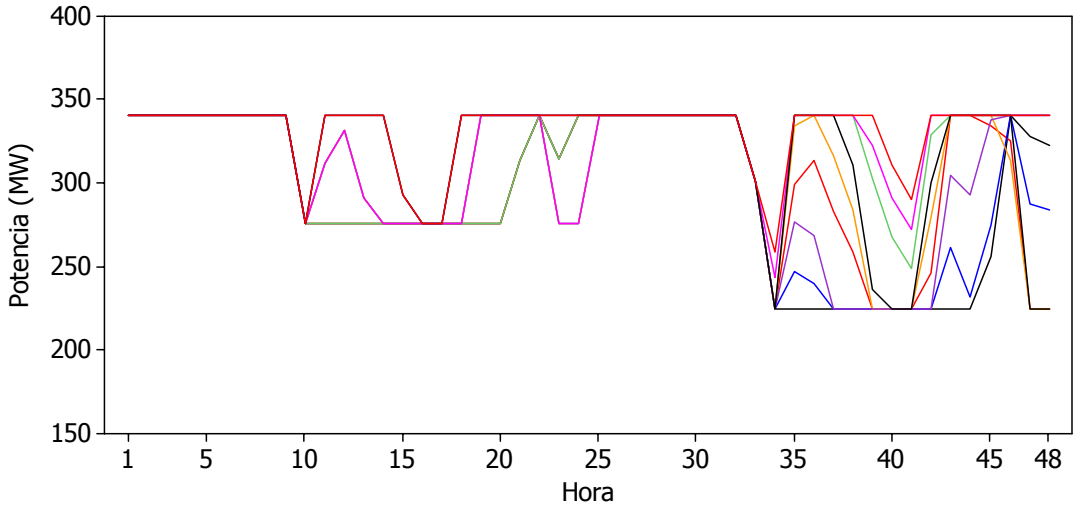


Figura 9*: Potencia de una unidad térmica en un grupo de escenarios de un árbol con 11 ramas*

Conclusiones

La generación de los árboles de escenarios es una pieza clave a la hora de resolver problemas donde hay variables aleatorias.

Mediante este caso se ha podido ver un ejemplo de cómo los modelos matemáticos de la investigación operativa nos dan herramientas para resolver problemas reales de toma de decisiones. Hemos visto cómo estos modelos ayudan a dar una respuesta a

problemas complejos de toma de decisiones como los que se plantean las compañías de producción de energía eléctrica:

- Cuánta energía ofrecen al mercado
- En qué periodos del día
- A qué precio

La gran ventaja de estas técnicas es que pueden ser aplicadas a un gran abanico de situaciones diferentes: tanto nos puede servir para obtener la oferta óptima en el mercado diario de la energía eléctrica como para diseñar una estrategia de inversión en bolsa o para saber cuánto dinero tienen que poner los bancos en sus cajeros automáticos. Las aplicaciones pueden ser infinitas, y los límites sólo los ponemos nosotros.

(Proyecto de la Diplomatura de Estadística, presentado en septiembre de 2007 con el título "Generació d'escenaris per a l'optimització de l'oferta al mercat elèctric")

A través de este ejemplo hemos podido ver las posibilidades de unas técnicas que se han aplicado para resolver problemas tan diversos como el diseño de las colas en los parques temáticos Disney, la optimización de las rutas de pequeños distribuidores locales o la mejora de los horarios de la tripulación de líneas aéreas. Podéis encontrar muchos más ejemplos en la web: www.scienceofbetter.org

Elisenda Vila estudió la Diplomatura de Estadística en la UPC y ahora está realizando el último curso de la Licenciatura y el Máster en Estadística e Investigación Operativa. Además de ser una estudiante brillante, participa y conoce como pocos la vida de la Facultad. Colabora como becaria en Ordenación de Estudios, es representante de los estudiantes en la Junta de Facultad y también en su Comisión Permanente, y participa en el curso de "Bioestadística per a no estadísticos" que se imparte en la Facultad. También ha formado parte de la comisión que ha diseñado los nuevos estudios del Grado Interuniversario de Estadística, organizado por la Universidad de Barcelona y la UPC.

8

Estudio de mercado para definir las características de una nueva variedad de galletas de chocolate

Proyecto realizado por: **Sara Solanes i Giner**
Dirigido por: **Lourdes Rodero de Lamo**

El lanzamiento de nuevos productos casi siempre requiere importantes inversiones de dinero (en diseño, nuevas líneas de producción, publicidad...) y hay que estar seguros de que tendrán la aceptación prevista. De lo contrario, las pérdidas pueden ser muy grandes. Tampoco es una buena noticia que la demanda sea mucho mayor de la esperada si la empresa no es capaz de atenderla. Seguramente este exceso de demanda, creada en la campaña de lanzamiento, será cubierta por otras empresas de la competencia. Mal negocio.

En este terreno, como en muchos otros, conviene tomar las decisiones basándose en información de calidad (qué tipo de producto apreciarán más los clientes, cuánto se venderá, qué tipo de envase gustará más, qué precio estarán dispuestos a pagar...). Esta información la pueden suministrar los expertos, con un grado de fiabilidad difícil de prever, o se puede obtener a través de un estudio de mercado, preguntando a los posibles consumidores las cosas que queremos saber.

Sin embargo, hay que escoger bien la muestra, y asegurarse de que los datos que se obtendrán serán correctos, también habrá que analizarlas de la forma adecuada para sacar la información que interesa con el grado de confianza previsto.

Objetivos del estudio

Se quiere obtener información que facilite la toma de decisiones en una empresa para el lanzamiento al mercado de una nueva variedad de uno de sus productos, que en este caso es una nueva galleta de chocolate.

En concreto, los objetivos son:

1. *Conocer si la nueva variedad incrementará las ventas y encontrar la combinación de variedades óptima.*

   En estos momentos el fabricante ya tiene 3 variedades en el mercado y quiere saber cuánto aumentarán las ventas al añadir una variedad nueva. También se quiere aprovechar el estudio para analizar qué pasaría si se lanzaran otras variedades distintas. Se aplicará el análisis TURF (*Total Unduplicated Reach Frequency*).

2. *Cuantificar la aceptación de la nueva variedad.*

   Escogida la variedad que se va a lanzar al mercado, se quiere conocer la aceptación que tendrá.

3. *Determinar la mejor alternativa de producto de la nueva variedad.*

   Existen 3 posibles alternativas de fabricación de la nueva variedad y se quiere saber cuál gustará más a los consumidores.

4. *Determinar el envase que tiene más aceptación entre el público consumidor.*

   Se quieren comparar dos tipos de envase para saber cuál gusta y llama más la atención.

5. *Determinar el precio óptimo.*

   Fijar el precio es una de las decisiones más difíciles y arriesgadas. Se quiere saber qué opinan los consumidores sobre el precio que debería tener este producto. Se aplicará el modelo PSM (*Price Sensitivity Model*).

Recogida de los datos

Para obtener la información que se necesita hay que recoger datos. Estos datos se obtienen mediante una encuesta realizada a una muestra de consumidores de galletas de chocolate.

Características de la muestra

Debido a la naturaleza de esta nueva galleta, se ha investigado sobre las siguientes poblaciones objetivo (*targets*, en el argot de estos estudios):

a) Madres de entre 30 y 55 años con hijos/as menores de 15 años. Esta población se ha dividido en dos intervalos de edad: de 30 a 40 años y de 41 a 55, y para cada intervalo se ha buscado el mismo número de participantes.

b) Niños/as entre 9 y 12 años, repartidos al 50% según el sexo.

Se han elegido madres ya que normalmente son las responsables de las compras en el hogar y niños/as porque son el consumidor final (el llamado *core target*). Ambos segmentos deben ser consumidores de galletas rellenas de chocolate y ser residentes en las ciudades de Barcelona, Madrid, Sevilla, Bilbao, Valencia y Zaragoza (una sexta parte de la muestra de cada ciudad).

Se podría haber planteado la recogida de datos de forma que no hiciera falta que fueran consumidores de este tipo de galletas, pero el objetivo era ofrecer un nuevo producto a los consumidores de productos similares. Ampliando el *target* a no consumidores habríamos obtenido valoraciones más bajas porque seguramente no les gustan este tipo de galletas.

Tamaño de la muestra

Los primeros cálculos para determinar el tamaño de la muestra se hicieron teniendo en cuenta las valoraciones que se querían pedir a los entrevistados.

Cada persona tenía que valorar seis conceptos. El concepto es la descripción de la galleta, por ejemplo: "galleta rellena de chocolate blanco con sabor a naranja", y esta valoración se tiene que hacer sin ver ni probar la galleta. Después, se valoran 3 variedades de uno de los conceptos probando, ahora sí, las galletas.

Se decidió que cada persona probaría 2 de las 3 galletas y no las 3, por una cuestión de duración y calidad de la entrevista. Hay que tener en cuenta que si una persona tiene que valorar muchos productos, cada vez lo hace con menos rigor, ya que la encuesta se hace pesada y baja la concentración de la persona encuestada.

Teniendo en cuenta todo esto se propuso una muestra de 300 individuos para poder obtener una muestra de 200 entrevistas por variedad (100 de madres y 100 de niños), ya que habitualmente se considera que es necesario un mínimo de $n = 100$

entrevistas por producto y target para tener un margen de error de $\pm 10\%$ y éste era el margen de error que el fabricante estaba dispuesto a aceptar.

Por cuestiones económicas, sin embargo, fue necesario reducir el tamaño de la muestra. Finalmente se planteó una muestra teórica de 225 entrevistas, obteniendo una muestra de 150 para cada variedad (75 de madres y 75 de niños/as).

Las 225 entrevistas se repartieron en las ciudades antes mencionadas, que son las que tienen más habitantes de España. También hay que tener en cuenta que distribuir la muestra en diferentes ciudades facilita la captación y hace que el ritmo del trabajo de campo sea más rápido. Para la selección de las madres se decidió tomar dos intervalos de edad, por no entrevistar, por ejemplo, sólo a las madres de 30 años, ya que estas pueden tener un punto de vista diferente al de las madres de 55 años. Además, aunque no es un objetivo importante de este estudio, si nos interesara podríamos detectar la existencia de diferencias significativas entre los dos intervalos de edad.

Los tamaños de muestra previstos, y los que finalmente se obtuvieron, son los que se recogen en la Tabla 1.

Tabla 1: *Tamaños de muestra real y prevista (en cursiva y entre paréntesis) para cada segmento*

		Barcelona	Madrid	Valencia	Sevilla	Bilbao	Zaragoza	TOTAL
Madres	30 a 40 años	11 *(10)*	10 *(*10)	11 *(*9)	9 *(*10)	10 *(*10)	15 *(*9)	66 *(*58)
	41 a 55 años	9 *(*10*)*	10 *(*9)	10 *(*10)	9 *(*9)	9 *(*9)	5 *(*10)	52 *(*57)
Niños/as	Niños	11 *(*10*)*	10 *(*9)	8 *(*9)	11 *(*9)	9 *(*9)	9 *(*9)	58 *(*55)
	Niñas	9 *(*9)	10 *(*10)	11 *(*9)	10 *(*9)	9 *(*9)	9 *(*9)	58 *(*55)
TOTAL		40 *(*39)	40 *(*38)	40 *(*37)	39 *(*37)	37 *(*37)	38 *(*37)	234 *(*225)

Estructura del cuestionario

El cuestionario empieza con unas preguntas filtro para de verificar que la persona seleccionada cumple con los requisitos necesarios para formar parte de la muestra. A continuación se realizan las preguntas sobre el concepto, mostrando las descripciones de 6 conceptos de producto, de los cuales 3 ya están al mercado (C1, C5, C6) y los otros 3 corresponden a las nuevas variedades (C2, C3, C4), para poder averiguar cuál es la combinación de variedades que maximizará las ventas (objetivo 1).

Después se realizan preguntas específicas sobre el concepto C4, el que se quiere introducir al mercado, para valorar la aceptación que tendrá (objetivo 2).

A continuación se pasa a la fase del producto, para identificar cual de tres posibles formas (L, H y G) de elaborar el concepto C4 gusta más. Cada persona entrevistada prueba dos productos y opina sobre ellos (objetivo 3).

Después de catar y valorar las galletas se muestran 2 formatos de envase (A y B) y se pregunta por la preferencia (objetivo 4). Finalmente, se hacen preguntas relativas al precio del producto (objetivo 5).

La entrevista acaba con algunas preguntas demográficas para poder clasificar a la persona encuestada.

Captación de personas para la muestra. Entrevistas

Las entrevistas se realizaron en salas de hoteles céntricos de cada una de las ciudades. La captación se realizó en la calle, cerca del hotel (siempre son lugares muy concurridos) donde los encuestadores buscan a personas que cumplan con las características que se precisan, en este caso madres y niños/as.

Una vez encuentran a alguna de estas personas, se les realizan unas preguntas filtro y si las pasan y la persona acepta colaborar en el estudio, se le acompaña a una sala del hotel donde está todo el material necesario para realizar la encuesta.

Las encuestas se llevan a cabo por el sistema llamado CAPI (*Computer Assisted Personal Interviewing*) que consiste en ir entrando las respuestas directamente a un ordenador. De esta forma se recogen los datos con los filtros y las consistencias controladas (detección automática de posibles errores en la introducción de datos).

Como ya se ha dicho anteriormente, cada una de las personas seleccionadas tendrá que valorar 6 conceptos. Cuando se califican tantos, las valoraciones a partir del segundo concepto están influenciadas por las valoraciones anteriores. Así que, con el fin de eliminar el efecto que tiene opinar de un concepto en primer lugar, en segundo lugar, etc., lo mejor que se puede hacer es aleatorizar el orden de las valoraciones. Es decir, los conceptos se presentaron de forma aleatoria a cada uno de los entrevistados.

Lo mismo pasa cuando se prueban dos productos, las valoraciones del segundo producto están influenciadas por las del primero. Con el fin de eliminar el efecto del orden, se va rotando el orden de prueba, es decir, la mitad de los encuestados probaron primero el producto L y después el producto H, por ejemplo, y la otra mitad probó primero el producto H y después el producto L.

Combinación de variedades que optimizan las ventas

Después de explicar cómo funcionará la entrevista y de hacer unas primeras preguntas para la clasificación del entrevistado, se le enseñan unas cartulinas con la descripción de los conceptos. Después de mostrar cada cartulina (el orden de presentación es aleatorio) el entrevistador pregunta:

> *"Pensando ahora en la variedad de galleta que acaba de leer, y en caso de que estuviera disponible en su tienda habitual al precio que Usted compra habitualmente, ¿podría decirme hasta qué punto estaría Usted dispuesta a comprar esta galleta?"*

Y las respuestas posibles, para cada concepto, son:

Seguro que la compraría	Probablemente la compraría	No sé si la compraría	Probablemente no la compraría	Seguro que no la compraría
☐	☐	☐	☐	☐

A cada una de estas respuestas se le asigna un valor de 5 ("Seguro de que la compraría") a 1 ("Seguro de que no la compraría") y la media de estos valores cuantifica la intención de compra para cada concepto.

Resultados de la intención de compra

La variedad que más gusta a las madres es la del concepto 5, con una diferencia que resulta significativa con respecto a todas las otras. Este era el resultado esperado, ya que esta variedad ya se encuentra en el mercado y es el producto estrella del fabricante. Después tenemos las variedades de los conceptos C6 y C1, sin diferencias significativas entre ellas. Estas dos variedades también son, junto con la C5, las que están en el mercado actualmente. Seguidamente encontramos el concepto C3 y después el C4, que es el que se quiere introducir al mercado y, finalmente, el C2.

Con respecto a los niños (Figura 1), éstos tienen dos variedades preferidas, que son las correspondientes a los conceptos C5 y C1 entre las cuales no hay diferencias significativas (recordamos que ambas ya están en el mercado). Seguidamente viene la variedad del concepto C4, que es la que se quiere introducir y que está especialmente enfocada a los niños. A continuación encontramos la variedad del concepto C6 (que también está en el mercado), aunque no hay ninguna diferencia significativa respecto a la anterior. Así pues, estadísticamente podemos considerar que los niños tienen la misma opinión de las variedades C4 y C6. Finalmente, tenemos el concepto C2 que es, con diferencia, el que menos gusta a los niños, igual que a las madres.

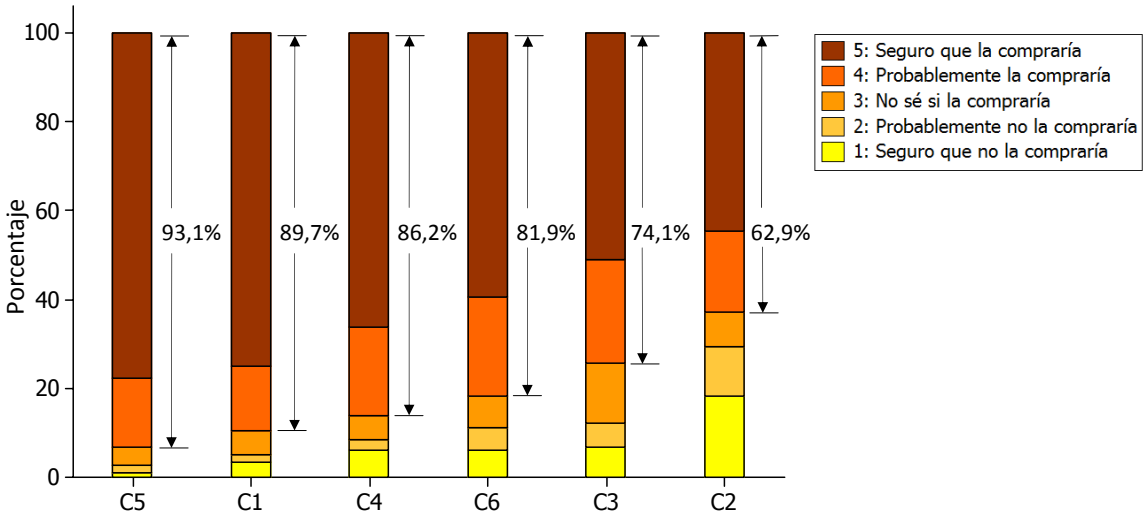


Figura 1: *Resultados de intención de compra. Grupo de niños/as (n = 116)*

Análisis TURF (Total Unduplicated Reach Frequency)

Este análisis tiene como objetivo encontrar la combinación de variedades de un producto que maximiza su penetración en el mercado. Partiendo de las respuestas a las preguntas sobre intención de compra de cada variedad se evalúa cuál es la combinación que consigue mayor el porcentaje global de intención de compra.

Por ejemplo, si queremos lanzar en el mercado dos variedades de un producto, de entre 3 posibles, la intención de compra para cada una de ellas podrían ser: V1: 60%, V2: 65% y V3: 50%.

A la vista de estos resultados parece que la mejor decisión sería no lanzar la variedad V3, ya que tiene menos intención de compra, pero si pudiéramos observar los porcentajes de los que han dicho que comprarían una variedad y después también han dicho que comprarían otra, estos resultados podrían ser:

> V1 + V2 = 85 % (es decir, el 85% de los entrevistados ha dicho que compraría conjuntamente la variedad 1 y la variedad 2)
>
> V1 + V3 = 80 %,
>
> V2 + V3 = 95 %.

Por lo tanto, la combinación que abarca más mercado es la formada por las variedades V2 y V3. Podría ser que V1 y V2 tengan intenciones de compra parecidas porque las galletas son parecidas, pero no las comprarían al mismo tiempo.

Aplicación del análisis TURF a los datos obtenidos

A partir de las respuestas obtenidas a las preguntas sobre intención de compra se puede construir una tabla como la que se presenta en la Figura 2, donde se indica el porcentaje de entrevistados que compraría alguna de las variedades indicadas. Sólo se ha reproducido la primera parte de la tabla, hasta las combinaciones de 2 variedades, pero la tabla completa contiene también las combinaciones de 3, 4, 5 y 6 variedades (en general, para k variedades el número total de combinaciones es $2^k - 1$). El análisis se ha hecho sólo con las respuestas que manifiestan intención de compra segura (responden "Seguro de que la compraría") estas respuestas son las que, en el lenguaje de esta metodología, se denominan las "*Top Box*".

	C5	C1	C6	C3	C4	C2	Alcance (%)	Media Favorable	Desviación Tipo
1 variedad	■						77,1	1	
		■					66,9	1	
			■				57,6	1	
				■			51,7	1	
					■		44,1	1	
						■	30,5	1	
2 variedades	■	■					85,6	1,68	0,47
	■		■				84,7	1,59	0,49
	■			■			84,7	1,52	0,50
	■				■		86,4	1,40	0,49
	■					■	83,1	1,30	0,46
		■	■				78,0	1,60	0,49
		■		■			78,0	1,52	0,50
		■			■		75,4	1,47	0,50
		■				■	72,0	1,35	0,48
			■	■			68,6	1,59	0,49
			■		■		72,0	1,41	0,50
			■			■	66,1	1,33	0,47
				■	■		68,6	1,40	0,49
				■		■	59,3	1,39	0,49
					■	■	57,6	1,29	0,46

... Continuación con los resultados de las combinaciones de 3, 4 , 5 y 6 productos simultáneamente.

Figura 2: *Tabla (parcial) con los resultados a las preguntas sobre intención de compra (madres)*

El alcance indica el porcentaje de la muestra que seguro que comprará alguna de las variedades indicadas. Por ejemplo, tenemos un 77,1% de las madres que dicen que seguro que comprarán el concepto C5, y el 85,6% han dicho que comprarían el concepto C5, el C1, o los dos simultáneamente.

"Media favorable" es la cantidad media de variedades que comprarán las personas que compren alguna de las variedades indicadas. Por ejemplo, de las madres que comprarán C5 y/o C1, tenemos algunas que sólo comprarán una variedad y otras que comprarán las dos; en promedio comprarán 1,68 variedades (con una desviación estándar de 0,47).

Mirando el gráfico de la Figura 3 (construido a partir de los resultados de la tabla de la Figura 2, pero completa) vemos que si sólo tuviéramos una variedad en el mercado ésta debería ser la C5 y tendríamos una penetración del 77,1%. Cuando añadimos otra variedad llegamos al 86,4% y la pareja que obtiene esta penetración es la formada por las variedades C5 y C4. Este resultado puede sorprender ya que C4 es la variedad que ha quedado en 5ª posición según la intención de compra individual, pero como se ha visto en la descripción del análisis TURF, esto es perfectamente posible.

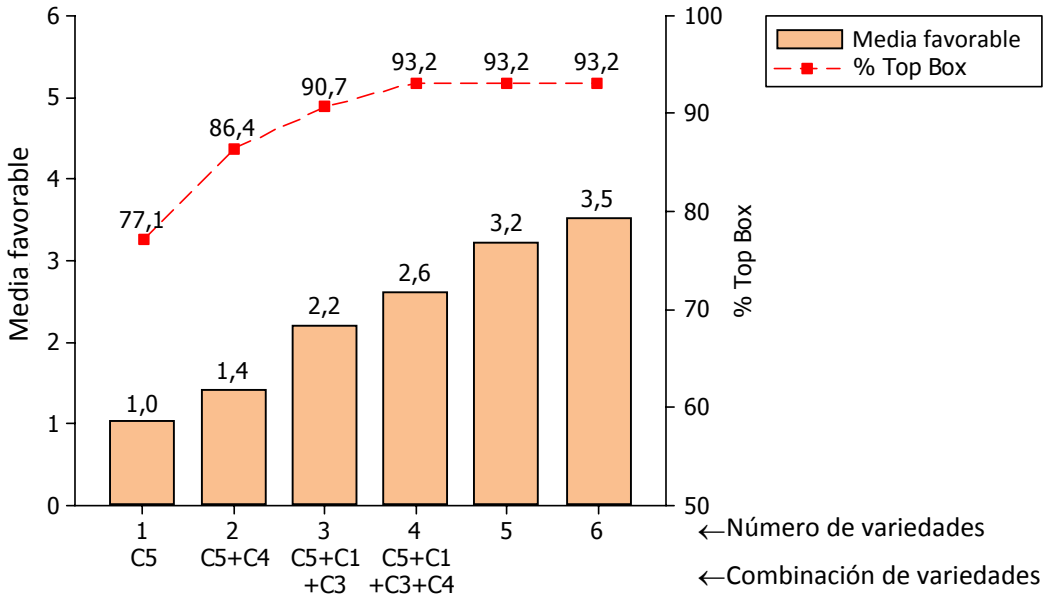


Figura 3: *Análisis TURF sobre intención de compra. Respuestas de las madres*

Al añadir una tercera variedad aumentamos la penetración hasta un 90,7% con las variedades C5, C3 y C1 (ha desaparecido C4). Seguidamente, si tenemos 4 variedades en el mercado, la penetración todavía aumenta pero menos que en los anteriores casos y llega a un 93,2%. A partir de aquí, tanto si tenemos en el mercado 5 o 6 variedades, la penetración ya no aumenta con respeto a tener 4.

Con respecto a los niños obtenemos valores de penetración más elevados que en las madres, por tanto, podemos volver a pensar que estamos hablando de un producto bastante enfocado a los niños. El 77,6% del mercado lo obtendremos con la variedad C5 y esta variedad entra en todas las combinaciones que nos maximizan la penetración.

Análisis de la aceptación de la nueva variedad

El fabricante ya suponía, por su experiencia y su conocimiento del mercado basado en estudios anteriores, que el concepto C4 era el más adecuado para completar su oferta, por lo que se quería valorar de forma cuantificada el nivel de aceptación de esta nueva galleta.

La valoración general es bastante buena, las madres dan una valoración media de casi 7 y los niños la valoran con una media superior a 8. La diferencia entre estas medias es estadísticamente significativa.

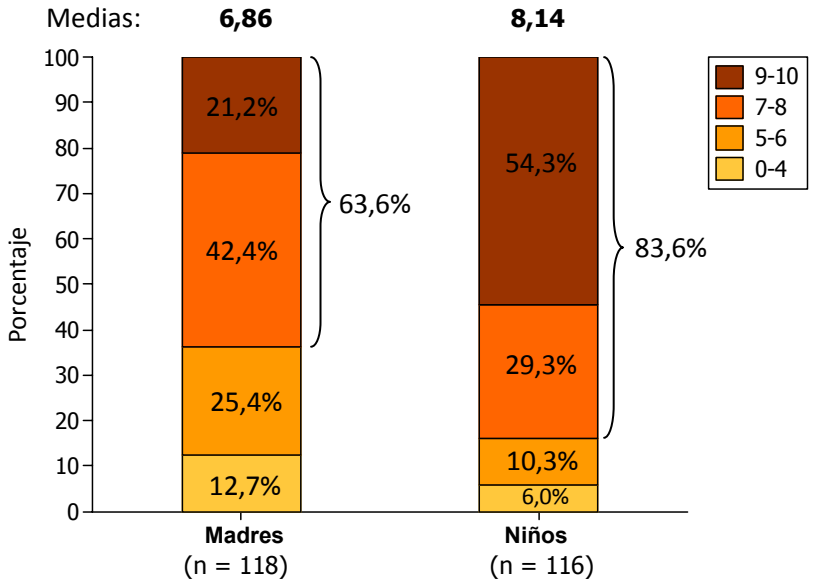


Figura 4: *"¿Qué valoración daría a la descripción del concepto de la galleta 4?"*

Las respuestas a la pregunta "¿Qué es lo que más le ha gustado del producto descrito?" están resumidas en la Figura 5. Observamos que lo que gusta más de esta nueva galleta, tanto a las madres como a los niños, es el tipo de relleno que lleva. El fabricante tenía la duda de si rellenar la galleta con chocolate o con crema de chocolate (chocolate rebajado con mantequilla y otros ingredientes) por lo que necesitaba saber si esa diferencia era importante para las madres. Las respuestas

ponen de manifiesto que casi el 85% de las madres consideran que es mucho mejor el chocolate que una crema de chocolate. El 13%, en cambio, prefiere la crema de chocolate y un 2,5% no se pronuncia.

Lo que **más** ha gustado:

	Madres	Niños	Total
El relleno	62	98	160
La galleta	14	25	39
Es de la marca X	19	10	29
Es nueva	17		17
El chocolate	8	3	11
El aspecto	5	3	8
El envase	7		7
El sabor	7		7
Nada	6		6
El diseño	4		4
Otras cosas	15	14	29

Lo que **menos** ha gustado:

	Madres	Niños	Total
Nada	68	80	148
El relleno	31	15	46
La galleta		5	5
NS/NC	8	7	15
Otras cosas	15	10	25

Figura 5: *Lo que más y lo que menos ha gustado del concepto C4*

Después de que el entrevistado haya profundizado más sobre el concepto 4 y lo conozca mejor, se le vuelve a preguntar sobre su intención de compra para ver si el mayor conocimiento sobre el producto aumenta o disminuye esa intención. En esta segunda parte se han obtenido mejores resultados que al inicio, tanto para las madres como para los niños.

Análisis de la mejor alternativa para la nueva variedad

Existen tres alternativas posibles (que llamamos L, H y G) para fabricar esta nueva variedad de galleta y cada persona entrevistada ha probado y opinado sobre dos de las tres variedades. La Figura 6 presenta un resumen de las respuestas sobre qué galleta ha gustado más. No hay ninguna diferencia significativa entre las madres y los niños.

Cuando analizamos la preferencia del producto en general entre las personas que han probado las galletas L y H, observamos que tanto las madres como los niños prefieren la galleta H y esta preferencia es estadísticamente significativa. Con respecto a las personas que han probado las galletas L y G, tanto las madres como los niños prefieren la galleta G y esta preferencia también es estadísticamente significativa. Finalmente si miramos la preferencia entre las personas que han probado las galletas H y G,

observamos que tanto las madres como los niños prefieren la galleta G, aunque en este caso la preferencia es significativa sólo para el grupo de los niños.

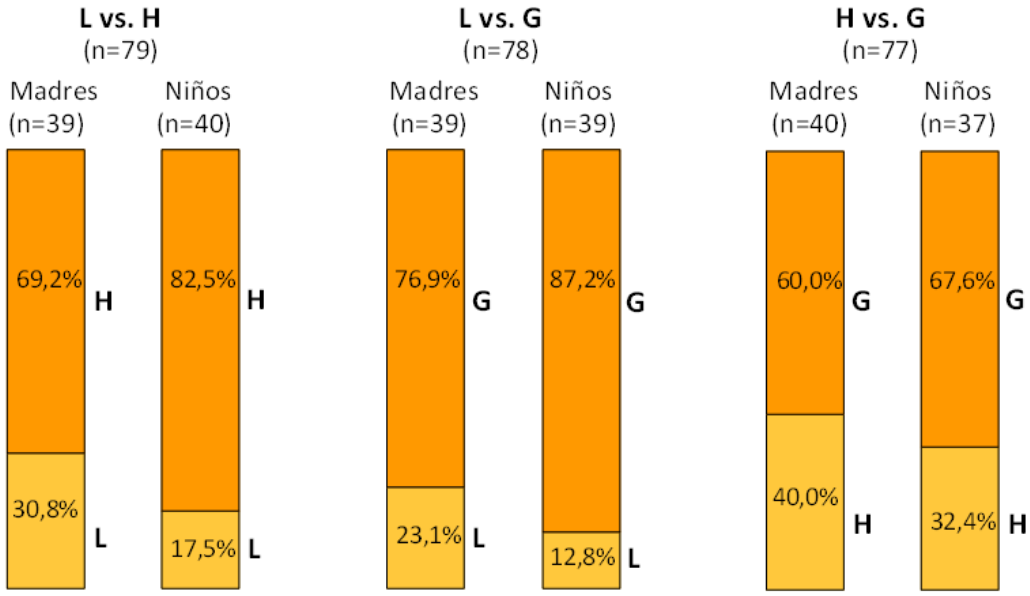


Figura 6: *"De las dos galletas que ha probado, en general, ¿cuál le ha gustado más?"*

Mejor alternativa para el envase

Recordemos que se muestran dos alternativas de envase para ver cuál gusta más a los posibles consumidores. El envase A (Figura 7, izquierda) es el típico envase de 12 galletas rellenas donde están todas juntas dentro de un envase cilíndrico, en cambio, el envase B (Figura 7, derecha) es una caja con 4 paquetes independientesde 4 galletas cada uno.

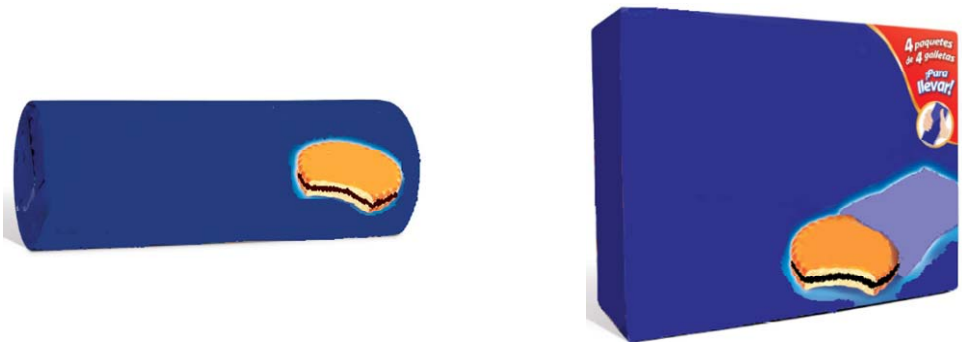


Figura 7: *Las dos alternativas consideradas para el tipos de envase*

Tanto a las madres como a los niños, se les ha preguntado sobre su preferencia entre estos dos envases. Tal como se ve en la Figura 8 gana muy claramente el envase B.

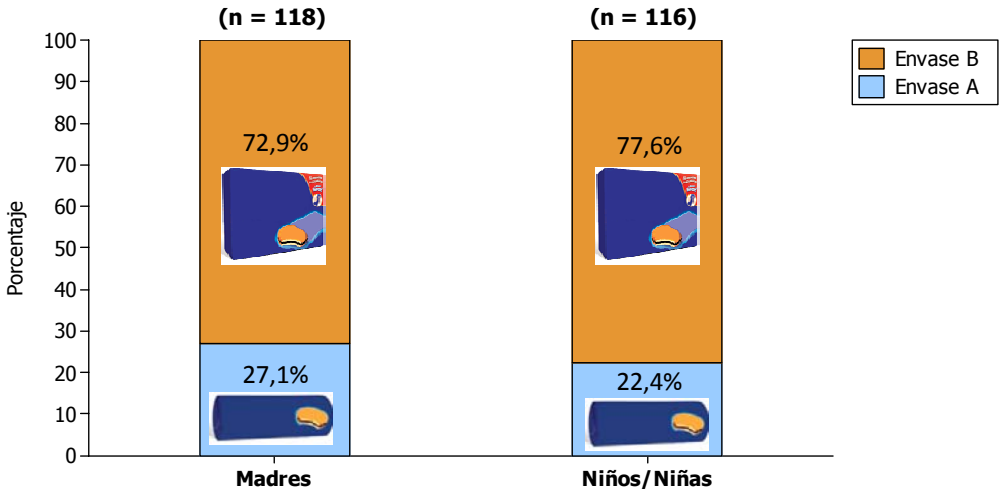


Figura 8: *Respuestas a la pregunta: "De los dos envases que le mostramos, ¿cuál le gusta más?"*

Análisis del precio óptimo

Fijar el precio es una de las decisiones más delicadas que se tienen que tomar cuando se lanza un nuevo producto al mercado. Un precio demasiado alto puede asustar, mientras que infravalorar el producto puede ser un error muy costoso. Es crucial, por lo tanto, tomar estas decisiones con la máxima información.

El análisis que se ha utilizado en este estudio fue propuesto por el economista alemán Peter Van Westendorp en el año 1970 y se conoce con el nombre de *Price Sensitivity Model*. Se basa en la realización a los encuestados de cuatro preguntas relacionadas con el precio del producto. Estas preguntas son:

1. ¿A qué precio considera que este producto es demasiado barato, y no puede ser bueno?
2. ¿A qué precio considera que este producto es barato?
3. ¿A qué precio considera que este producto es caro?
4. ¿A qué precio considera que este producto es demasiado caro, y no vale lo que se tiene que pagar por él?

A partir de las respuestas obtenidas se pueden construir gráficos como el de la Figura 9, que representa la proporción de personas que consideran que el producto es demasiado barato en función de su precio. Por ejemplo, si el precio es de 0,5 €, el 95% de las madres encuestadas piensan que es demasiado barato, por 1,1 € sólo el 5% considera que se demasiado barato, y a partir de 1,6 € prácticamente ya nadie lo considera así. Estos datos son los obtenidos en la encuesta cuando se preguntaba por el precio del producto en el envase A.

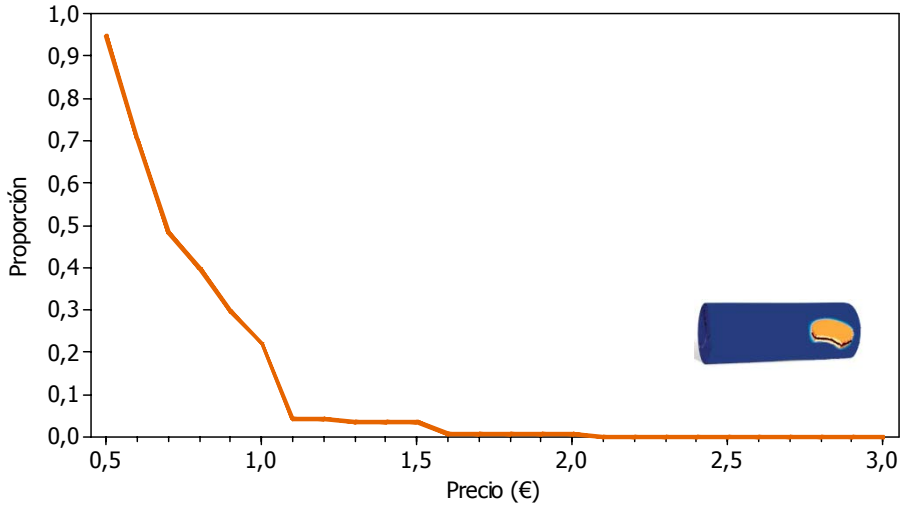


Figura 9: *Proporción de encuestados que consideran que el producto es demasiado barato en función del precio, cuando se presenta en el envase A*

Superponiendo en este gráfico la curva que representa la proporción de encuestados que consideran el producto demasiado caro en función del precio, se obtiene la Figura 10. El punto de cruce de las dos curvas es el que se considera precio óptimo (*Optimum Pricing Point*: OPP).

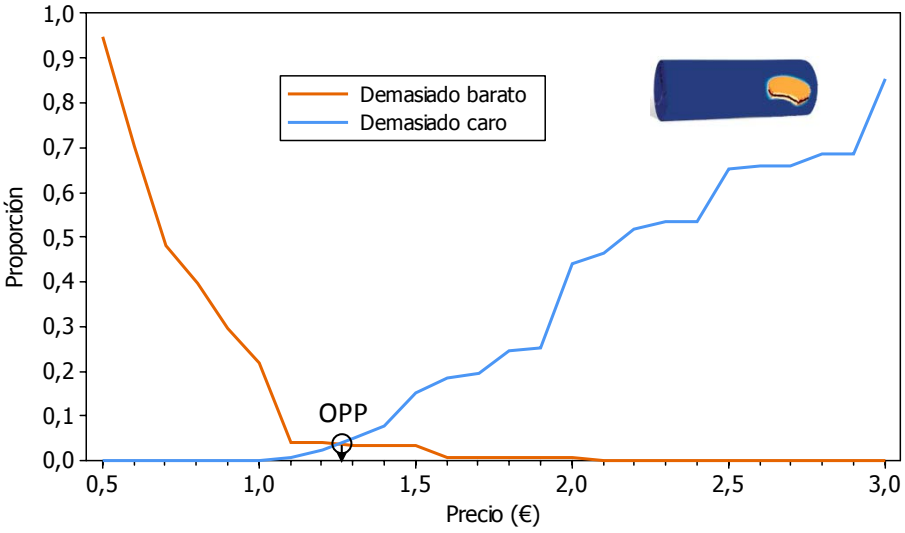


Figura 10: *Determinación del OPP (Optimum Pricing Point) para el producto en el envase A*

Las Figuras 11 y 12 incluyen también las curvas que representan las proporciones de encuestados que consideran el producto simplemente caro o barato en función del precio, en los envases A y B respectivamente. Los puntos de intersección de estas curvas proporcionan información sobre qué precio conviene establecer. Además del OPP (*Optimum Pricing Point*), estos puntos son:

- IPP (*Indifference Price Point*): Según Van Westendorp normalmente representa o la mediana del precio actual pagado por los consumidores o el precio del producto de una importante marca líder del mercado.

- PMC (*Point of Marginal Cheapness*) y PME (*Point of Marginal Expensiveness*): Estos dos puntos marcan el rango de precios aceptable para el producto. Según Van Westendorp, en mercados asentados, pocos productos competitivos tienen precios fijados fuera de este rango.

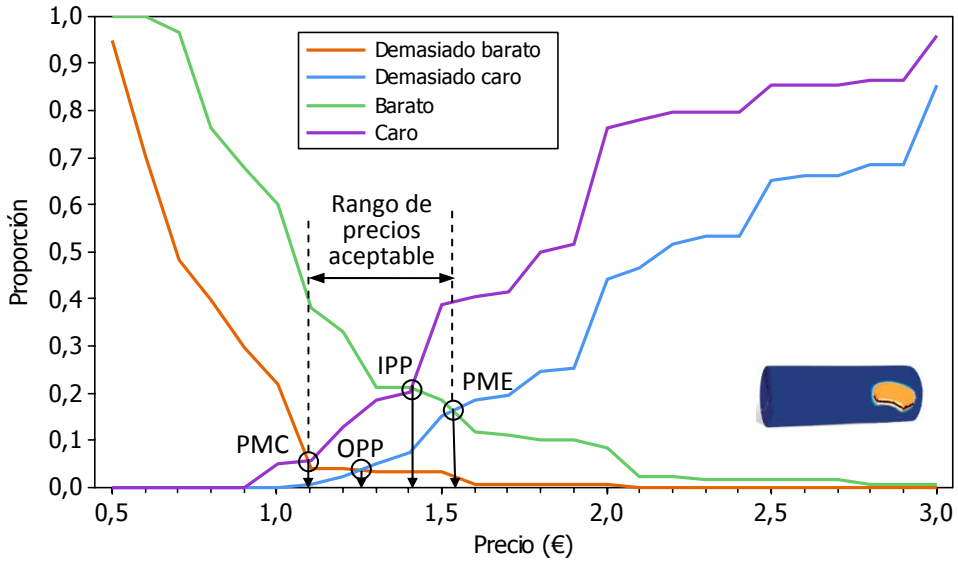


Figura 11: *Curvas y puntos de intersección cuando se utiliza el envase A*

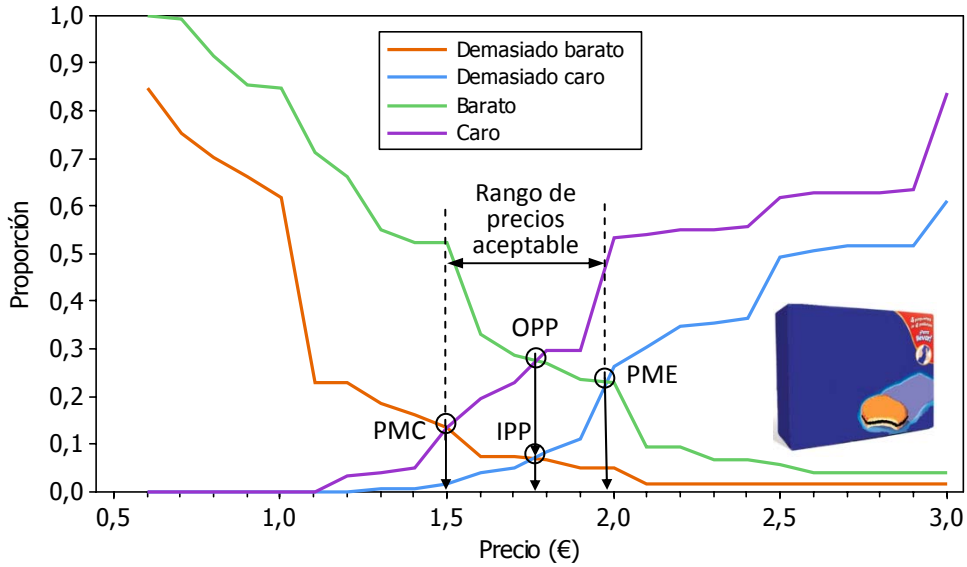


Figura 12: *Curvas y puntos de intersección cuando se utiliza el envase B*

En el caso del envase A, obtenemos un rango de precios aceptable entre 1,09 € y 1,53 €, por tanto, el precio teórico que había pensado el fabricante de 1,19 € está dentro de este rango. El precio óptimo que se obtiene es de 1,25 € que es un poco superior al precio establecido por el fabricante.

En el caso del envase B el rango de precios aceptable está entre 1,50 € y 1,98 €, por lo tanto, el precio de 1,44 € que había pensado el fabricante está por debajo de este rango. El precio óptimo que se obtiene es de 1,77 €, más de 30 céntimos superior a lo que en principio se había establecido. Esta información fue muy valiosa para fijar el precio final del producto.

(Proyecto de la Licenciatura de Estadística, presentado en febrero de 2008 con el título "Estudi d'una nova varietat de galeta des de la perspectiva de concepte, producte, preu i envàs")

Nos dicen que seguro de que hemos probado alguna vez estas galletas, y que seguro de que nos han gustado. No es extraño, lo estudian todo tan bien...

Sara Solanes estudió la Diplomatura de Estadística en la Universidad de Barcelona y después realizó la Licenciatura en la UPC. Ha completado su formación con diversos cursos orientados a profesionales sobre técnicas estadísticas y paquetes de software especializados. Cuando acabó la diplomatura empezó a trabajar en una empresa dedicada a realizar estudios sociológicos y de mercado. Después cambió, dentro del mismo sector, y ahora es coordinadora y técnica del departamento de proceso de datos en IPSOS Operaciones S.A., que pertenece a un grupo internacional líder en estudios comerciales y de opinión. Una de les sus aficiones es cantar en un coro de Gospel (Twocats pel Gospel).

Mejora de la calidad

9

Reducción de defectos en el proceso de fabricación de un componente de automóvil

Proyecto realizado por: **Mª del Mar Costa Vaghi**
Dirigido por: **Alexandre Riba Civil**

La estadística tiene un inmenso campo de aplicación en el mundo industrial, donde siempre es vital tener la mejor información para poder tomar las mejores decisiones. Y para tener la mejor información hay que saber utilizar las técnicas estadísticas más adecuadas en cada caso.

El objetivo de este proyecto se centra al reducir el porcentaje de defectos detectados en las pruebas de control final de un producto. No se trata de evitar que llegue producto defectuoso al cliente, que en este caso no llega, sino de evitar las pérdidas económicas que ocasiona el tener que reparar los productos y repetir las pruebas. Para conseguir esta reducción hace falta identificar cuáles son las causas que provocan el defecto, seleccionar aquéllas en que conviene centrar los esfuerzos, aplicar las técnicas adecuadas para descubrir la raíz de los problemas y plantear los cambios necesarios para solucionarlos.

El proyecto está enmarcado en un programa de mejora "Seis Sigma", metodología que se está consolidando como la más utilizada por las grandes empresas para implantar sus programas de mejora y que tiene en la aplicación de técnicas estadísticas uno de sus pilares fundamentales.

La empresa, el producto y su proceso de producción

La fábrica donde se ha llevado a cabo este proyecto forma parte de una multinacional que fabrica componentes por el sector del automóvil. Sus clientes son los grandes fabricantes de coches como BMW, Ford, Nissan, Opel, Peugeot, Citroën, Renault, Suzuki y Volvo, entre otros. La fábrica está localizada en Cataluña y tiene una plantilla de unos 1.600 trabajadores.

El producto que se fabrica forma parte del motor del coche y está formado por un conjunto de componentes que hay que montar, algunos comprados al exterior y otros producidos en la misma fábrica. El proceso de montaje se realiza dentro de una cámara totalmente aislada en la cual sólo se puede entrar con un uniforme especial, que incluye un gorro como si se tratara de un quirófano. Esto es así para evitar que puedan entrar partículas durante el proceso, por pequeñas que sean, y queden dentro del producto final, cosa que provocaría un mal funcionamiento.

Una vez montado, el producto sale de la cámara y pasa por una serie de controles. Si el producto es rechazado quiere decir que tiene algún tipo de defecto y pasa a la sección de reparaciones, donde intentan arreglarlo y lo devuelven para repetir los controles. Si el producto es aceptado, significa que es bueno y entonces pasa a las fases finales, que simplemente consisten en añadir algunos tapones exigidos por el cliente. De ahí pasa a embalaje y finalmente a expediciones.

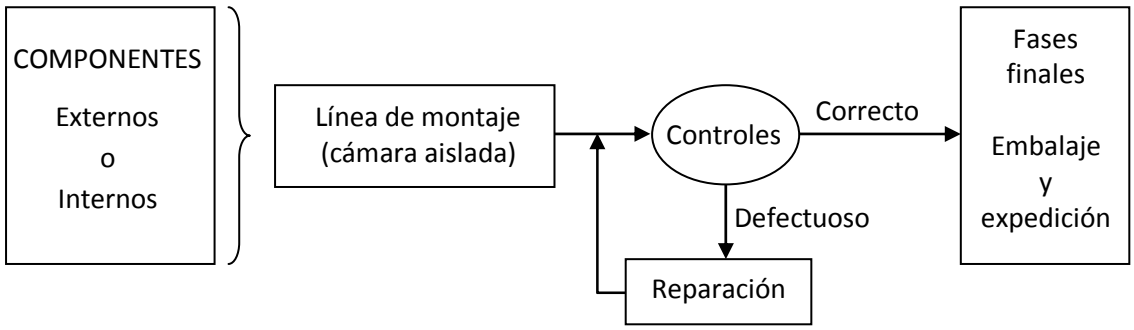


Figura 1: *Esquema del proceso de producción*

El control final consiste en simular que el producto se encuentra funcionando dentro del motor del vehículo. Se realiza introduciendo el producto de forma automática en el banco de test y conectándolo de forma que el resultado es un simulacro del sistema real. A continuación se somete a diferentes condiciones de velocidad y presión que simulan las diferentes situaciones en las que se puede encontrar un vehículo, algunas correspondientes a condiciones normales de funcionamiento y otras a condiciones extremas. Cada una de estas condiciones se conoce como operación.

En cada una de las condiciones en que se prueba se mide el caudal y la depresión, y se mira si el valor obtenido está dentro de los límites de tolerancia definidos. Si lo está, se da el producto como bueno (es decir, funciona); sino quiere decir que el producto tiene algún problema en el caudal y/o la depresión y, por lo tanto, se rechaza.

Estos problemas de rechazo ocasionan importantes pérdidas económicas a la empresa, razón por la cual se decidió poner en marcha este proyecto de mejora.

La metodología de mejora "Seis Sigma"

Las empresas tienen que mejorar continuamente para seguir siendo competitivas. Tienen que adaptarse a las nuevas demandas de los clientes, mejorar su satisfacción y fidelización con los productos o servicios que le suministran. También hay que mejorar para reducir costes haciendo los procesos más eficientes (disminuir defectos, producir menos chatarra, eliminar tiempos muertos,...).

Los programas de mejora sistemática incluyen aspectos organizativos (liderazgo de la dirección, responsabilidad y participación del personal), metodológicos (qué etapas se tienen que seguir) y recomendaciones con respecto a qué herramientas de resolución de problemas hay que utilizar.

Aun así, también hay "mejores maneras de mejorar". Los programas "Seis Sigma" son, desde finales de los años 90, los que se han impuesto, primero en las grandes empresas multinacionales norteamericanas, y después en todo el mundo.

La empresa en qué se ha llevado a cabo este proyecto utiliza la metodología Seis Sigma. Sus aspectos clave son:

- Ligar las mejoras a los valores de unos indicadores que hay que medir de forma objetiva ("mejorar es cambiar el valor de un indicador en la dirección que interesa").
- Cuantificar siempre el impacto económico (ahorros) de las mejoras introducidas. Estos ahorros se tienen que ver reflejados finalmente en la cuenta de resultados de la empresa.
- Dar la máxima relevancia a los datos para tomar decisiones (no se pueden tomar decisiones basadas en intuiciones, suposiciones, impresiones...). Tienen, por tanto, un gran protagonismo las estrategias de recogida y análisis de datos (la estadística, en una palabra).
- La dirección de la organización lidera y se implica en los proyectos de mejora.

- Utilizar una metodología definida, que consta de 5 etapas. Los aspectos fundamentales de cada una de las etapas son:
  - Definir: Establecer los objetivos con los indicadores y las métricas adecuadas. Cuantificar el impacto económico de la mejora.
  - Medir: Obtener los datos necesarios para acotar el problema, diagnosticarlo y orientar su resolución.
  - Analizar: Sacar conclusiones a partir de los datos recogidos.
  - Mejorar: De acuerdo con toda la información obtenida, plantear y poner en marcha las estrategias de mejora.
  - Controlar: Establecer los procedimientos de control adecuados para asegurar que se mantienen los nuevos niveles de calidad obtenidos.

Etapa DEFINIR. Objetivos del proyecto de mejora

Al inicio del proyecto, el porcentaje de rechazo con respecto a la producción total era del 4,34%. El objetivo marcado por la dirección era pasar a un rechazo del 1%.

Impacto económico

El porcentaje de defectos inicial se puede dividir en dos partes: un 2% que no necesita reparación y un 2,34% que sí la necesita. Como se consideró que sería más fácil reducir el porcentaje que no necesita reparación, se planteó que el 1% final estaría compuesto por un 0,25% que no necesitaría reparación y un 0,75% que sí la necesitaría.

Sobre una producción anual prevista de 1 millón de unidades, la reducción sería de 33.400 unidades defectuosas menos (Tabla 1).

Tabla 1: *Reducción del número de defectos*

	Defectos	No necesitan reparación	Necesitan reparación
Situación inicial	43.400	20.000	23.400
Situación final	10.000	2.500	7.500
Reducción	33.400	17.500	15.900

Repetir el test tiene un coste de 0,8€ y hacer la reparación 7,8€. De acuerdo con los valores anteriores, la consecución de los objetivos reducirá en 33.400 el número de tests que se tendrán que repetir y en 15.900 el número de reparaciones. Por tanto, los ahorros serán los indicados en la Tabla 2.

Tabla 2: *Ahorros estimados reduciendo el número de defectos*

	Reducción	Ahorro unitario (€)	Ahorro total (€)
Repetición de tests	33.400	0,8	26.720
Reparaciones	15.900	7,8	124.020
TOTAL			150.740

Alcanzar el objetivo supondrá, por lo tanto, un ahorro anual de 150.740 €.

Etapa MEDIR. Obtención de datos para diagnosticar el problema

En el test final se realizan toda una serie de operaciones, pero los rechazos sólo se producen en 5 de estas operaciones y las causas son siempre debidas a que el caudal o la depresión están fuera de tolerancias. En una de las cinco operaciones sólo puede estar fuera de tolerancias el caudal y en otra sólo la depresión (Tabla 3), en las otras tres lo puede estar cualquiera de las dos.

Tabla 3: *Operaciones que pueden provocar rechazo en el test final*

Código operación	Prueba	Posible causa de rechazo	
15	Funcionamiento a velocidad = 2.000 y depresión = 1.600	Caudal	Depresión
16	Funcionamiento a velocidad = 1.000 y depresión = 1.200	Caudal	Depresión
17	Funcionamiento a velocidad = 400 y depresión = 230	Caudal	Depresión
18	Funcionamiento a velocidad = 50 y depresión = 230	Caudal	---
200	Conexión automática de la pieza al banco de pruebas	---	Depresión

Se tienen, por tanto, 8 posibles causas de rechazo, pero intentar atacarlas todas desde el principio no acostumbra a ser una buena estrategia. Lo que se hizo fue identificar cuáles eran las pocas causas que provocaban la mayor parte del problema (principio de Pareto). Para hacer este estudio se tomaron datos durante 6 semanas (periodo lo bastante largo para ser representativo del funcionamiento general) y a partir de estos datos se construyeron los diagramas de Pareto de la Figura 2. Con respecto a los defectos por caudal, la operación con mayor porcentaje de rechazo era la 17. Con respecto a los defectos por depresión, la operación con más porcentaje de rechazo era la 200, que suponía el 67% del rechazo total. Si tenemos en cuenta también la operación 17, que ha salido la más importante en el rechazo de caudal, entonces estaremos teniendo en cuenta el 87% del rechazo total por depresión.

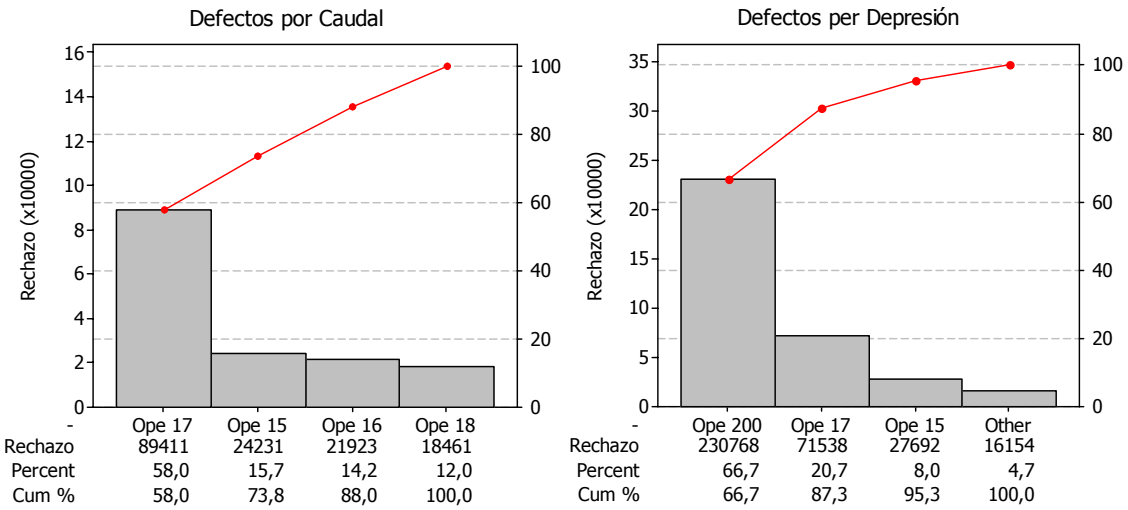


Figura 2: *Diagramas de Pareto del número de defectos por caudal y por depresión, según la operación en que se han detectado*

Por tanto, las operaciones en las que había que centrarse en primer lugar eran las que provocaban un mayor rechazo:

- La operación 200 comprueba la conexión física que existe entre la pieza D (componente del producto) y el banco de test. Si la conexión es incorrecta, el producto es rechazado por código 33 (depresión fuera de tolerancias). Este error es un error físico del banco de test y no significa que el producto sea defectuoso. Al rechazo detectado en la operación 200 le llamaremos rechazo por mala conexión.

- La operación 17 somete el producto a una velocidad de 400 rpm y una presión de 230 bares. Cuando se estabiliza se toman medidas del caudal y de la depresión. Las dos medidas tienen unos límites de tolerancia asignados, si la medida está dentro de los límites se acepta el producto como bueno, mientras que si la medida del caudal o de la depresión está fuera de límites se rechaza el producto.

Así pues, se centrarán las acciones en tres respuestas. La respuesta Y_1 mide el rechazo por mala conexión (2% de rechazo con respecto a la producción total), la Y_2 el rechazo por caudal en la operación 17 (0,78%) y la Y_3 el rechazo por depresión en la operación 17 (0,62%). Las tres medidas juntas representan el 3,4% de rechazo respecto de la producción total.

Situación de partida de la respuesta Y_1: Rechazo por mala conexión

Cuando un producto entra en el banco de test, lo primero que hace es conectarse automáticamente. Una de estas conexiones se realiza a través de la pieza D. Antes de

empezar lo que sería propiamente el test de funcionalidad del producto se comprueba que estas conexiones están bien hechas. Cuando la conexión entre la pieza D y el banco de test no es correcta, la depresión que se crea dentro de esta pieza es demasiado pequeña y el producto se rechaza por mala conexión. En este tipo de rechazo, el producto no es defectuoso, el problema se encuentra sólo en la conexión con el banco. La pieza que provoca que la conexión sea mala es la junta de conexión, la cual se debe cambiar cada 24 horas.

El primer paso, por tanto, es ver si existe alguna relación entre el desgaste de esta junta y los rechazos por mala conexión, pero no existía ningún registro que permitiera saber cuándo se había realizado el cambio de junta. Por eso, se añadió un formato al banco de test donde el operario anotaba cada vez que realizaba un cambio de la junta.

La Figura 3 muestra que la mayoría de los rechazos no son casos aislados sino que de golpe se empiezan a rechazar muchos productos por mala conexión. Cuando pasa esto es debido a que la junta de conexión se ha roto. Pero, además, este gráfico también nos permite ver que la causa de que esta junta se rompa no es el desgaste, ya que encontramos una crisis importante de rechazo cuando se acaba de cambiar la junta.

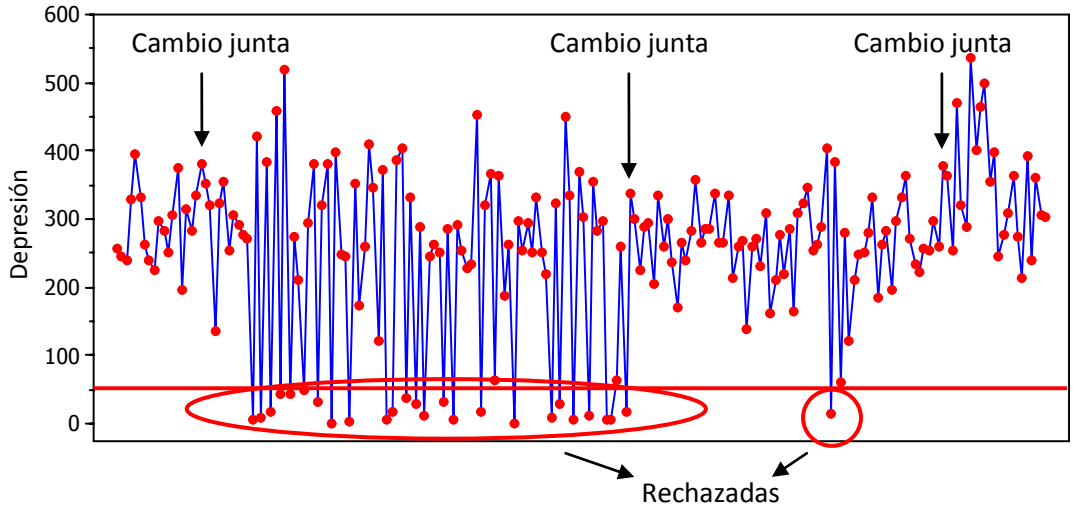


Figura 3: *Evolución de la depresión en la operación 200 y cambio de la junta de conexión*

Se llega a la conclusión de que la causa del rechazo por mala conexión es la rotura de la junta de conexión y eso provoca crisis de rechazo, pero el hecho de que esta junta se rompa no es debido al desgaste. Por tanto, controlar que se cambien las juntas periódicamente no garantiza que no se produzcan crisis.

Situación de partida de la respuesta Y_2: Rechazo por caudal en la operación 17

Hacía falta conocer cuál era la situación de los valores del caudal en esta operación, así que durante una semana se tomaron datos para hacer un estudio de capacidad.

La Figura 4 nos permite ver que la campana no está centrada dentro de las tolerancias. El índice de capacidad *Cp* es una comparación de la variabilidad con la que se produce ("anchura de la campana") con las tolerancias del producto (cómo tenemos que producir en realidad), de manera que nos muestra la posibilidad de producir dentro de tolerancias cuando el proceso está centrado (produce con media en el valor que nos interesa). Con respecto a este índice de capacidad, el *Cp* obtenido de 1,47 indica que el proceso es capaz de cumplir con las especificaciones. Ahora bien, esto no nos garantiza que estemos produciendo bien, porque puede pasar de que el proceso esté descentrado. Para comprobar este hecho se define el *Cpk*. Si el proceso está centrado en el valor nominal el *Cpk* coincide con el *Cp*; cuanto más se desvía el valor del *Cpk* del que tiene el *Cp*, más descentrado está el proceso. El *Cpk* obtenido de 0,64 se aleja mucho del 1,47 del *Cp* lo cual es un indicador de que el proceso está descentrado.

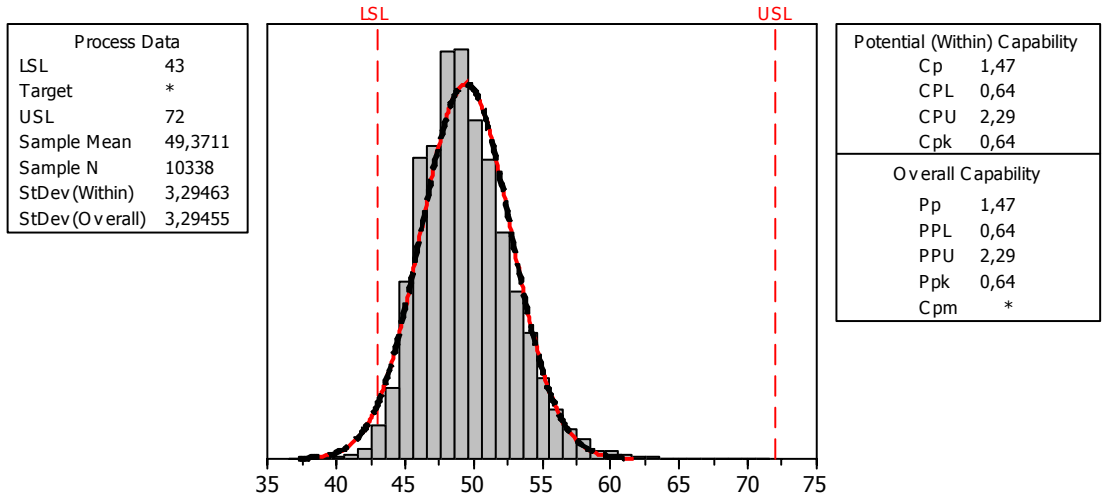


Figura 4: *Estudio de capacidad para el caudal en la operación 17*

Está claro que centrando el proceso reduciremos considerablemente el rechazo, pero esto no es sencillo porque el caudal no es una característica que se pueda manipular ajustando una máquina sino que es una consecuencia de muchas variables del producto.

Situación de partida de la respuesta Y_3: Rechazo por depresión en la operación 17

En este caso sólo se tiene límite de tolerancia superior y por tanto el *Cp* no se puede calcular. El valor del *Cpk* de 0,72 nos indica que el proceso está demasiado cerca de la tolerancia superior.

La depresión tampoco es una característica que se pueda manipular mediante un ajuste sino que también es una consecuencia de otras variables, concretamente es una consecuencia del caudal. Esta relación entre el caudal y la depresión ya se conoce, pero será interesante verificarla y cuantificarla.

Etapa ANALIZAR (1): Mejora de la conexión al banco de test

Para solucionar definitivamente los problemas de conexión al banco de test había que identificar las causas que provocaban la rotura de las juntas. Pero mientras no se identificaban estas causas, había que evitar que cuando una junta se rompiera, el banco de test siguiera funcionando y rechazando por mala conexión hasta que un operario se diera cuenta de ello. Así pues, paralelamente a la busca de las causas raíz se implantó un control estadístico del proceso para identificar lo más rápidamente posible la aparición de rachas de rechazo. La pieza que hacía la conexión entre la pieza D y el banco de test es la siguiente:

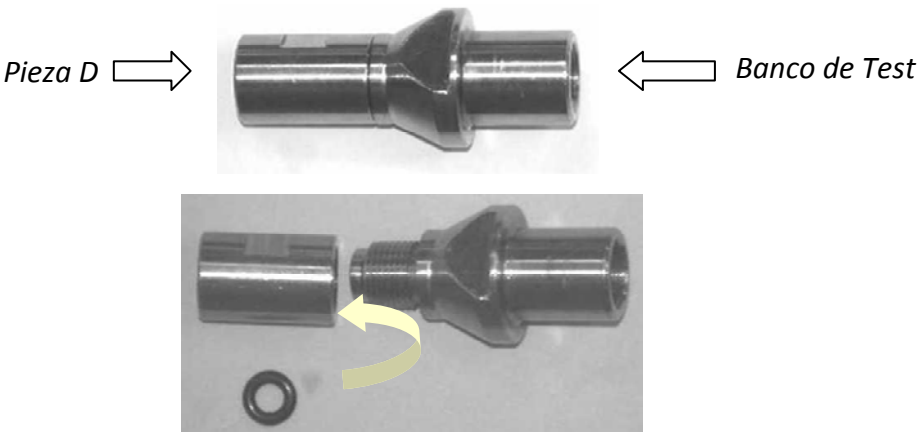


Figura 5: *Pieza de conexión, con el detalle de la situación de la junta*

El análisis detallado de las juntas de conexión puso de manifiesto que las juntas que se rompían siempre tenían el diámetro interior más pequeño. Esto provocaba que al entrar la pieza D el material de la junta se tuviera que estirar más de lo que inicialmente estaba previsto y se debilitara, de forma que el movimiento que hace la pieza D al entrar, rozando ese material más débil, provocaba la rotura.

Para solucionar este problema se debía reducir la variabilidad del diámetro interior de forma que no se encontraran diámetros pequeños, pero estas juntas se compraban a un proveedor que no era capaz de reducir esta variabilidad. La única alternativa era realizar un nuevo diseño para la conexión entre la pieza D y el banco de test. Por tanto, se sustituyó la actual junta de 'vitón' con forma de aro, por una nueva pieza de 'teflón', un material mucho más resistente, con forma de tronco de cono. De esta manera, la pieza D entraba por el lado más abierto del cono y se deslizaba por el interior de la nueva pieza hasta el lado más cerrado sin tener que hacer ningún movimiento brusco.

De acuerdo con el nuevo diseño se construyeron dos piezas piloto y se montaron cada una en un banco de test. Durante una semana se controlaron estos dos bancos. Ninguno de los dos rechazó ni un solo producto por mala conexión de la pieza D. Posteriormente, se hicieron cuatro piezas más y se montaron en cuatro bancos, siguiendo con el resultado de cero rechazos por mala conexión. Por tanto, la acción de mejora prevista para reducir los rechazos por mala conexión fue cambiar el sistema de conexión de la pieza D en todos los bancos de test.

Etapa ANALIZAR (2): Relación entre el caudal y la depresión

La variable Y_3, depresión, se refiere a la depresión que se crea dentro de lo que se ha denominado pieza D. Para crearla se aprovecha el caudal de Y_2, por lo que es lógico pensar que la depresión y el caudal estarán relacionadas. Ambas respuestas están directamente ligadas con nuestros objetivos.

Para estudiar la forma de esta relación se utilizaron los datos correspondientes a una semana de producción, durante la cual se produjeron 8.681 unidades. Con la ayuda del siguiente gráfico podemos hacernos una idea del tipo de relación que existe.

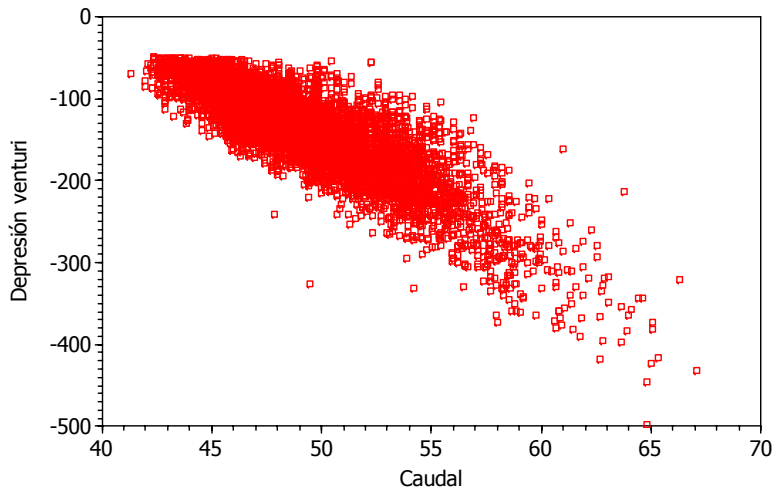


Figura 6: *Relación entre Depresión y Caudal*

Se ve que, efectivamente, cuanto más caudal hay, menor es la depresión. En el gráfico ya se intuye la presencia de curvatura en la nube de puntos, como si la relación fuera cuadrática. Después de estudiar diferentes posibilidades, se llegó a la conclusión que la mejor manera de explicar la depresión creada en la pieza D a partir del caudal era mediante el modelo:

$$\text{Depresión} = -136 + 11{,}7\ \text{Caudal} - 0{,}237\ \text{Caudal}^2$$

que explica el 71,6% de la variación total de la depresión.

Etapa ANALIZAR (3): Diseño de experimentos

Nos interesa aumentar el caudal (acercándolo todo lo posible a su valor nominal de 57,5 litros/hora) y disminuir la depresión (cuanto más pequeña sea mejor). Por lo que se ha podido ver, la relación que existe entre la depresión y el caudal nos beneficia, ya que aumentando el caudal disminuye la depresión. Ahora bien, aumentar el caudal o disminuir la depresión es más complicado de lo qué puede parecer a primera vista.

Es necesario identificar cuáles son las variables que pueden tener algún efecto sobre el caudal (y, por lo tanto, también sobre la depresión). Con la colaboración de los ingenieros del producto se consideró que éstas eran: el diámetro del agujero L, el juego entre las piezas P y el juego entre las piezas E.

Para ver cómo estas variables afectaban a la respuesta era necesario hacer pruebas, pero cada prueba era muy costosa ya que implicaba fabricar piezas fuera de la producción normal, controlar el flujo de estas piezas y colocarlas todas de una forma especial en la cadena de montaje. Por tanto, había que utilizar una estrategia de experimentación que proporcionara la máxima información con el mínimo número de pruebas. Por este motivo se decidió utilizar el diseño de experimentos, más concretamente se usó un diseño factorial 2^3. A cada factor se le dieron dos valores diferentes (horquillando los valores actuales) y se hicieron las pruebas en todas las combinaciones de valores de los factores, en este caso 8.

Tabla 4: *Factores y niveles de experimentación (valores para las pruebas)*

Factor	Descripción	Valores (mm)		
		Actual	Pruebas	
		(0)	(–1)	(+1)
L	Orificio de lubricación de la pieza por donde pasa el caudal	0,72	0,69	0,75
P	Distancia entre dos piezas que van insertadas en el interior	0,025	0,015	0,035
E	Holgura entre unas piezas y el agujero en que van encajadas	0,003	0,002	0,005

Se montaron 8 productos diferentes y cada uno se sometió al test funcional que proporcionaba los valores de caudal y depresión. El orden en que los productos pasaron el test funcional fue aleatorizado para evitar efectos que no somos capaces de controlar. Después de realizar los 8 experimentos, los resultados obtenidos estan representados en la Figura 7, donde los vértices de los cubos representan los valores de las respuestas en cada condición de experimentación. Estas dos representaciones sirven para hacernos una primera idea visual de las conclusiones a las que se llegará después de analizar los resultados de forma más detallada.

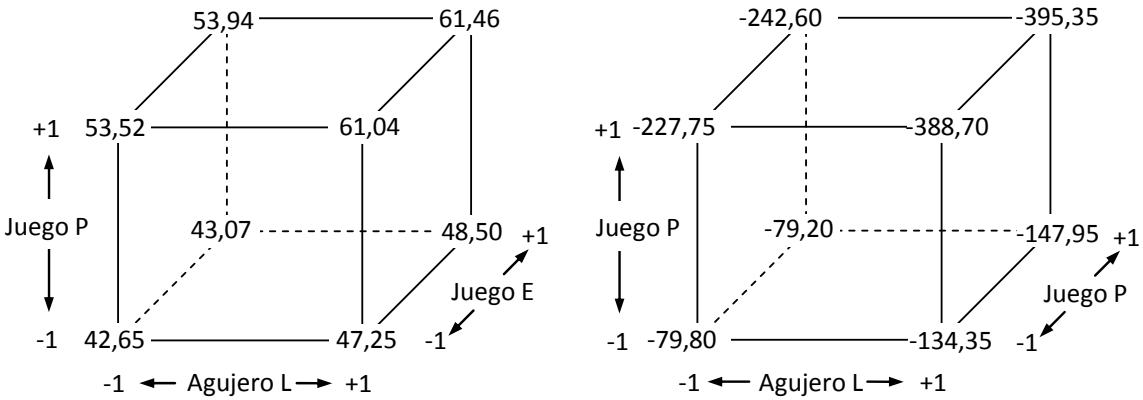


Figura 7: *Representación gráfica de los resultados del caudal (izquierda) y de la depresión (derecha)*

El estudio detallado de la influencia de los factores sobre la respuesta (cálculo de los efectos y análisis de su significación estadística) permitió sacar las siguientes conclusiones:

- El juego E, que de entrada se había tenido en cuenta, no tiene ningún efecto en las respuestas analizadas (en el rango de valores que se ha estudiado)
- El juego P y el diámetro del agujero L sí que afectan y, además, interaccionan entre ellos; es decir, la influencia de un factor depende del valor que toma el otro.

Para llegar al valor objetivo del caudal (57,5 litros/hora), vimos que interesa trabajar al nivel alto (+1) del diámetro del agujero L y a un punto medio entre el valor actual (0) y el nivel alto (+1) del juego P. Con respecto a la depresión (tiene que ser la mínima posible), interesa trabajar al nivel +1 de los dos factores. En la práctica, modificar el juego P es bastante complicado, ya que implica muchos cambios en el proceso, mientras que cambiar el diámetro del agujero L es mucho más sencillo porque sólo supone cambiar la broca con la cual se realiza el mecanizado de ese agujero. Por este

motivo se modificó sólo el diámetro del agujero L, que, por lo que hemos visto a lo largo del análisis conviene ponerlo a nivel alto, lo cual corresponde a 0,75 mm.

Trabajando al nivel +1 del diámetro del agujero L y al nivel 0 del juego P, esperaremos obtener un caudal de unos 57,48 litros/hora y una depresión de unos -313,6 mbar, condiciones lo bastante próximas al objetivo.

Etapa MEJORAR

Cambio en el sistema de conexión de la pieza D

Antes de empezar a cambiar todas las conexiones se realizaron dos piezas piloto y se montaron cada una en un banco de test diferente. Durante una semana se controlaron estos dos bancos y ninguno rechazó productos por mala conexión. Al introducir estas nuevas piezas en cuatro bancos diferentes se observó que el resultado era igual de bueno que en la prueba piloto. Por tanto, a la vista de estos resultados, se dio el diseño como bueno y se introdujeron las nuevas piezas en todos los bancos de test.

Durante las semanas 43 a la 46 (20/10/03 al 14/11/03) se cambiaron las juntas de conexión a medida que se hacían. A partir de la semana 46 (15/11/03), los 32 bancos de test ya tenían la nueva junta de conexión instalada. Ver Figura 8.

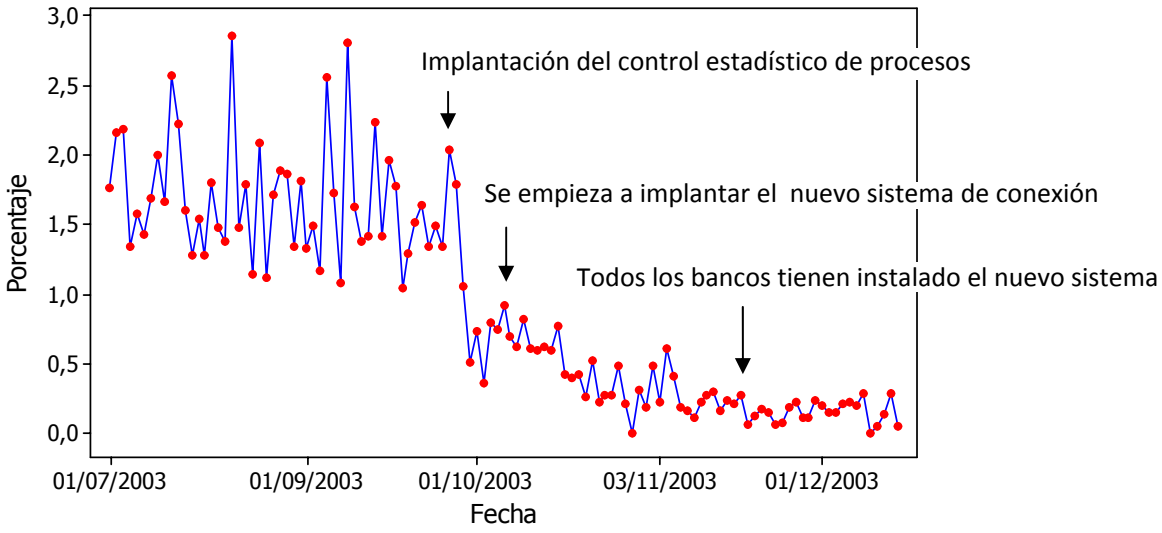


Figura 8: *Evolución del rechazo por mala conexión en el período julio-diciembre de 2003*

Cambio en el mecanizado del agujero L

Antes del cambio de broca definitivo, se fabricó un lote de productos mecanizados con la broca de 0,75 para comparar los resultados obtenidos con un lote de productos fabricados en las condiciones anteriores. Un estudio de capacidad de cada lote permitió ver cómo se distribuyen los datos entre las tolerancias con cada diámetro de broca (ver Tabla 5).

Tabla 5: *Valores del caudal y de la depresión antes y después del cambio de broca*

	Antes: Ø = 0,72 mm			**Después: Ø = 0,75 mm**	
	Caudal	Depresión	→	Caudal	Depresión
Media	52,49	-186,26		53,90	-212,10
Desviación tipo	3,28	51,13		2,95	52,73
Tamaño de la muestra	2964	2964		3422	3422
Cp	1,47	--	MEJORA	1,64	--
Cpk	**0,96**	**0,89**	→	**1,23**	**1,02**

Confirmados los buenos resultados previstos, se decidió realizar el cambio definitivo de la broca en el proceso de producción.

¿Objetivo alcanzado?

Se pasó del 4,34% de rechazo inicial a un 0,7%, lo cual supone una reducción del 83,8% de la producción defectuosa y, por tanto, se superó el objetivo fijado de llegar al 1% (que suponía la reducción del 77% de la producción defectuosa).

Tabla 6: *Resumen de resultados antes y después de las mejoras introducidas*

Concepto	Antes	Después
% Rechazo per mala conexión	2%	0,20%
% Rechazo por caudal	1,34%	0,30%
% Rechazo por depresión	1%	0,20%
TOTAL	**4,34%**	**0,70%**

Las acciones de mejora permitieron alcanzar sobradamente el objetivo fijado. Evidentemente, consideramos los posibles efectos colaterales: el cambio en la junta de conexión no supuso ninguna alteración en el funcionamiento normal del banco ni en el test funcional, y el cambio en el diámetro del agujero L tampoco supuso ningún inconveniente para el correcto funcionamiento del producto ni hizo aumentar ningún otro tipo de rechazo.

Etapa CONTROLAR

Llegados a este punto, tan importante como todo lo que se había hecho hasta el momento era asegurar que el proceso no daría ningún paso atrás. Así pues, había que tener controlados los porcentajes de rechazo, de forma que, en el caso que alguno de ellos incrementara, se pudiera detectar rápidamente y actuar en consecuencia.

Para hacerlo, se creó un programa que tomaba los datos directamente de los bancos de pruebas, donde quedan almacenados los resultados del test funcional. Esta información se actualiza cada 5 minutos, que quiere decir prácticamente en cada test, ya que éste tiene una duración media de 4 minutos.

El programa tiene implantados unos límites de control y da señales de alarma si los resultados que se obtienen hacen sospechar que el proceso puede estar fuera de control. También permite que, desde cualquier ordenador y a cualquier hora del día, las personas interesadas puedan consultar el estado de un rechazo determinado y ver si se alejan del valor objetivo alcanzado.

(Proyecto de la Licenciatura de Estadística, presentado en febrero de 2004 con el título "Reducció del rebuig obtingut en el test de final de línia d'un procés industrial mitjançant la metodologia Sis Sigma de resolució de problemes")

Los expertos coinciden en que la mejora continua basada en la selección de proyectos concretos, el seguimiento de una metodología ordenada, el apoyo de la dirección (tiempo, recursos, atención ...) y la toma de decisiones basada en datos (usando la estadística) son un elemento clave para mantener la competitividad. Pero desde el punto de vista del responsable del proyecto, siendo la estadística una herramienta fundamental, no es la única que hay que saber aplicar. Constancia, habilidades de comunicación y de liderazgo o capacidad para el trabajo en equipo son también fundamentales para conseguir los objetivos.

Mar Costa estudió la Diplomatura y la Licenciatura de Estadística en la UPC. Posteriormente completó su formación con un curso de postgrado sobre Programas de Mejora Seis Sigma en la Fundació UPC, dos cursos sobre metodología Shainin de resolución de problemas y uno sobre Ingeniería de Calidad. Desde que acabó los estudios, su actividad profesional ha estado vinculada a empresas industriales, especialmente en temas relacionados con el control y la mejora de la calidad. Actualmente trabaja en el departamento de calidad de una importante empresa del sector de la automoción, como responsable de proyectos de mejora Seis Sigma y Shainin.

Nuevos medicamentos y mejora del sistema sanitario

10

Análisis de la eficacia de un nuevo fármaco contra el SIDA

Proyecto realizado por: **Raquel López Blázquez**
Dirigido por: **Guadalupe Gómez Melis** y **Núria Porta Bleda**

El SIDA es una enfermedad relativamente nueva. Se empezó a diagnosticar en el año 1981 y rápidamente se convirtió en una de las más temidas y de las que causa más muertos, especialmente en el África subsahariana. En los países de nuestro entorno se ha conseguido estabilizar e incluso reducir su propagación con medidas de prevención como el sexo seguro o las jeringuillas de un solo uso. Pero es necesario ir con cuidado porque cualquier relajación provoca inmediatamente un repunte del número de infectados.

Por otra parte, se está realizando una intensa investigación con el fin de obtener tratamientos que mejoren la esperanza y la calidad de vida de las personas infectadas. Esta investigación ha dado resultados muy importantes y hoy en día las personas diagnosticadas a tiempo pueden llevar una vida prácticamente normal.

En nuestro país uno de los centros más activos en la investigación contra esta enfermedad es la "Fundación para la Lucha Contra el SIDA". (www.flsida.org). En el marco de sus programas de investigación se realizó este proyecto, que tiene como objetivo evaluar la evolución de la carga viral de pacientes infectados al añadir al tratamiento convencional un nuevo fármaco: el Enfurvitide. Se ha demostrado que este fármaco es eficaz para reducir rápidamente la carga viral en pacientes tratados con antirretrovirales. En este proyecto se estudió su efectividad en pacientes que todavía no se habían medicado con antirretrovirales.

Contexto: el virus y la enfermedad del SIDA

La enfermedad del SIDA es causada por el Virus de la Inmunodeficiencia Humana (VIH), que es del tipo lentivirus, lo cual significa que puede permanecer mucho de tiempo en estado latente y, por tanto, puede pasar un largo intervalo de tiempo desde que se produce la infección hasta que se presentan síntomas serios. El VIH destruye las células inmunológicas llamadas CD4 y esto puede provocar que diversas infecciones y cánceres entren en el cuerpo sin defensa. A este tipo de infecciones y cánceres se les denomina enfermedades o infecciones oportunistas.

La carga viral –cantidad de virus existente– funciona como un indicador del avance y pronóstico de la enfermedad, mientras que la cantidad de células CD4 es un indicador de cuánto daño ha causado ya el VIH.

La variabilidad en la celeridad de la progresión de la infección por el VIH se ha explicado en diversas ocasiones haciendo alusión a la metáfora de un tren. En ésta, la infección VIH es el propio tren, la carga viral es la velocidad que lleva el tren y la longitud de los raíles que forman el trayecto son los linfocitos CD4 que tiene el paciente. La estación final, hacia la que el tren avanza, es el desarrollo del SIDA, pero necesitará un tiempo para recorrer el trayecto. Este tiempo dependerá de la cantidad de virus (velocidad) y, por tanto, cuanto más baja sea más tiempo tardará en recorrer el trayecto; si la carga viral es indetectable la velocidad del tren será prácticamente nula pero el tren nunca se para del todo, por lo que hasta hoy se conoce, como si existiera una ligera pendiente hacia abajo y no se llegara a frenar del todo.

Cuando la carga viral sea constante, el tiempo dependerá de la distancia que se tenga que recorrer. Si el sistema inmunitario es estable, el trayecto será muy largo y le costará más tiempo llegar al destino; por el contrario, conforme la cifra de CD4 caiga, el recorrido será menor y, a igual velocidad, el tiempo empleado en recorrerlo será menor.

El VIH está presente en todos los fluidos de la persona infectada, tanto internos como externos, pero solamente algunos de ellos tienen capacidad de infección. La infección sólo se puede producir cuando una cantidad suficiente de virus que se encuentra en la sangre, el semen, las secreciones vaginales y la leche materna de las personas afectadas penetra en la sangre mediante heridas, pinchazos, lesiones en la piel, en la mucosa vaginal, en la mucosa anal o en la mucosa bucal. El virus VIH sobrevive poco tiempo fuera del organismo, por eso tiene que penetrar en el torrente sanguíneo de la persona expuesta. Además, esta transmisión necesita una cantidad mínima para

provocar la infección. Por debajo de este umbral el organismo consigue liberarse del virus e impide que se instale.

El VIH es un virus de forma esférica, de unos 80-100 nm de diámetro. La información genética está compuesta por dos cadenas de ARN (ácido ribonucleico) que contienen los diferentes genes que permiten producir las proteínas virales.

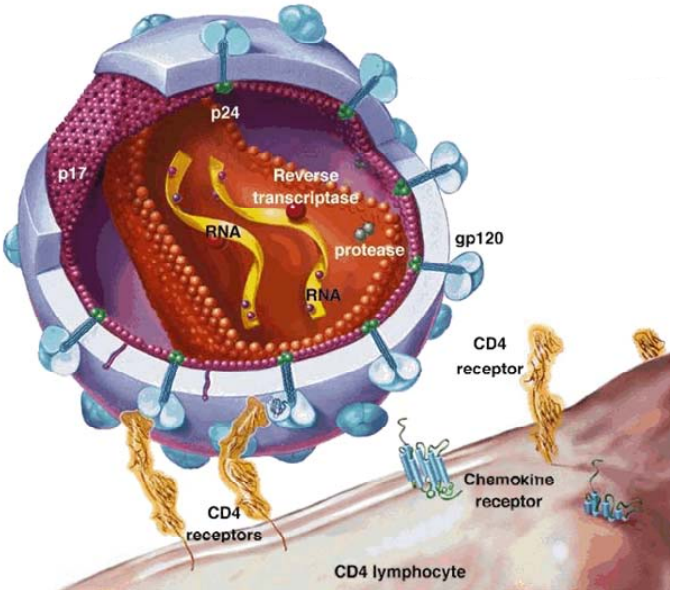


Figura 1: *Estructura del VIH*

La intensa investigación en el terreno del VIH nos ha dotado de una buena colección de fármacos que son capaces de bloquear el ciclo vital del virus. Lo más habitual es utilizar una combinación de tres fármacos que se conoce como tratamiento HAART (que son las siglas en inglés de Terapia Antirretroviral Altamente Activa). Cuando un paciente no ha iniciado ningún tratamiento antirretroviral se dice que es *naïve*.

El Enfurvitide es un fármaco que bloquea la entrada de los virus en las células huésped uniéndose a una proteína de la superficie del VIH llamada GP41. Una vez lo hace, puede adherirse a la superficie de las células CD4 evitando que el virus infecte las células sanas.

Se ha demostrado que el Enfurvitide tiene una gran potencia antirretroviral y una buena tolerancia por parte de los pacientes. Estudios ya realizados han mostrado que añadir el Enfurvitide al tratamiento óptimo (HAART) de pacientes muy tratados con antirretrovirals hace bajar su carga viral muy rápidamente. Pero este hallazgo puede no ser extrapolable a pacientes *naïves*.

Objetivo del proyecto

Este proyecto parte de los datos y de un estudio previo[1] que se realizó para valorar si el Enfurvitide aumentaba la actividad antiviral de una combinación de cuatro antirretrovirales (ARV) en pacientes ARV-*naïves* e infectados por el VIH.

En este primer estudio se estimaron las tasas individuales de decrecimiento de las cargas virales durante la primera fase, del día 1 al día 6, a partir de:

- CV_t: carga viral en el instante de tiempo t
- CV_0: carga viral en el momento inicial
- k: constante relativa a la tasa de decrecimiento

Mediante la siguiente expresión:

$$CV_t = CV_0 \cdot e^{-kt}$$

El estudio plantea comparar la tasa de decrecimiento k entre el grupo de los que habían tomado el Enfurvitide y el grupo de los que no. La metodología utilizada en este primer estudio no tenía en cuenta la dependencia entre las diferentes medidas de un mismo paciente.

Este proyecto pretende mejorar el análisis y, con este objetivo, se ha aplicado una nueva metodología (modelos de efectos mixtos) adecuada para el tipo de datos que se tienen: la misma información (la carga viral) para cada uno de los individuos recogida en diferentes instantes a lo largo de un cierto periodo de tiempo. Además, se tienen recopilados datos complementarios que se incorporarán a los modelos con el fin de explicar la dinámica viral de una forma más detallada.

Diseño del estudio y recogida de datos

Los datos que se han utilizado en este proyecto pertenecen a un estudio piloto, abierto, prospectivo (se diseña y realiza en el presente, pero los datos se analizan cuando ha transcurrido un determinado tiempo, en el futuro) y aleatorizado realizado en la Fundación Lucha contra el Sida del Hospital Universitario *Germans Trias i Pujol*.

[1] J. Moltó et al. (2006): "*Increased antiretroviral potency by the addition of enfurvitide to a four-drug regimen in antiretroviral naive, HIV-infected patients*". Antiviral Therapy; **11**: 47-51

Se estableció que los pacientes incluidos en el estudio debían satisfacer los siguientes criterios (criterios de inclusión):

- Tener más de 18 años y el test VIH positivo,
- Tener una carga viral de alrededor de 10.000 copias/ml,
- Recuento de células CD4 alrededor de 100 células/mm^3,
- Capacidad para seguir el periodo del tratamiento,
- En caso de ser mujer, no ser potencialmente fértil (definido como post menopáusica al menos desde hace 1 año o esterilizada quirúrgicamente) o haberse comprometido a utilizar un método anticonceptivo de barrera durante el estudio,
- Haber aceptado y firmado un informe de consentimiento.

Por otra parte, para poder entrar en el estudio, no debían tener ninguna de las siguientes características (criterios de exclusión):

- Exposición a algún tratamiento antirretroviral previo,
- Alergias conocidas a alguno de los fármacos de tratamiento o similares,
- Sospecha de la no-adherencia al tratamiento,
- Incremento del AST/ALT (relación de los valores de dos tipos de transaminasas) más de 5 veces el límite superior de normalidad,
- Ser mujer y estar embarazada o en periodo de lactancia,
- Presencia de infecciones oportunistas o tumores en los tres meses anteriores a la inclusión.

Después de ser incluidos en el estudio, los pacientes eran asignados aleatoriamente a uno de los dos grupos: el grupo de control (tratamiento habitual) y el grupo enfurvitide (posible mejora).

Tabla 1: *Medicamentos tomados por el grupo tratado y por el grupo de control*

Grupo	Medicamento	Posología
CONTROL	Lamivudine (Epivir®)	1 pastilla (300mg) / 24h
	Tenofovir (Viread®)	1 pastilla (300mg) / 24h
	Efavirenz (Sustiva®)	1 pastilla (600mg) / 24h
	Lopinavir/ritonavir (Kaletra®)	4 pastillas (533/133mg) / 12h (4 primeras semanas)
ENFURVITIDE	Lamivudine (Epivir®)	1 pastilla (300mg) / 24h
	Tenofovir (Viread®)	1 pastilla (300mg) / 24h
	Efavirenz (Sustiva®)	1 pastilla (600mg) / 24h
	Lopinavir/ritonavir (Kaletra®)	4 pastillas (533/133mg) / 12h (4 primeras semanas)
	Enfurvitide T-20 (Fuzeon®)	90mg / 12h, subcutáneo (4 primeras semanas)

Como se puede ver en el listado de fármacos (Tabla 1), en la semana 4 se realizó una simplificación del tratamiento, igual para los dos grupos. Por este motivo, los datos que se han tenido en cuenta para los sucesivos análisis son solo hasta este cambio de tratamiento (los 30 primeros días). Las observaciones que se tienen de tiempos posteriores no aportan información adicional, ya que pasado este periodo todos los pacientes pasan a seguir el mismo tratamiento.

La asignación de los pacientes a un grupo u otro (control o Enfurvitide) se hizo aleatoriamente, de forma que se puede asumir que los individuos de ambos grupos tienen características similares. Si no se hubiera hecho de esta manera (por ejemplo, si se hubieran asignado los más jóvenes a un grupo) y se detectaran diferencias entre grupos, nunca podríamos estar seguros de si estas diferencias son debidas al tratamiento que estamos probando o a las diferentes características de los pacientes de cada grupo. Aleatorizando la asignación de pacientes, si los resultados son diferentes, esta diferencia sólo se puede atribuir a la incorporación del nuevo fármaco, ya que todas las otras variables que pueden influir lo hacen de la misma forma.

La respuesta analizada es la carga viral de los dos grupos, y se evaluará en diferentes instantes de tiempo.

Descripción de la muestra

La recogida de datos ha tenido una duración de 2 años, comprendiendo el periodo entre el 29 de septiembre de 2003 y el 7 de marzo de 2005. En el estudio ha participado un total de 15 pacientes a los cuales se les ha realizado un seguimiento de 4,28 semanas en mediana. En el grupo de control han quedado incluidos 7 individuos y 8 en el grupo Enfurvitide. El número de participantes es reducido ya que, como ya se ha comentado, se trata de un estudio piloto.

Se han recogido las siguientes variables socio-demográficas y clínicas en el momento basal (tiempo cero): sexo (Hombre/Mujer), coinfección con virus hepatitis C (Sí/No), recuento células CD4 (cels/mm3), carga viral (copias/ml), edad (años) y tiempo desde la infección por VIH (meses).

Realizando los tests correspondientes se desprende que no hay diferencias estadísticamente significativas entre los grupos en estudio por lo que queda validada su homogeneidad a nivel basal. Es decir, podemos suponer que la tipología de individuos es similar respecto de las características socio-demográficas y clínicas en el momento basal en los dos grupos. Sin embargo, si la aleatorización ha sido realizada correctamente, esta homogeneidad ya tiene que quedar garantizada.

Análisis descriptivo de la dinámica viral

El perfil que describe la dinámica viral, o su farmacodinámica, acostumbra a tener una primera fase de crecimiento, es decir, un periodo de tiempo en el cual el tratamiento todavía no ha hecho su efecto y la carga viral sigue incrementándose. Acto seguido, aproximadamente un día después del inicio del tratamiento, empieza una primera fase de decrecimiento muy rápida que dura más o menos una semana. Finalmente, hay una segunda fase de decaimiento bastante más lenta hasta que se llega a una carga viral indetectable (por debajo de las 50 copias por ml). Eso lo podemos reflejar, de una forma lineal y muy simplificada, en el siguiente diagrama (Figura 2):

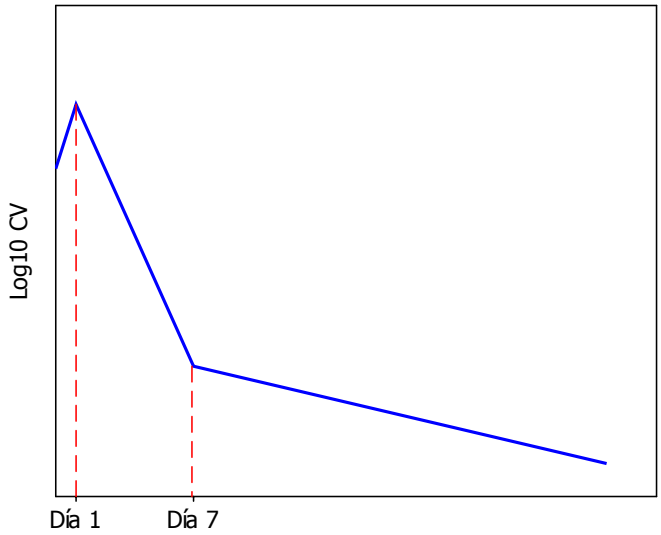


Figura 2: *Esquema de la dinámica viral*

Asimismo, la dinámica viral ha sido modelada en otras ocasiones mediante un modelo biexponencial con dos fases de decrecimiento, tal como se expresa con la función:

$$Ae^{-Bt} + Ce^{-Dt}$$

Las medidas de la carga viral de cada uno de los pacientes se han realizado según el siguiente patrón: del día 0 al 3 se han tomado medidas cada 6 horas, del día 4 al 7 cada 24 horas, y a partir del día 8 se ha tomado una medida diaria los días 9, 12, 16, 19, 23, 26 y 30.

A continuación se muestra un gráfico con los perfiles de cada uno de los individuos separados por grupos de tratamiento (Figura 3), entendiendo como perfiles la evolución del logaritmo en base 10 de la carga viral en cada instante de tiempo en que se ha hecho una analítica para medirla.

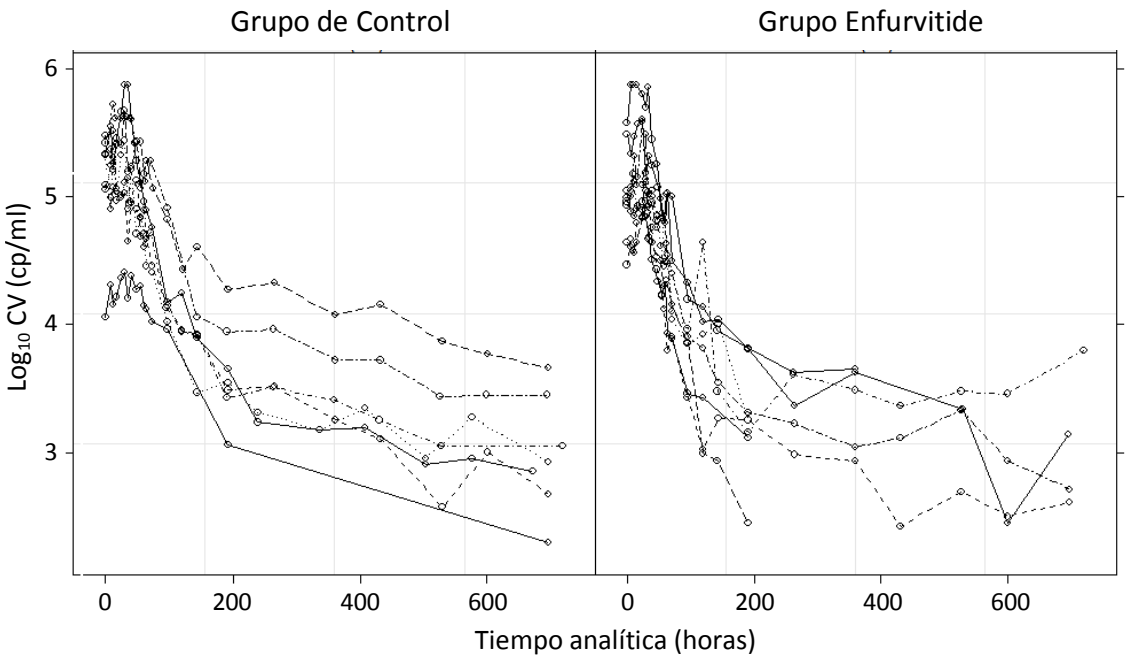


Figura 3: *Perfil de la dinámica viral para grupos y para pacientes*

A la vista de los gráficos de los perfiles de los pacientes (Figura 3) podemos ver que más o menos se sigue el patrón mencionado: una subida inicial seguida de un descenso rápido y, finalmente, un decrecimiento más lento. Por esta razón los modelos que ajustaremos seguirán esta forma.

Métodos

Datos longitudinales y medidas repetidas

A menudo nos encontramos con estudios que recogen una misma información sobre la misma unidad o individuo, teniéndose medidas repetidas de la variable respuesta. Si, además, esta recopilación de datos se realiza a lo largo de un cierto periodo de tiempo los datos estarán ordenados temporalmente, es lo que se llaman datos longitudinales. El hecho de tener una misma medida recogida diversas veces durante un cierto espacio temporal para cada una de las unidades de observación (pacientes, en nuestro caso) hace que éstas estén correlacionadas entre ellas, y no se puede usar la teoría clásica de los modelos de regresión. Hay que pensar, pues, en modelos específicos como los que se utilizan en este estudio.

Modelo Biexponencial

El modelo biexponencial es un caso particular de modelo no lineal y es apropiado para captar la forma de la dinámica viral. La ecuación para la respuesta Y_{ij} del individuo i-ésimo en el instante de tiempo t_j viene dada por:

$$Y_{ij} = \phi_{1i} \exp\left[-\exp(\phi_{2i})t_j\right] + \phi_{3i} \exp\left[-\exp(\phi_{4i})\, t_j\right] + \varepsilon_{ij}$$

Este modelo, al ser un modelo biológico que describe el descenso en la carga viral, es fácilmente interpretable en términos clínicos. Así pues, ϕ_{2i} y ϕ_{4i} son las tasas de descenso de la carga viral en la primera y segunda fase de decrecimiento. Por otra parte, $\phi_{1i} + \phi_{3i}$ representa la carga viral en el instante de tiempo cero. Intuitivamente se puede ver que ϕ_{1i} representaría el intercepto, o punto de partida, de la primera exponencial y ϕ_{3i} el de la segunda. ε_{ij} representa la variabilidad aleatoria que no se puede modelizar.

Modelización

Aplicando las técnicas estadísticas adecuadas y utilizando los datos a partir del momento en que la carga viral es máxima en cada uno de los pacientes, se llega al siguiente modelo (se trata de un modelo llamado "de efectos mixtos"):

$$\begin{aligned}\log_{10} CV_{ij} = \log_{10}\{&(\exp(\beta_{10} + \beta_{11} Grupo_{ENF} + \beta_{12} CV_{basal} + \beta_{13} Edad) \cdot \\ &\cdot \exp\left[-\exp(\beta_{20} + \beta_{21} Grupo_{ENF} + \beta_{22} CD4_{basal} + \beta_{23} Edad)t_j\right] + \\ &+\exp\left(\beta_{30} + \beta_{31} Sexo_{mujer} + \beta_{32} CD4_{basal}\right) \cdot \exp\left[-\exp\left(\beta_{40} + \beta_{41} T_{inf}\right)t_j\right]\} + \varepsilon_{ij}\end{aligned}$$

En este modelo tenemos los parámetros (las betas) y las variables que se ha comprobado que tienen influencia sobre la respuesta. Los valores estimados para los parámetros se encuentran en la Tabla 2 y la descripción de las variables está en la Tabla 3.

Tabla 2: *Valores estimados para los parámetros del modelo escogido*

Parámetro	Estimación	Parámetro	Estimación	Parámetro	Estimación
β_{10}	8,170	β_{20}	-4,167	β_{30}	9,022
β_{11}	0,649	β_{21}	0,361	β_{31}	1,215
β_{12}	0,0000035	β_{22}	0,000316	β_{32}	-0,00186
β_{13}	0,112	β_{23}	0,0175	β_{40}	-6,245
				β_{41}	0,00388

Tabla 3: *Variables que entran en el modelo*

Variable	Descripción	Valores
$Grupo_{ENF}$	Grupo asignado	control=0; Enfurvitide=1
CV_{basal}	Carga Viral inicial	copies/ml
$Edad$	Edad	años
$Sexo_{mujer}$	Sexo	hombre=0; mujer=1
T_{inf}	Tiempo de infección	meses
t_j	Tiempo en tratamiento	horas

Si se representa una curva para cada grupo, con los siguientes valores de las variables: hombre con unos CD4 basales de 401,57 (cels/mm3), una CV basal de 158.489,32 (copias/ml), de 37 años y con un tiempo de infección por VIH de 30 meses (valores medios del grupo control), se obtiene el gráfico de la Figura 4 donde podemos ver que el grupo que toma el Enfurvitide tiene un decrecimiento de la carga viral en la primera fase más rápido que el grupo control.

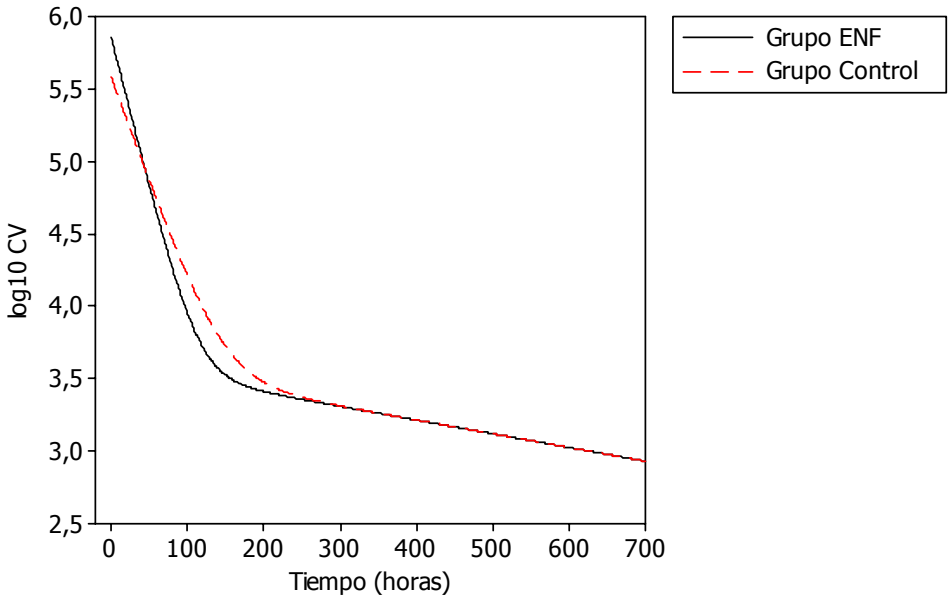


Figura 4: *Gráfico de la previsión de la dinámica viral del grupo de control (línea discontinua) y del grupo ENF (línea continua)*

Conclusiones

El objetivo principal era ver si el Enfurvitide aumentaba la efectividad de un tratamiento antirretroviral formado por cuatro fármacos en pacientes que nunca han seguido este tipo de tratamiento. Además, al disponer de las características de los

pacientes a nivel basal, se ha querido ver cuáles de estas características aportaban información sobre el descenso de la carga viral.

Con el modelo construido se ha comprobado que el grupo de pacientes que han tomado Enfurvitide muestran un decrecimiento de la carga viral en la primera fase más rápido que el grupo de control. En este modelo, aparte del tratamiento, se ha visto que en esta primera fase influyen el CD4 basal y la edad del individuo en el momento de la inclusión, de forma que cuanto más alto se tenga el recuento de células CD4 basal o cuanto mayor sea el paciente, más rápido será el descenso de la carga viral. La segunda fase de decrecimiento sólo se ve afectada por el tiempo desde la infección por VIH, en el sentido de que cuanto más tiempo haga de la infección, más rápido es el descenso de la carga viral.

(Proyecto de la Licenciatura de Estadística, presentado en diciembre de 2006 con el título "Models no-lineals d'efectes mixtes per estudiar la dinàmica viral de pacients infectats pel VIH")

Este proyecto se realizó en el marco de un convenio de colaboración que la Fundación Lucha contra el SIDA y la UPC firmaron el año 2002. Un equipo de la UPC, dirigido por la profesora Guadalupe Gómez, da soporte en las técnicas estadísticas de recogida y análisis de datos para la investigación clínica que se está llevando a cabo en la lucha contra el SIDA. Como parte de estos trabajos, un becario, normalmente estudiante de los últimos cursos, realiza sus prácticas en el hospital Germans Trias i Pujol*, integrado en equipos de médicos, psicólogos y otros profesionales aportando sus conocimientos de análisis de datos para la toma de decisiones.*

Raquel López Blázquez estudió la Diplomatura de Estadística y también la Licenciatura en Ciencias y Técnicas Estadísticas en la UPC. Fue una alumna muy brillante y obtuvo el Premio extraordinario final de carrera concedido por el Ministerio de Educación y Ciencia. En la foto, la ministra Sra. María Jesús San Segundo entregándole el premio (Diciembre 2005).

Realizó el proyecto final de carrera de la Licenciatura como becaria de la *Fundació Lluita contra la SIDA* y actualmente trabaja en equipos interdisciplinarios realizando investigaciones en el ámbito de la medicina en el Instituto Guttmann de Barcelona.

11

Comportamiento de la demanda de urgencias hospitalarias y factores meteorológicos asociados

Proyecto realizado por: **Jordi Real Gatius**
Dirigido por: **Josep Anton Sánchez Espigares** y **Aureli Tobías Garcés**

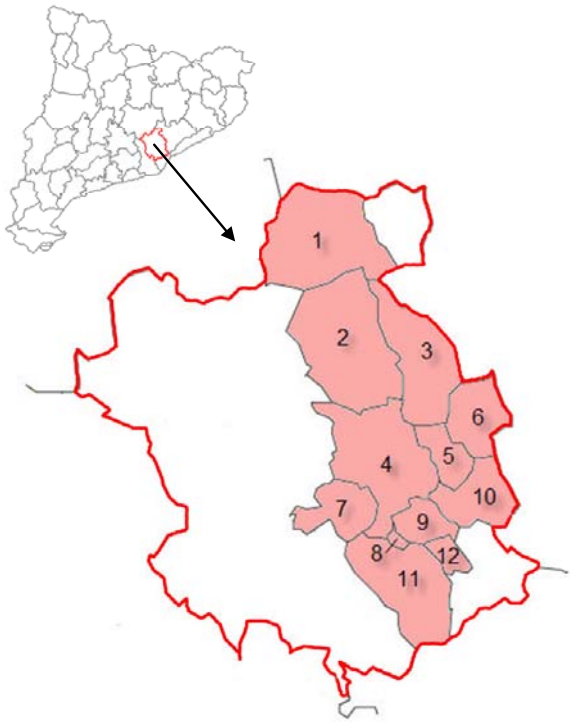
La demanda de los servicios de urgencias ha ido aumentando en los últimos años. Los factores que a menudo se citan como causa de este aumento son: el crecimiento de la población, el cambio en la actitud de los usuarios (que cada vez utilizan más este servicio) y el envejecimiento de la población.

Por otra parte, el aumento de recursos no siempre ha ido en paralelo al aumento de la demanda, de forma que en muchos casos ha disminuido la calidad asistencial, se han sufrido dificultades organizativas y en alguna ocasión incluso se ha llegado al colapso. Se sabe que la demanda de los servicios de urgencias fluctúa con el tiempo (depende de la hora del día, del día de la semana, de la época del año...) y que algunas patologías pueden estar relacionadas con factores climatológicos o ambientales.

Un buen conocimiento de los patrones de comportamiento de la demanda y de las variables que estan relacionadas puede ayudar a prever el número de visitas y a planificar mejor los servicios, optimizando el uso de los recursos disponibles y dando una mejor calidad asistencial. Este proyecto trata sobre esta problemática en la zona de influencia del hospital de Sabadell.

Ámbito del estudio. Objetivos

Este proyecto estudia la evolución del número de personas atendidas por el servicio de urgencias del Hospital de Sabadell durante el periodo 1998-2004. El Hospital de Sabadell pertenece a la Corporación Sanitaria Parc Taulí[1] y su área de influencia comprende 12 municipios ubicados en la comarca del Vallès Occidental (Figura 1).



	Municipio	Habitantes
1	St. Llorenç Savall	2.210
2	Castellar del Vallès	19.475
3	Sentmenat	6.628
4	Sabadell	193.338
5	Polinyà del Vallès	5.855
6	Palau-Solità i Plegamans	12.499
7	Sant Quirze del Vallès	15.729
8	Badia del Vallès	14.313
9	Barberà del Vallès	27.202
10	Santa Perpètua de Mogoda	20.844
11	Cerdanyola del Vallès	56.065
	TOTAL	407.763

Figura 1: *Área sanitaria de referencia del Hospital de Sabadell*

Los tres objetivos que se plantean son:

1. Describir el volumen de visitas al servicio de Urgencias Hospitalarias (SUH) durante el periodo de estudio y las características de los pacientes que se han atendido en cuanto a diagnóstico, edad, sexo y número a veces que han acudido.

2. Estudiar la evolución de la demanda a lo largo del tiempo estratificando por segmentos de edad y por patología asociada. Se quiere identificar, en el caso de que existan, los patrones de comportamiento estacional (diarios, semanales...) y la tendencia general que muestra cada uno de los estratos estudiados.

[1] www.tauli.cat

3. Describir las series de las variables meteorológicas y la evolución demográfica de la población de referencia para determinar la posible existencia de una asociación entre las series meteorológicas y el número de visitas, según cuál sea el grupo de edad (65 años o más y menores de 65) y el tipo de diagnóstico (patologías de tipo circulatorio, de tipo respiratorio y otras).

Los datos

Registros del Servicio de Urgencias del Hospital

El Servicio de Urgencias del Hospital de Sabadell mantiene un registro de los ingresos hospitalarios que es actualizado diariamente desde 1995. A partir del año 1998 se implantó un sistema de codificaciones y de apoyo al personal administrativo para la entrada de códigos al sistema informático, con control de calidad de la información entrada, de forma que estos datos se puedan considerar bastantes fiables en el periodo a partir del cual se realiza este estudio. Las variables consideradas son:

- Número total de visitas al SUH.
- Número de visitas de pacientes de 65 años o mayores.
- Número total de visitas con diagnóstico de tipo circulatorio.
- Número de visitas con diagnóstico de tipo circulatorio y pacientes de 65 años o más.
- Número total de visitas con diagnóstico de tipo respiratorio.
- Número de visitas con diagnóstico de tipo respiratorio y pacientes de 65 años o más.

Datos meteorológicos

Se dispone de un registro diario de las variables meteorológicas. Las variables consideradas son: temperatura media, humedad relativa media, temperatura húmeda media, presión atmosférica media, velocidad media del viento, radiación solar acumulada y precipitación (variable dicotómica). La fuente de estos datos es el Servicio de Meteorología de Cataluña (MeteoCat), que tiene tres estaciones próximas al hospital: Sabadell, Vacarisses y Vallirana. Se tomaron como valores de referencia los de la estación de Sabadell por ser la más próxima al mayor municipio del territorio cubierto por el Hospital.

Algunos días, en concreto 206 días (un 8%), presentaban valores no observados por la estación de Sabadell con respecto a la temperatura, la humedad, la presión

atmosférica, el viento y la precipitación acumulada, normalmente causados por errores del sistema o apagones de electricidad. Se observaron altas correlaciones positivas (coeficientes de correlación de Pearson > 0,95) entre las mismas observaciones registradas en una estación y otra con un ligero desfase. Es decir, los datos de la estación de Sabadell eran bastantes parecidos a los de las otras. Así, las dos estaciones restantes (Vallirana y Vacarisses), sirvieron para la asignación de los valores perdidos en la estación de referencia (Sabadell). La imputación de datos no observados se hizo mediante modelos de regresión lineal basados en los valores de las otras estaciones meteorológicas con datos válidos para aquel día.

Datos demográficos

Como variables demográficas de la población de referencia se tomó la información demográfica de los municipios a los que el Hospital da cobertura, partiendo de las estimaciones intercensales anuales de la pirámide de edad de todos los municipios de Cataluña que calcula el Instituto de Estadística de Cataluña (IDESCAT).

Datos epidemiológicos

Para cada día se conoce si se había declarado un brote en el área cubierta y si fue declarado como epidemia de gripe. Estos datos provienen del Departamento de Salud de la Generalitat de Catalunya que publica el Boletín Epidemiológico de Cataluña donde se hace una recopilación de información sistemática sobre las enfermedades de declaración obligatoria, brotes epidémicos y notificación microbiológica, con las fechas correspondientes y la localización de los casos.

Toda esta información (registros del hospital, datos meteorológicos, demográficos y epidemiológicos) se incorporaron a una sola base de datos que ha sido analizada detalladamente. Cada registro corresponde a un día y como el periodo de estudio es de 7 años (desde el 1 de enero de 1998 hasta el 31 de diciembre de 2004) se han recogido 2.557 registros, tantos como días tiene el periodo.

Descripción de la demanda

Durante el periodo estudiado, el Servicio de Urgencias del Hospital de Sabadell ha atendido 1.305.423 visitas de 763.459 pacientes diferentes.

Un 65% de los pacientes han ido una sola vez al SUH y el resto (35%) han acudido dos o más veces (Tabla 1). El grupo de edad que más ha repetido visita son los menores de 5 años (un 56% ha repetido visita) con una media de 2,5 veces frente al grupo de edad adulta, con una media de 1,6 veces.

Tabla 1: *Frecuencias y porcentajes del número de visitas por paciente*

Pacientes que han sido atendidos …	Número total de visitas que representa	Porcentaje
1 vez	496.868	65,1 %
2 veces	147.468	19,3 %
3 o más veces	119.123	15,6 %
Total	763.459	100 %

Tabla 2: *Frecuencias (n) y porcentajes (sobre el total del grupo de edad) del número de visitas por paciente y grupo de edad*

Grupo de edad (años)	1 vez		2 veces		3 o más		Total
	n	%	*n*	%	*n*	%	
≤ 5	41.816	43,5	21.915	22,8	32.454	33,7	96.185
6-14	51.479	64,8	16.874	21,2	11.139	14,0	79.492
15-64	327.891	70,2	83.679	17,9	55.512	11,9	467.082
65-74	37.792	66,1	11.204	19,6	8.210	14,4	57.206
≥75	37.890	59,7	13.796	21,7	11.808	18,6	63.494
Total	496.868	65,1	147.468	19,3	119.123	15,6	763.459

El 98% de las visitas recibidas durante estos 7 años ha correspondido a pacientes de origen nacional, mientras que el 2% ha sido de origen extranjero. Este porcentaje de visitas de origen extranjero se ha ido incrementando a lo largo de los años, especialmente en los últimos, pasando de un 0,11% el año 1998 hasta el 5 y el 6% en el 2003 y el 2004, respectivamente.

La edad media de las visitas es de 33 años. En el histograma de frecuencias por edades se observan 3 poblaciones claramente diferenciadas: población pediátrica, población adulta y población mayor. Por sexos, un 53% de la demanda es femenina.

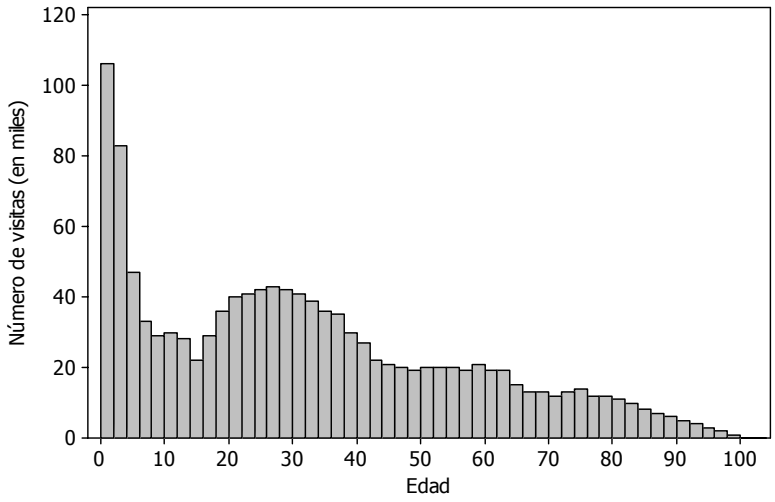


Figura 2: *Histograma de frecuencias por edades de los pacientes atendidos*

Por diagnóstico, las visitas más frecuentes son las lesiones seguidas por las de tipo respiratorio, las del aparato locomotor y las del aparato digestivo. Las de tipo respiratorio representan un 14% del total, variando en función de la época del año. Así, en los meses de invierno representan un 19% y en los meses de verano disminuye a la mitad, representando un 8,6%.

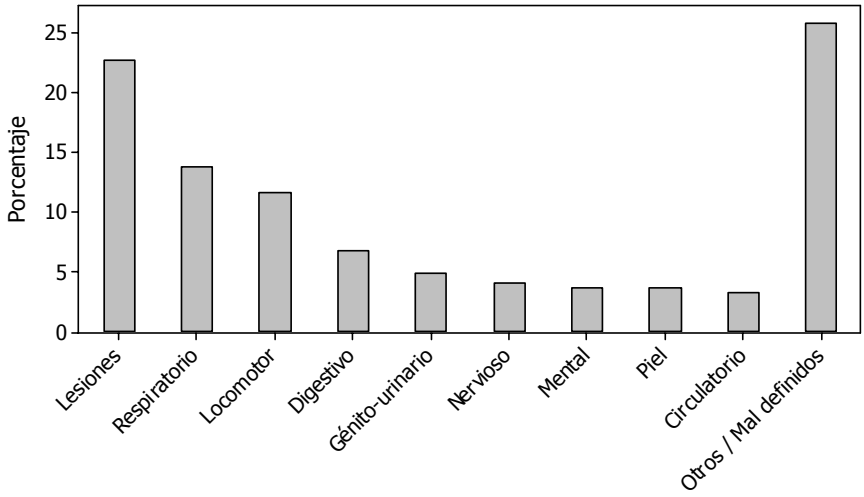


Figura 3: *Porcentaje de visitas según el diagnóstico que se realiza*

La distribución de la edad depende del tipo de visita. Las visitas por razones de tipo cardiovascular provienen de una población más envejecida (con una media de 66 años), mientras que las de tipo respiratorio son mayoritariamente muy jóvenes (media = 20 años) y la mayoría son de edades pediátricas (mediana = 4 años).

Puede sorprender el hecho de que siendo la población cubierta unos 400.000 habitantes, durante 7 años hayan pasado por el Servicio de Urgencias 736.459

personas distintas. Seguramente, muchos demógrafos lo encontrarían lógico, razonando que en un territorio, durante un tiempo determinado, su población es dinámica. Hay nacimientos y defunciones, emigrantes e inmigrantes, personas en tránsito, personas no empadronadas, turistas y, también, pacientes derivados de otros territorios y hospitales. Precisamente, el fenómeno de la inmigración está al orden del día en nuestro entorno, hecho que también se ve reflejado en la procedencia de las visitas de origen extranjero que aumentan año tras año.

Por otra parte, el SUH recibe una gran variedad de pacientes, en cuanto a edad, patología asociada y gravedad. Con respecto a la edad se observan tres poblaciones: la pediátrica y/o neonatal, que merecería un tratamiento diferenciado; la del grupo adulto, situado entre los 20 y 40 años, con unos motivos probablemente relacionados con accidentes laborales y de tráfico; y la del grupo de edad avanzada con patología asociada al envejecimiento. Pero básicamente se pueden distinguir dos grandes grupos: los menores de 65 años y los de 65 o más.

Todo esto ya justifica realizar el análisis separado por edad y diagnóstico, pero también señala que valdría la pena, en un futuro, analizar otros grupos más específicos como el grupo de edad neonatal y pediátrico, que, muy probablemente, tiene una casuística mucho más diferenciada que el resto.

Evolución temporal

La Figura 4 muestra la evolución de la demanda a lo largo del periodo estudiado. Se puede observar una cierta estacionalidad, pero no una tendencia a crecer ya que la demanda anual ha aumentado poco, sólo un 4% el 2004 respeto a 1998.

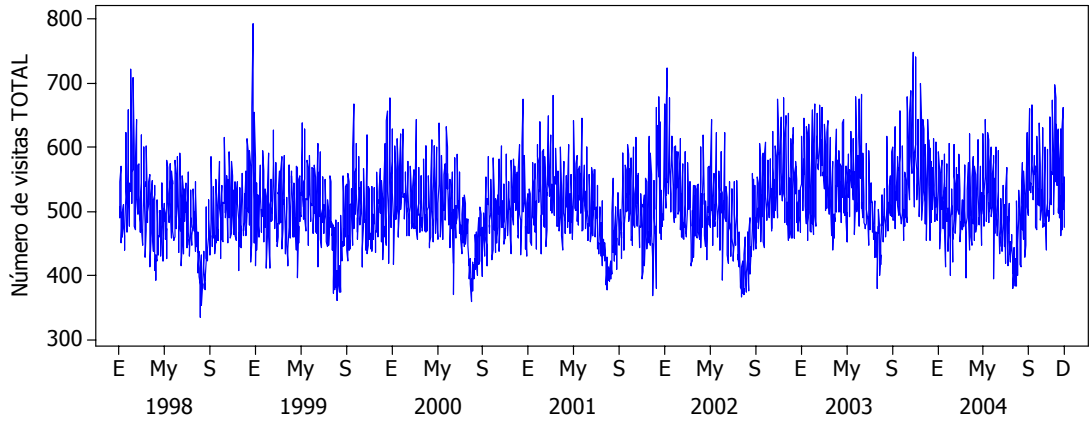


Figura 4: *Evolución de la demanda en el período estudiado*

La evolución temporal pone de manifiesto que en verano, y especialmente en el mes de agosto, la demanda baja y tiene menos variabilidad que en el resto de meses (seguramente debido al menor número de afecciones respiratorias y de enfermedades epidémicas).

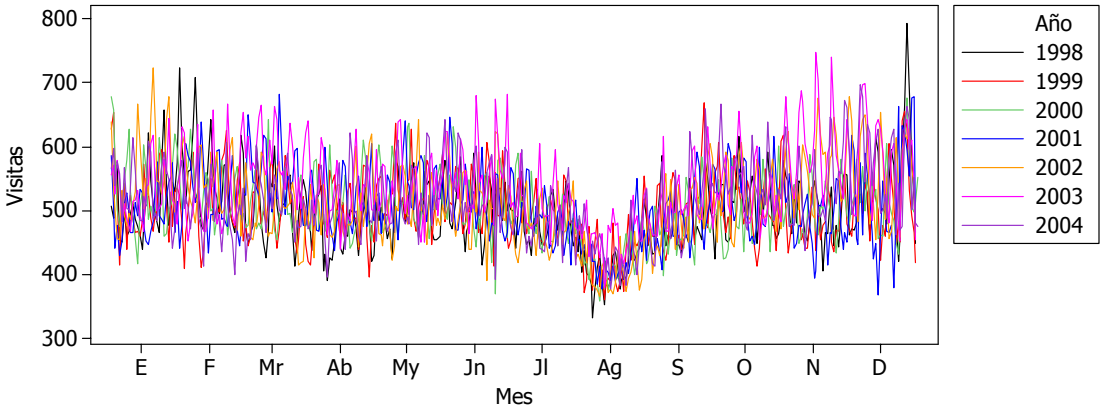


Figura 5: *Evolución anual superpuesta de los 7 años del periodo estudiado. La marca del mes se ha situado en cada día 15*

Se ha observado que semanalmente la máxima demanda se da los lunes y martes mientras que la mínima la tenemos jueves, viernes y sábados.

Si estratificamos por tipo de diagnóstico se observa que cuando está relacionado con el aparato circulatorio prácticamente no hay estacionalidad pero sí una tendencia creciente (Figura 6). El número de visitas por este diagnóstico el año 1998 fue de 4.583 y en el 2004 de 6.024, lo que representa un incremento del 31%. Otro aspecto relevante es que el 63% de estas visitas las han protagonizado pacientes mayores de 65 años.

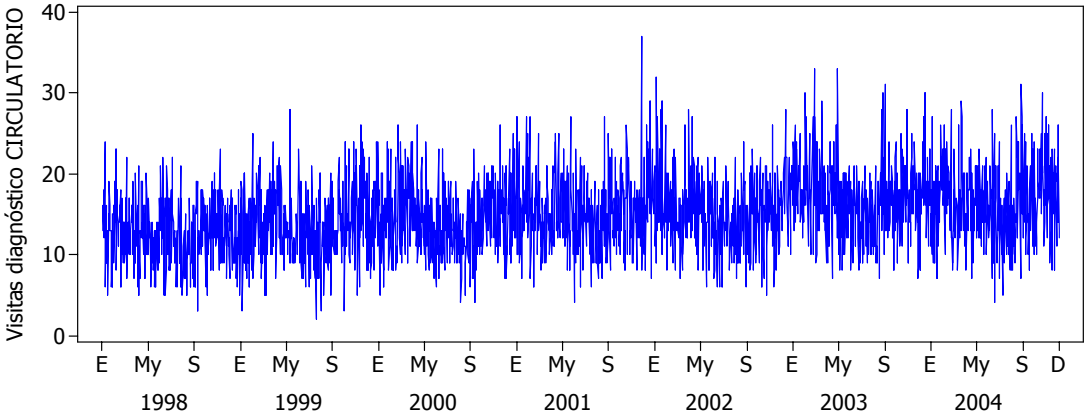


Figura 6: *Evolución del número de visitas con un diagnóstico relacionado con el aparato circulatorio.*

Respecto al número de visitas que tienen como causa el aparato respiratorio, la estacionalidad es clarísima, y en este caso los mayores de 65 años sólo representan un 13% del total, mientras que los niños de 4 años o menores representan el 50%.

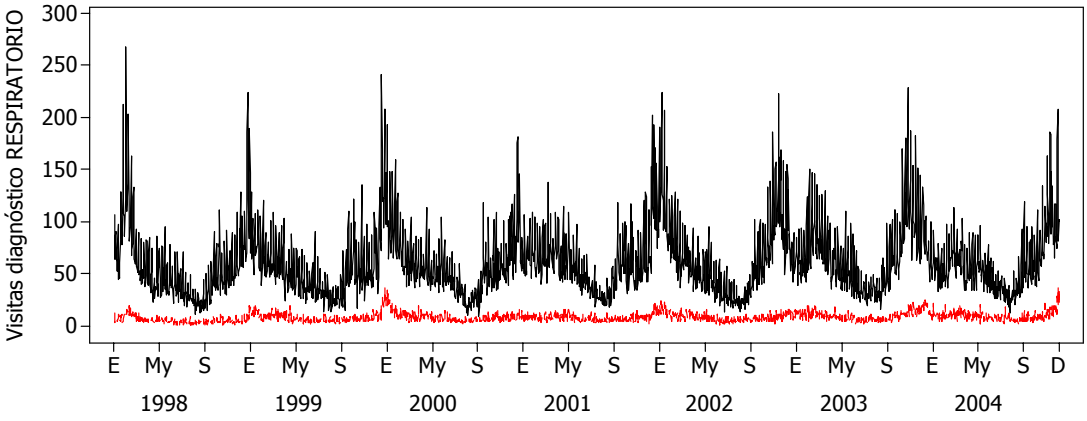


Figura 7: *Evolución del número de visitas, total (negro) y de mayores de 65 años (rojo), con un diagnóstico relacionado con el aparato respiratorio.*

Tanto la temperatura media como el número de visitas con diagnóstico relacionado con el aparato respiratorio tienen un comportamiento claramente estacional. La temperatura media disminuye cada invierno y aumenta en cada verano mientras que el número de visitas aumenta cada invierno y disminuye cada verano (Figura 8).

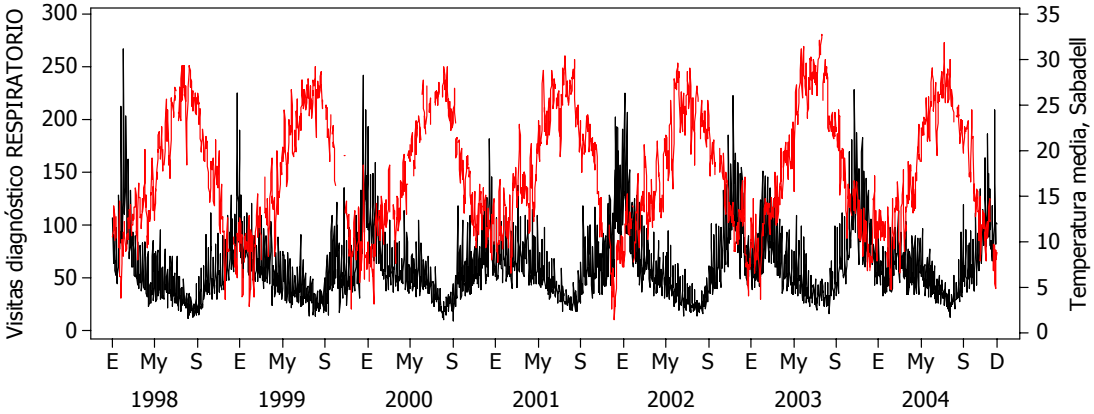


Figura 8: *Evolución del número de visitas con un diagnóstico relacionado con el aparato respiratorio (negro) y la temperatura media diaria en Sabadell (rojo)*

El número de visitas originadas por problemas del aparato respiratorio presenta una diferencia importante entre días laborables y festivos (Tabla 3). Esta diferencia también se observa, pero es menor, en los que tienen 65 años o más. Algunos expertos ven en esta diferencia una prueba de la banalidad de algunas visitas "urgentes" que se rigen por el calendario laboral. Cuando el diagnóstico está relacionado con el aparato circulatorio, las diferencias son pequeñas y a favor de los días laborables.

Tabla 3: *Número medio de visitas según sea día laborable o festivo*

Diagnóstico	Segmento de edad	Número medio de visitas	
		Laborables	Festivos
Respiratorio	< 65	51,0	84,3
	≥ 65	7,47	9,25
Circulatorio	< 65	5, 49	4,83
	≥ 65	9,36	8,89

Descomposición de la variabilidad

A continuación nos centraremos únicamente en las visitas de pacientes de 65 o más años que acuden al SUH por motivos respiratorios.

La Figura 9 representa lo que se denomina descomposición de la variabilidad, ya que a partir de la serie original se construyen tres nuevas series mediante las cuales se podría volver a reconstruir la original. La primera representa la tendencia de la serie, la segunda la estacionalidad y la última contiene la parte residual.

Así pues, observamos una tendencia creciente de forma no lineal durante todo el periodo. Aumenta más los primeros años, para después reducir la velocidad de crecimiento hasta el 2004. También se observa la existencia de un patrón que se repite a lo largo del tiempo (hay estacionalidad). Se puede ver cómo este patrón estacional que se va repitiendo a lo largo de los años es mensual. La parte residual es la que no somos capaces de explicar a través de la tendencia y la estacionalidad.

Hay que tener en cuenta que las escalas de estos gráficos no son las mismas y eso puede conducir a una interpretación incorrecta de la importancia de éstas componentes. Naturalmente, existen procedimientos para valorar la significación estadística de la tendencia y la estacionalidad. En este caso, ambas componentes son significativas.

La Figura 10 ayuda a valorar mejor la estacionalidad. Agrupa las medias mensuales de todos los meses estudiados y las representa en forma de serie temporal. También para cada mes se representa la media para todos los años. Podemos observar que el mes de enero tiene una demanda que ha variado bastante a lo largo de los 7 años estudiados y es al mismo tiempo uno de los meses con más demanda. En general, se observa que la variabilidad aumenta al aumentar el valor medio lo cual es habitual en las variables aleatorias que, como ésta del número de visitas mensuales, se puede modelar a través de una distribución de probabilidad llamada de Poisson.

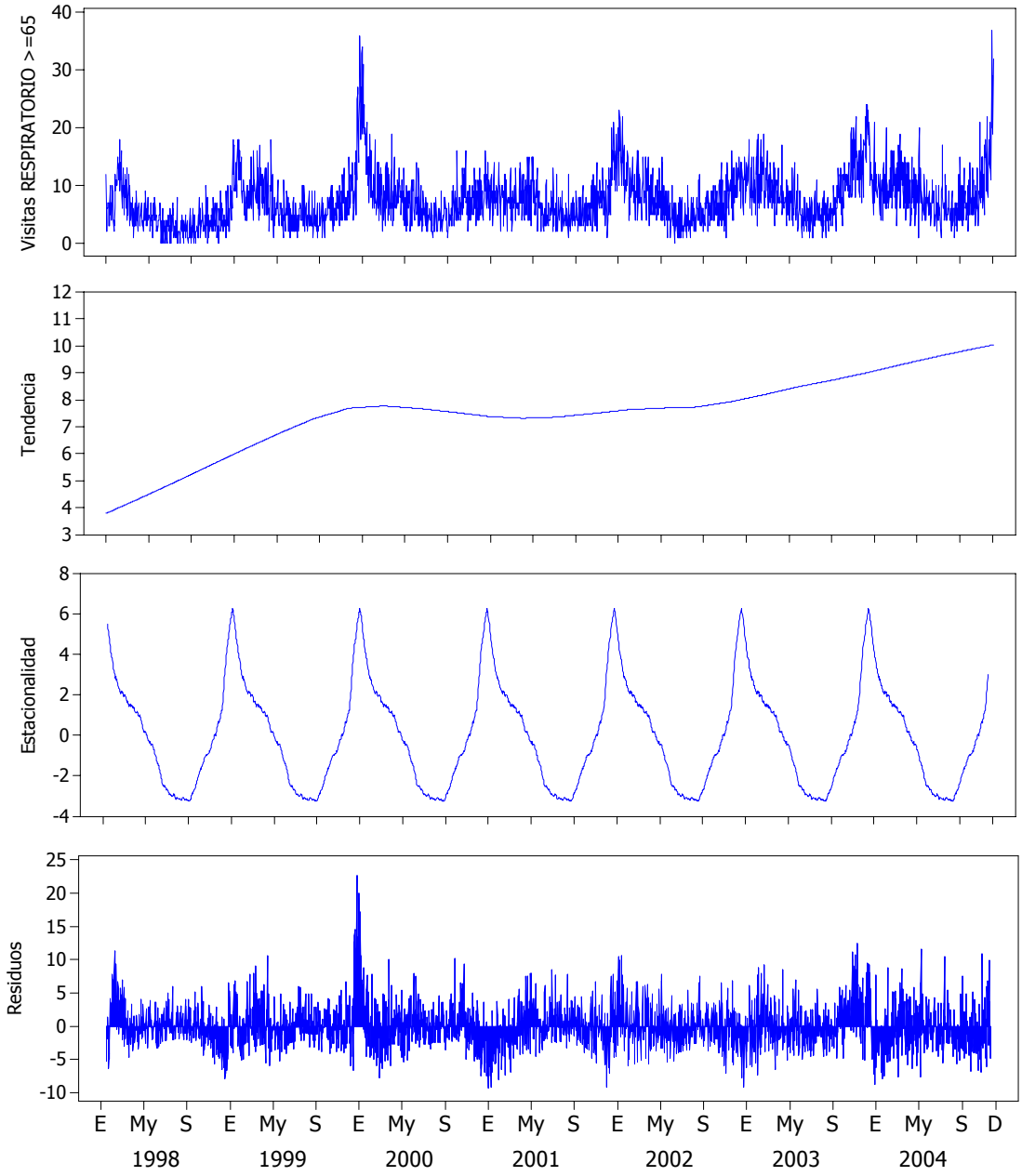


Figura 9: *Descomposición de la variabilidad de la serie temporal correspondiente a los mayores de 64 años con diagnóstico respiratorio*

La estacionalidad que se observa es muy clara y tiene forma de U, denotando que los meses más fríos son los de mayor demanda mientras que los cálidos tienen menos. Tal como habíamos visto en otros gráficos, se observa que las demandas máximas se corresponden con los meses de diciembre y enero, y las mínimas coinciden con el periodo de verano (julio, agosto y septiembre).

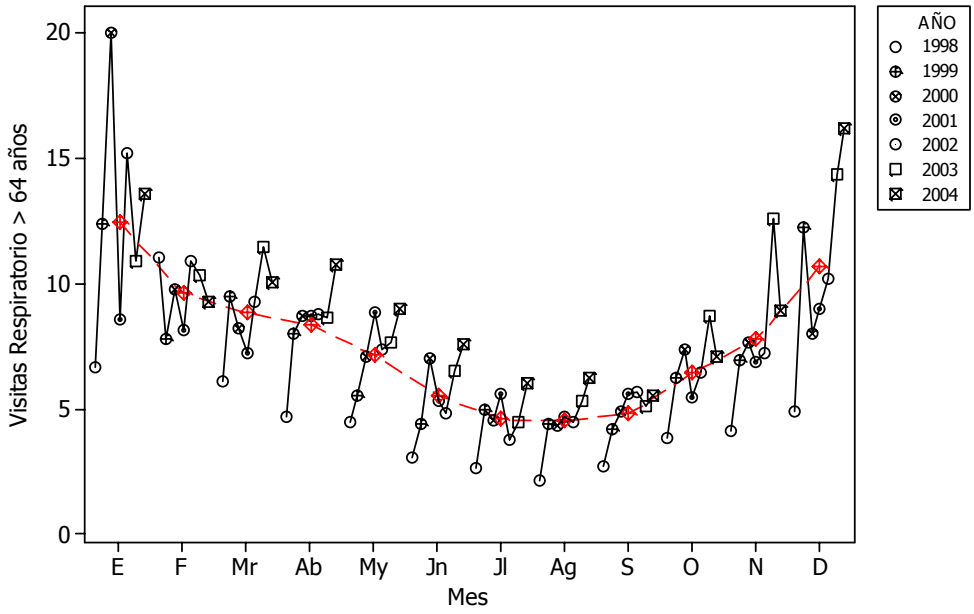


Figura 10: *Valores medios del número de visitas mensuales de los mayores de 64 años con diagnóstico respiratorio. Para cada mes se indican los valores correspondientes a cada año*

Todas las series, además de la analizada, presentan una estacionalidad en forma de V o bien de U, con mayor frecuencia en los meses de invierno y menos en los meses de verano. La serie total tiene una caída de la demanda más pronunciada en julio, agosto y septiembre. El resto del año se mantiene irregular, excepto el mes de abril, que baja un poco (puede coincidir con el periodo de vacaciones de Semana Santa). La serie de mayores de 64 años presenta una estacionalidad más progresiva, disminuyendo a medida que se acercan los meses de verano y aumentando a medida que se acercan los meses de invierno.

Efecto de la climatología: Asociación de diagnósticos respiratorios en pacientes de más de 64 años con factores climatológicos

Más allá del análisis descriptivo de los datos, se quiere obtener un modelo que explique la demanda de este grupo de la población en función de variables ambientales (temperatura, viento y presión atmosférica), incluyendo también el efecto del día de la semana y de si se trata de un día laborable o festivo.

Se ha considerado que lo más adecuado para esta situación es un modelo GAM (*Generalized Additive Models*) y se propone el siguiente para explicar la demanda (más concretamente el logaritmo del riesgo relativo de acudir a urgencias por causas respiratorias en pacientes mayores de 64 años):

$$\log(y_t) = \beta_0 + \beta_1 Festivo + \beta_2 Lunes + \beta_3 Martes + \beta_4 Miércoles + \beta_5 Jueves + \beta_6 Viernes + \beta_7 Sábado + \beta_8 Gripe + \beta_9 f_1(Tiempo) + \beta_{10} f_2(Temperatura) + \beta_{11} f_3(Viento) + \beta_{12} f_4(Presión)$$

Donde, *Festivo*, *lunes*, *martes,...*, *sábado*, *gripe*: toman el valor 1 en caso de que el día lo sea o haya un brote de gripe, y 0 en caso contrario, y f_1, f_2, f_3, f_4: son funciones de las variables tendencia, temperatura, viento y presión, respectivamente.

La Tabla 4 muestra el resultado del ajuste del modelo elegido, incluyendo como variables predictoras: la temperatura, la presión atmosférica, el viento y la existencia de gripe. No han entrado en el modelo la temperatura húmeda, la humedad, la radiación solar y precipitación, dado que han resultado ser estadísticamente no significativas.

Tabla 4: *Coeficientes del modelo de regresión*

Variable	β	Variable (continuación)	β
Constante	4,7465	Festivo (sí)	0.1303
f_1(Tiempo)	0,00015	Gripe (sí)	0.1630
Día de la semana*		f_2(temperatura)	0.0101
Lunes	-0.1120	f_3(viento)	-0.0293
Martes	-0.1657	f_4(Presión)	-0.0029
Miércoles	-0.1479		
Jueves	-0.0972		
Viernes	-0.1759		
Sábado	-0.2077		

*El día de referencia es el domingo.

Que los coeficientes que acompañan a las variables viento y presión sean negativos significa que la presencia de viento y altas presiones tienen efectos inversamente proporcionales sobre la demanda. O sea, a más viento y más presión menos demanda de la esperada. Con la temperatura pasa lo contrario, aunque la función no es del todo lineal. Dependiendo del instante de que se trate, el riesgo de acudir al servicio de urgencias es más elevado. Se observa un cierto patrón mensual que se va repitiendo (la afluencia no es igual todos los meses del año) pero que también varía en función del año (tiene una tendencia creciente).

La categoría de referencia del día de la semana (el día que no aparece en el modelo) es el domingo. El resto de días tienen asociados unos coeficientes que son negativos en todos los casos. Por lo tanto, el efecto de cada día de la semana es negativo sobre la

demanda respecto del domingo. En otras palabras, el domingo es el día con más demanda, mientras que el sábado es el que tiene menos (coeficiente más negativo).

Por otra parte, los días declarados como días con epidemia de gripe y los días festivos incrementan la demanda esperada.

En la modelización se ha seguido el método de la construcción de modelos paso a paso. Se ha empezado con un modelo basal con los efectos del calendario y la estructura temporal y a continuación se han ido añadiendo los predictores que resultan significativos y mejoran el ajuste respecto al modelo anterior. Este método no asegura que el resultado final sea independiente del orden de entrada de los predictores, ya que si se hubiera evaluado primero la temperatura húmeda o la radiación solar antes que la temperatura, muy probablemente el modelo resultante habría sido otro. Por tanto, el modelo presentado no puede considerarse como el único válido entre todos los posibles.

Aun así, reconforta ver que la relación encontrada entre la temperatura y las visitas por motivos respiratorios es muy parecida a la que determinan otros autores al evaluar asociaciones entre contaminantes ambientales y las visitas a un servicio de urgencias hospitalario por motivos de asma, donde utilizaban variables meteorológicas como control. Por el contrario, la presión atmosférica determina el buen o mal tiempo y también está relacionada con el viento, y el viento es un buen predictor de contaminantes ambientales. Así, sin profundizar en otras interpretaciones, no es de extrañar la existencia de estas asociaciones en este análisis.

Las otras variables evaluadas (humedad, radiación solar, temperatura húmeda y precipitación) no aportan suficiente información para mejorar el ajuste, son redundantes o simplemente no afectan.

De todas formas, para entrar más a fondo en interpretaciones haría falta la intervención de un experto para evaluar si las asociaciones encontradas pueden tener algún sentido clínico/biológico o social, o pueden están ligadas a otros factores desconocidos que no se han evaluado en este análisis.

Conclusiones

La aplicación de herramientas gráficas de series temporales ha permitido detectar los patrones de comportamiento de las series de visitas al Servicio de Urgencias Hospitalarias de un hospital (en este resumen se ha presentado sólo la serie de visitas de mayores de 64 años con diagnóstico respiratorio). Estos patrones son diferentes en

función del motivo de diagnóstico y del grupo de edad. En general, la demanda de las visitas a urgencias sigue un patrón estacional, con tendencia positiva.

Las series de visitas relativas a gente mayor han crecido mucho más a lo largo de los años que el resto. Esta tendencia ha supuesto un incremento del 28% en 7 años en la serie de mayores de 64 años. Este incremento no se corresponde al envejecimiento de la población de referencia.

Las visitas por motivos respiratorios tienen una variabilidad estacional más pronunciada que las de los otros motivos analizados. Es decir, el efecto del mes tiene más influencia en las causas respiratorias que en las otras causas.

Parte de las fluctuaciones de demanda durante el año están relacionadas con movimientos migratorios debidos a periodos de vacaciones. También se ha puesto de manifiesto que una parte de la demanda de urgencias hospitalarias, seguramente las más banales, se rigen por el calendario laboral. De hecho, la existencia de la patología banal queda reflejada en la gran variabilidad existente entre los días laborables y festivos.

(Proyecto de la Licenciatura de Estadística, presentado en noviembre de 2006 con el título "Anàlisi del comportament de la demanda d'urgències hospitalàries i factors meteorològics associats")

Seguramente todos los que alguna vez nos hemos pasado horas esperando en un servicio de urgencias (sí, horas y urgencias) hemos dedicado alguno de esos ratos de aburrimiento e impaciencia a pensar en qué habría que hacer para mejorar el servicio.

Quizá por esa razón estos trabajos se miran siempre con buenos ojos. Toda la información que dé pistas sobre cómo conviene distribuir los recursos para mejorar la calidad del servicio será bienvenida. ¡Falta hace!

Jordi Real realizó los estudios de la Diplomatura de Estadística en la Universidad Autónoma de Barcelona, donde muy pronto empezó a colaborar como analista de datos en diversos proyectos relacionados con la salud (para uno de estos trabajos realizó una estancia de 3 meses en Nicaragua). Más adelante mientras trabajaba como consultor estadístico en el Hospital Parc Taulí cursó la Licenciatura de Estadística en la UPC. De la estadística le gusta que tenga aplicaciones en ámbitos muy distintos. Es por esto que se considera un estadístico aplicado y, especialmente, descriptivista. Actualmente enseña bioestadística en la Universidad de Lleida y trabaja como asesor de profesionales sanitarios en una unidad de soporte a la investigación en el ámbito de la atención primaria (IDIAP-Jordi Gol-ICS Lleida).

12

Análisis de la mortalidad por tumores malignos de mama y estómago en Cataluña

Proyecto realizado por: **Xavier Puig Oriol**
Dirigido por: **Josep Ginebra i Molins**

Las diferencias en la distribución geográfica de las causas de mortalidad son una información de gran interés para luchar contra ellas. Las primeras hipótesis sobre el origen de muchas enfermedades han sido establecidas a partir de la identificación de una mayor frecuencia de aparición en ámbitos geográficos donde hay presencia o ausencia de ciertos factores, sean tipos de hábitos, alimentación, exposiciones ambientales u otros.

Además, conocer el patrón de distribución geográfica de cualquier causa de muerte ya tiene valor por sí mismo, ya que puede servir para la toma de decisiones en el ámbito de la gestión sanitaria y de la salud pública, mostrando las áreas donde es más prioritario intervenir, así como para evaluar la efectividad de algunas intervenciones o programas sanitarios implantados en las diferentes zonas.

Por otra parte, conocer la evolución a lo largo del tiempo de las causas de mortalidad aporta también una información valiosa para identificar tendencias, planificar recursos y evaluar los resultados de las acciones que se van desarrollando.

Introducción

El objetivo del proyecto era proporcionar una información complementaria a los indicadores de mortalidad elaborados anualmente por el Registro de Mortalidad de Cataluña, presentando un análisis temporal y un análisis espacial. En este resumen se ilustran algunos de los resultados para el tumor maligno de mama en mujeres y para el tumor maligno de estómago tanto en hombres como en mujeres.

Con respecto a los tumores que aquí se analizan, el tumor maligno de mama femenina era la causa de mortalidad más frecuente por cáncer entre las mujeres en el momento del estudio. En Cataluña, este tipo de cáncer causaba unas 1.000 muertes anuales y las tasas de mortalidad eran de las más altas de España. El abordaje terapéutico de que está siendo objeto en los últimos años está teniendo un impacto muy importante en la supervivencia, motivo por el cual el análisis de su evolución temporal es de gran interés. Se ha investigado mucho sobre la etiología de este tumor, aunque los factores de riesgo conocidos explicarían menos de la mitad de los casos observados. También se ha descrito la existencia de una susceptibilidad genética (mayor frecuencia en mujeres con familiares afectados) y la relación con diferentes factores hormonales y reproductivos (menopausia tardía o primer hijo engendrado a edad avanzada, entre otros). La obesidad, el alto consumo de grasas y proteínas animales, y el consumo de alcohol también se asocian con la mayor frecuencia de aparición de este tumor.

La mortalidad por cáncer de estómago en hombres ocupó el cuarto lugar entre las causas de mortalidad por cáncer durante el periodo estudiado, después del cáncer de pulmón, próstata y colon. Los factores de riesgo del cáncer de estómago en las mujeres son los mismos que en los hombres: los hábitos alimenticios y el consumo de alcohol. Éste último podría explicar en parte la mayor frecuencia de aparición de este cáncer en los hombres. Hasta la realización del proyecto, la dieta como factor de riesgo había tenido un claro patrón espacial y por este motivo tiene un especial interés el estudio de la distribución geográfica de esta enfermedad.

Los datos

Los datos necesarios para realizar un análisis de la mortalidad son el número de defunciones y el tamaño de la población. Tanto las defunciones como la población se pueden categorizar de acuerdo con el sexo, la edad, el año de defunción y la comarca de residencia.

Los datos de defunciones provienen del Registro de Mortalidad de Cataluña, del Departamento de Salud de la Generalitat. En este análisis se consideran las defunciones de residentes en Cataluña ocurridos en este territorio durante el periodo 1986-2000. No se incluyen los residentes en Cataluña muertos fuera del territorio catalán, que representan aproximadamente el 1% del total.

Los datos de población provienen de las estimaciones intercensales y postcensales a 1 de julio elaboradas por el Instituto de Estadística de Cataluña (IDESCAT) a partir de los censos y padrones de los años 1986, 1991 y 1996. Las unidades geográficas seleccionadas para el análisis han sido las 41 comarcas catalanas.

Indicadores para el análisis temporal

Tasa bruta de mortalidad

Existen diferentes formas de evaluar las tasas de mortalidad. Una de ellas es la "tasa bruta de mortalidad" (*TB*) que se calcula como el cociente entre el número de defunciones en un periodo de tiempo y la población en el mismo periodo. La *TB* expresada como defunciones por 100.000 personas-año es:

$$TB = \frac{D}{P} \times 100.000$$

Donde,

D: es el número total de defunciones durante el periodo de tiempo determinado
P: son las personas-año del periodo

El valor del denominador, personas-año del periodo, se obtiene como la suma de la población en riesgo de sufrir la enfermedad (por ejemplo: sólo mujeres en el caso del cáncer de mama) en cada uno de los años considerados. Éste puede ser un valor controvertido y el criterio más aceptado es utilizar la suma de las estimaciones de la población el 1º de julio de todos los años que forman el periodo.

Tabla 1: *Tasas brutas por 100.000 habitantes para cáncer de mama en Catalunya*

	1986-1990	1991-1995	1996-2000
Nº defunciones	4.623	5.351	4.883
Personas-año	15.369.836	15.530.721	15.685.455
Tasas brutas	30,08	34,45	31,13

Esta tasa bruta de mortalidad ignora la distribución de la población por edad, por lo que resulta poco útil a la hora de hacer comparaciones. Así, si la incidencia de la mortalidad es más elevada en las edades avanzadas, un área más envejecida que otra tendrá una *TB* superior. Es conocido que la edad es un factor determinante en el comportamiento de la mortalidad, pero es un factor no modificable. Para comparar la mortalidad de áreas con diferentes estructuras de población es necesario eliminar el efecto confusor de la edad.

Tasas específicas de mortalidad

Las tasas específicas de mortalidad en lugar de tomar el conjunto de las defunciones y el conjunto de la población sólo consideran una parte de éstos. En nuestro caso, se pueden calcular tasas específicas para cada tramo de edad. Su expresión para 100.000 personas-año y para el grupo de edad x, será:

$$m_x = \frac{d_x}{p_x} \times 100.000$$

Donde:

d_x: es el número total de defunciones en el intervalo de edad x,
p_x: son las personas-año en el intervalo de edad x

Las tasas específicas también se pueden dar por sexo, por causa de mortalidad o por comarca. En la Tabla 2 se muestran las tasas específicas por grupos de edad quinquenales para cáncer de mama durante los tres periodos. Está clara la existencia de un gradiente que aumenta con la edad, exceptuando al primer grupo que se aleja de ese patrón y que podría tratarse de casos mal codificados de la causa o de la edad. Las tasas brutas son una media ponderada de las tasas específicas, donde los pesos son la proporción de población en cada subgrupo de edad.

Tasas de mortalidad estandarizadas

La Tasa de Mortalidad Estandarizada (*TME*) trata de responder a la pregunta: ¿si las dos poblaciones tuvieran la misma estructura de edades, cuál tendría una mayor mortalidad? Es decir, hace que la tasa de mortalidad sea independiente de las pirámides de edad de las poblaciones que se comparan.

La *TME* se calcula tomando como referencia la estructura de edades de una población tipo. Una posibilidad es utilizar como población tipo la población mundial, que permite realizar comparaciones internacionales, pero esta población tiene una estructura muy

joven y produce una gran modificación de las tasas cuando se aplica a poblaciones más envejecidas. Una posibilidad alternativa es utilizar otras poblaciones estándar como la europea, o bien en nuestro caso la población catalana.

Tabla 2: *Tasas específicas de mortalidad por cáncer de mama en Cataluña*

EDAD	PERIODO			TOTAL
	1986-1990	1991-1995	1996-2000	
<1	4,29	0	0,73	1,70
1-4	0,00	0	0,00	0,00
5-9	0,00	0	0,00	0,00
10-14	0,00	0	0,13	0,03
15-19	0,00	0	0,10	0,03
20-24	0,26	0,25	0,16	0,22
25-29	1,59	1,01	0,65	1,07
30-34	5,12	4,71	3,00	4,24
35-39	15,19	10,96	8,13	11,29
40-44	22,01	26,04	18,58	22,15
45-49	35,99	37,49	28,21	33,72
50-54	48,50	50,44	36,87	44,85
55-59	60,80	60,32	47,41	56,49
60-64	64,98	72,24	58,26	65,31
65-69	73,55	76,95	63,34	71,07
70-74	85,78	89,31	89,09	88,21
75-79	103,15	111,2	101,61	105,15
80-84	123,04	146,03	122,32	130,59
85-89	154,58	176,92	151,36	160,78
90-94	184,69	223,04	234,78	219,81
>94	222,46	244,1	263,41	248,24
Total	30,08	34,45	31,13	31,89

Hay diversas formas de calcular la *TME*, en este resumen utilizaremos la llamada forma directa que, para 100.000 personas-año, se calcula con la siguiente fórmula:

$$TME = \frac{\sum_{x=1}^{J} m_x \times \Pi_x}{\sum_{x=1}^{J} \Pi_x}$$

Donde,

m_x: es la tasa específica de mortalidad para 100.000 personas–año en el intervalo de edad x

Π_x: es la población tipo en el intervalo de edad x

J: es el número de intervalos de edad

A partir de los datos de la Tabla 3 se puede determinar la *TME*. Para el periodo 1991-95 en Cataluña, tomando como población de referencia la de 1991, se obtiene una *TME* de:

$$TME = \frac{184.848.870{,}4}{6.059.494} = 30{,}51$$

Taula 3: *Cálculos para la determinación de la tasa de mortalidad estandarizada por cáncer de mama en Cataluña en el período 1991-95. Como población de referencia se ha tomado la de 1991*

j	Segmentos de edad	Población de referencia, Π_x	Tasa específica per segmento de edad, m_x	Producto $m_x \times \Pi_x$
1	0-4	280.083	0	0,0
2	5-9	337.829	0	0,0
3	10-14	458.366	0	0,0
4	15-19	512.091	0	0,0
5	20-24	487.215	0,25	121.803,8
6	25-29	469.141	1,01	473.832,4
7	30-34	446.777	4,71	2.104.319,7
8	35-39	412.802	10,96	4.524.309,9
9	40-44	407.726	26,04	10.617.185,0
10	45-49	367.460	37,49	13.776.075,4
11	50-54	320.004	50,44	16.141.001,8
12	55-59	357.360	60,32	21.555.955,2
13	60-64	335.329	72,24	24.224.167,0
14	65-69	296.938	76,95	22.849.379,1
15	70-74	215.013	89,31	19.202.811,0
16	75-79	168.887	111,2	18.780.234,4
17	80-84	112.547	146,03	16.435.238,4
18	85-89	54.548	176,92	9.650.632,2
19	90-94	16.061	223,04	3.582.245,4
20	>94	3.317	244,1	809.679,7
Total		6.059.494		184.848.870,4

La ventaja de este indicador es que concentra en un solo número la tasa de mortalidad tomando en consideración la estructura de edades y permitiendo comparar datos provenientes de diferentes tipos de población (siempre y cuando se calculen tomando la misma población de referencia).

Pero presenta una limitación cuando se trabaja con poblaciones pequeñas, como pasa en algunas comarcas, ya que hay pocas defunciones y eso hace que las *TME* presenten

una variabilidad elevada que dificulta su interpretación. Una forma de contemplar esta situación es considerar que el número de defunciones es una variable aleatoria (concretamente, una Poisson), y a partir de los datos disponibles estimar su variabilidad, lo que permite calcular intervalos de confianza por la *TME* tal como se presentan en la Tabla 4.

Taula 4: *TME por 100.000 habitantes en Cataluña, con sus intervalos de confianza del 95%*

Periodo	*TME*	Intervalo de confianza del 95%
1986-1990	28,81	(27,98 – 29,65)
1991-1995	30,51	(29,68 – 31,33)
1996-2000	25,29	(24,25 – 26,02)

Análisis temporal de la moralidad por cáncer de mama en mujeres

La primera impresión visual de la evolución de la mortalidad por cáncer de mama, Figura 1, ya sugiere que el efecto del año de defunción no ha sido lineal, sino más bien cuadrático con un incremento importante de las tasas de mortalidad hasta el año 1991, y a continuación un descenso. Por tanto, se introduce un término cuadrático para el año de defunción en el ajuste del modelo, con el fin de recoger este efecto parábola, tal como se ha hecho en el ajuste que se presenta en la figura.

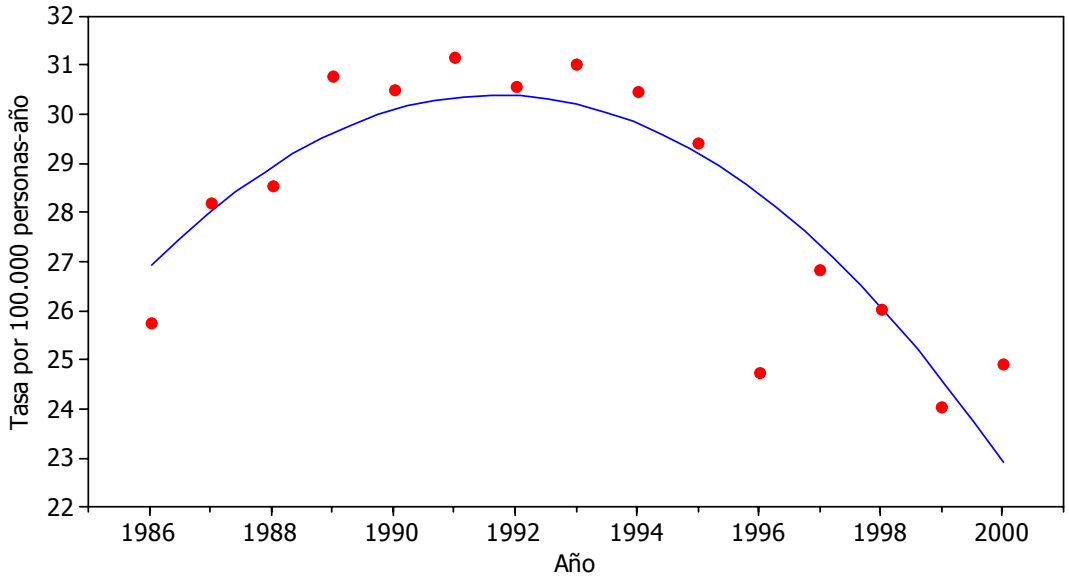


Figura 1: *Tasas ajustadas de mortalidad para cáncer de mama en Cataluña durante el período 1986-2000 tomando como referencia la población de Cataluña del censo de 1991*

Lo que nos planteamos seguidamente fue si la evolución es similar para los diferentes grupos de edad. Al observar el gráfico de las tasas específicas por edad (Figura 2) está claro que existe un gradiente de las tasas con la edad, es decir, que a mayor edad las tasas son más altas. Lo que no se observa tan claramente (sólo fijándonos en los puntos y no en las líneas) es si hay efecto cuadrático.

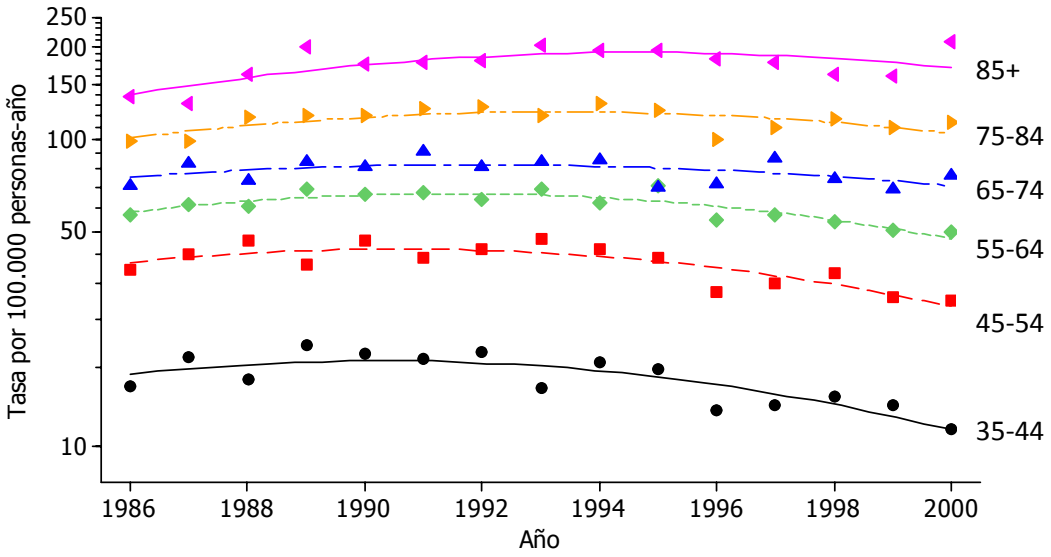


Figura 2: *Tasas de mortalidad específicas para edad, observadas y estimadas por el modelo ajustado de cáncer de mama en Catalunya durante el periodo 1986-2000*

Tanto al escoger el modelo estadístico más adecuado como al validar la bondad del ajuste se han usado criterios estadísticos. El mejor modelo será el más simple de entre los que representen la realidad de forma adecuada. Una vez aplicadas las herramientas estadísticas se ha concluido que el modelo más adecuado es el que contiene como variables explicativas de la mortalidad por cáncer de mama: *edad*, *periodo*, *período al cuadrado* y *edad×período*. El hecho de tener el periodo al cuadrado nos indica que el modelo tendrá curvatura. Por otra parte, observamos que la edad interacciona con el periodo, lo cual quiere decir que el valor de una variable dependerá del valor que tome la otra. Este modelo nos indica que una misma parábola con la misma curvatura sirve para explicar la evolución de todos los grupos de edad pero trasladada a diferentes puntos para cada uno de ellos, cosa que a simple vista no éramos capaces de determinar con seguridad.

A partir del modelo ajustado se puede estimar el año en que se alcanza la tasa máxima para cada grupo de edad, y el valor de la tasa previsto. Estos valores estimados se presentan en la Tabla 5.

Tabla 5: *Estimación del año en que se alcanza la tasa máxima de mortalidad, y valores previstos de dicha tasa por grupos de edad en mujeres con cáncer de mama en Cataluña en el periodo 1986-2000*

Grupo de edad	Año en que se alcanza la tasa máxima de mortalidad	Valores previstos para las tases máximas de mortalidad para 100.000 personas-año
35-44	1989	18,824
45-54	1990	43,591
55-64	1991	65,861
65-74	1992	84,836
75-84	1993	123,923
>84	1994	192,248

Es interesante ver cómo el año en que se alcanza el máximo varía gradualmente con la edad, así el inicio en el descenso experimentado en la mortalidad se observa de forma más tardía a medida que incrementa la edad. Además, las tasas de mortalidad por cáncer de mama son mucho mayores en los grupos de edad más avanzada.

Análisis temporal de la mortalidad por cáncer de estomago en hombres

En Cataluña, la mortalidad por cáncer de estómago ha seguido la misma tendencia descendente que la observada en todos los países desarrollados, tanto en hombres como en mujeres. El descenso observado en la Figura 3 se repite en la Figura 4, que representa la evolución de las tasas específicas por grupos de edad y en la que se observa que las tasas más altas de mortalidad por cáncer de estómago en hombres se dan en los grupos de edad más avanzada.

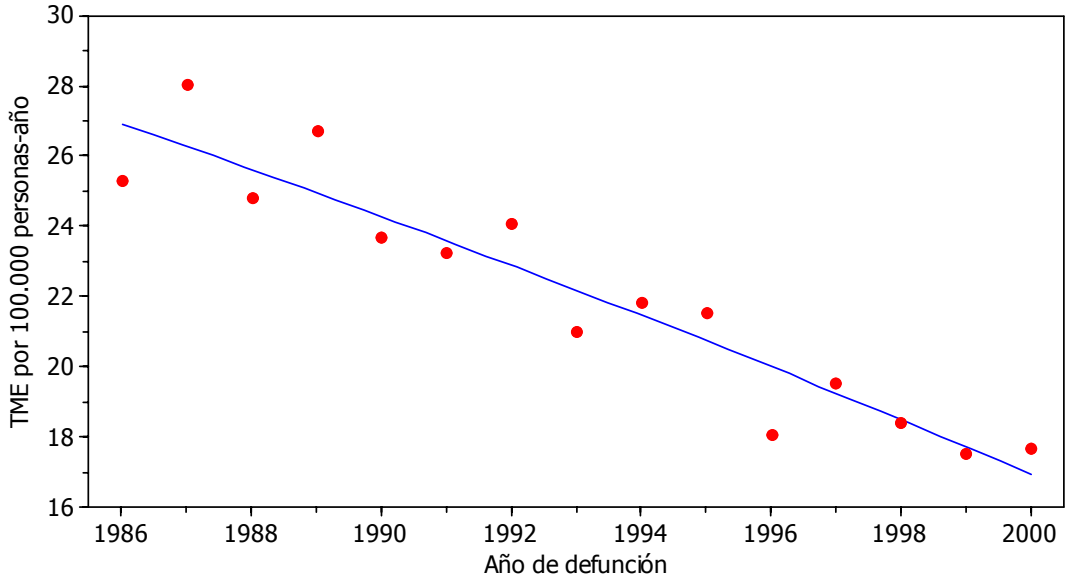


Figura 3: *Evolución de la tasa de mortalidad estandarizada, según la población de Cataluña en el censo de 1991 por cáncer de estómago en hombres en Cataluña durante el periodo 1986-2000*

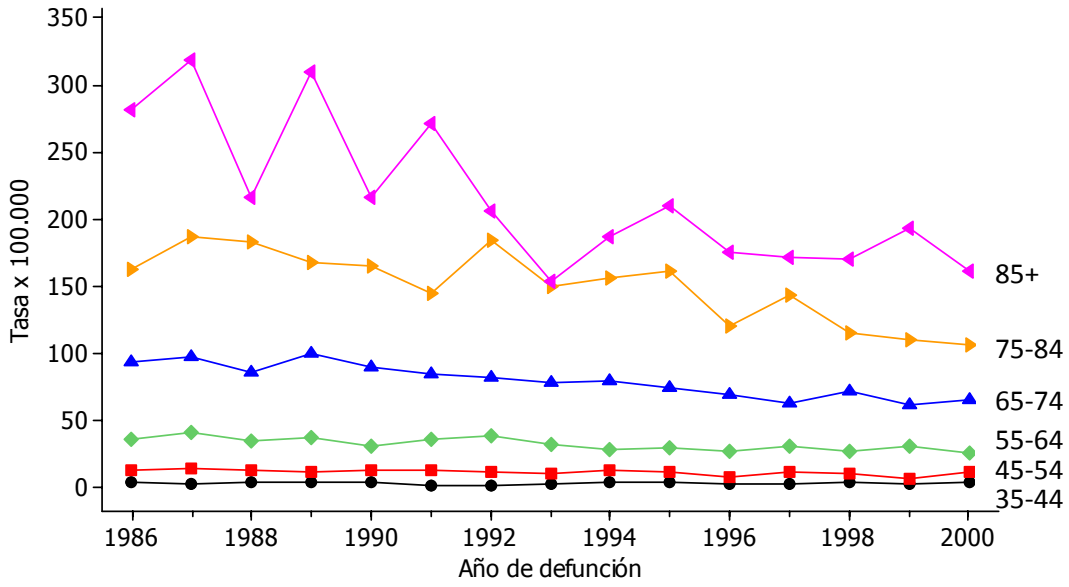


Figura 4: *Evolución de les tasas específicas por edad y año de defunción por cáncer de estómago en hombres en Cataluña durante el período 1986-2000*

Aunque las tasas más elevadas corresponden al grupo de edad «85 y más», éstos representan tan sólo los 8,4% de los casos el año 1986, y el 10,5% el año 2001. Eso podría explicar parte de la variabilidad en las tasas observadas en este grupo de edad.

Haciendo determinadas hipótesis sobre el comportamiento de los datos[1] se ha llegado a la conclusión de que el mejor modelo para explicar el valor esperado del número de defunciones por edad y año es de tipo log-lineal (el logaritmo de la variable respuesta se explica a través de una función lineal). A partir de este modelo se puede deducir que el valor esperado para el porcentaje de cambio anual es del -3,13%. Es decir, hay una tendencia a la disminución de la mortalidad por esta causa.

Análisis temporal de la mortalidad por cáncer de estómago en mujeres

Igual que sucedía con los hombres, la mortalidad por cáncer de estómago en las mujeres ha seguido una tendencia decreciente similar a la observada en el resto de países europeos.

Tal como podemos ver en la Figura 6 las tasas son más altas a mayor edad, y el descenso global de la mortalidad observado en la Figura 5 se repite para todos los grupos de edad, si bien para las edades más jóvenes no se aprecia con claridad, básicamente por una cuestión de escala.

De la misma forma que sucedía en el caso de los hombres, ha aumentado la proporción de casos en los mayores de 85 años a lo largo de todo el periodo observado. No obstante, esta concentración de casos es mucho más acentuada en las mujeres. En el año 2000 el número de muertos por esta causa en las mujeres del grupo de edad mayor representaba el 22% de los casos (recordemos que en los hombres era del 10,5%).

[1] Básicamente que el número de muertos esperados por año y segmento de edad corresponde a una variable aleatoria con distribución de Poisson.

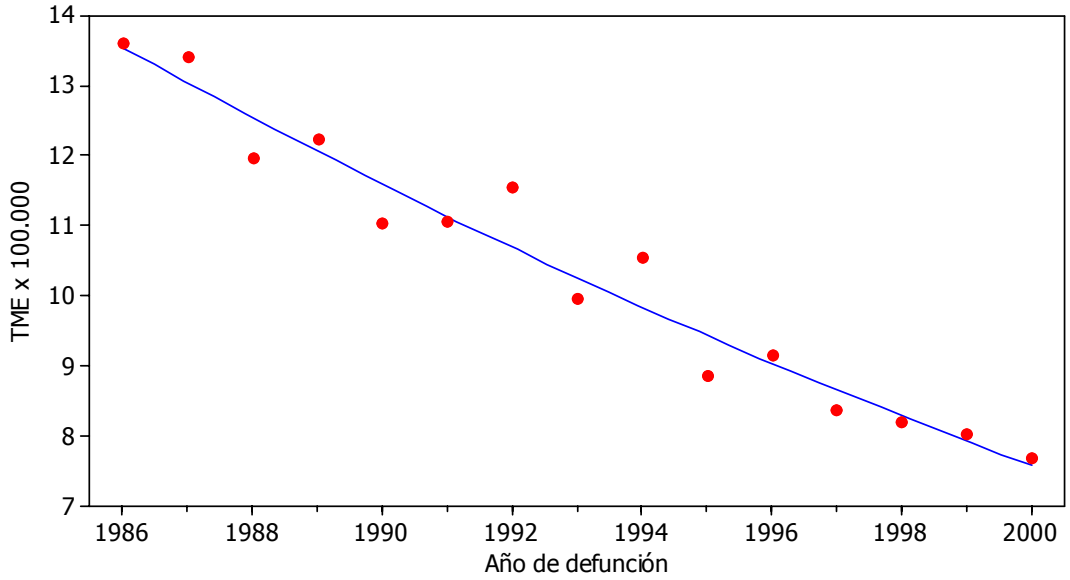


Figura 5: *Evolución de la tasa de mortalidad estandarizada, según la población de Cataluña en el censo de 1991 para el cáncer de estómago en mujeres en Cataluña durante el periodo 1986-2000*

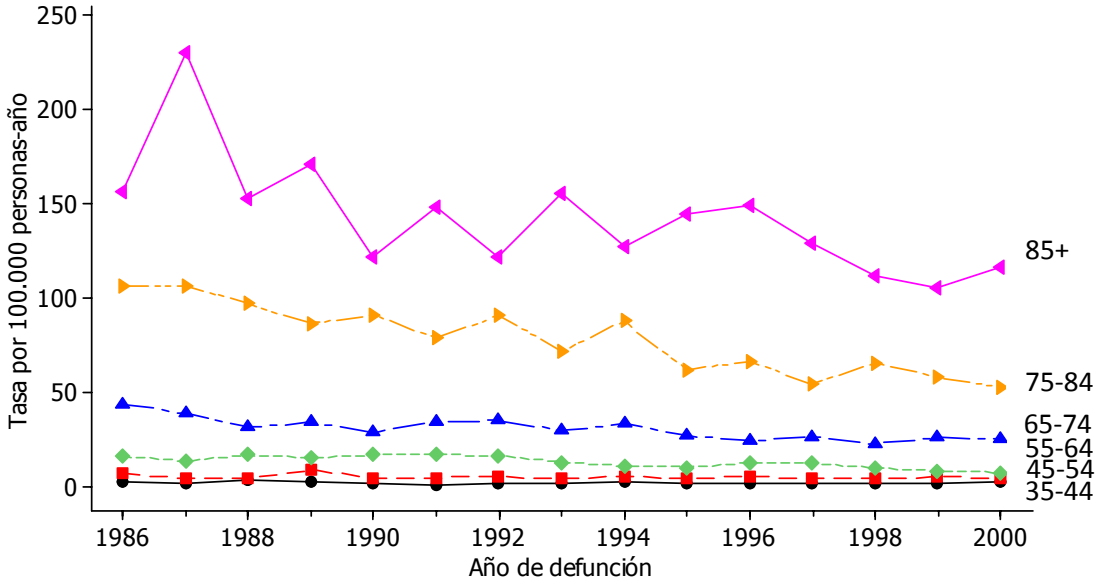


Figura 6: *Evolución de les tasas específicas por edad y año de defunción para cáncer de estómago en mujeres en Cataluña durante el periodo 1986-2000*

El modelo que mejor ajusta estos datos también es de tipo log-lineal y el porcentaje de cambio anual es del -3,91%, así pues la reducción de la mortalidad ha sido ligeramente más marcada en las mujeres que en los hombres (recordemos que éstos presentaban una reducción del 3,13%).

Análisis espacial

El análisis espacial tiene como finalidad el representar los indicadores mediante mapas. La ventaja de los mapas sobre las tablas es que aportan una información suplementaria sobre la contigüidad y proximidad o lejanía entre las áreas, lo cual permite una visualización e integración espacial de la información.

Es preferible que el indicador seleccionado para realizar la representación sea una medida global, como podría ser la *TME*. No obstante, y siguiendo el criterio de la gran mayoría de trabajos al respecto, se usará la razón de mortalidad estandarizada (denotada *RME*), ya que es más fácil de interpretar y presenta mayor estabilidad.

El procedimiento de cálculo de la *RME* se conoce como estandarización indirecta. El concepto es el mismo que el de la directa pero en este caso se aplican una serie de tasas específicas de mortalidad a las diferentes estructuras de población, con lo que se obtieneun número de defunciones esperadas para cada una de ellas. La *RME* surge del cociente entre las defunciones observadas y las defunciones esperadas, y su expresión es:

$$RME_i = \frac{\text{Observados en el área } i}{\text{Esperados en el área } i} = \frac{\sum_{x=1}^{J} d_{x,i}}{\sum_{x=1}^{J} \mathrm{s}_x \cdot \mathrm{p}_{x,i}}$$

donde:

$d_{x,i}$: es el número de defunciones en el intervalo de edad x del área i
$\mathrm{p}_{x,i}$: son las personas-año por edad en el intervalo de edad x del área i
s_x: es la tasa específica estándar en el intervalo de edad x
J: es el número de intervalos de edad

Las tasas específicas estándar son las estimadas a partir del conjunto de las poblaciones que se pretenden comparar, en nuestro caso Cataluña. De esta forma la *RME* se interpreta como un riesgo relativo respecto a la media de las poblaciones a comparar (en este caso comarcas). A pesar de ser una medida relativa, la *RME* es de fácil interpretación como indicador de riesgo, ya que cuando la RME_i es mayor que 1 quiere decir que la mortalidad en la comarca i-ésima es superior a la media de todas las comarcas, mientras que cuando es menor que 1 quiere decir que la mortalidad es menor.

Para enfermedades poco frecuentes y/o áreas pequeñas, la variabilidad que presentan los datos dentro del área puede ser tan grande que no permite considerar como

significativas las diferencias que se puedan observar. Para solucionar estos problemas el proyecto plantea diferentes alternativas metodológicas. Los resultados que aquí se comentan corresponden al enfoque denominado bayesiano, que el autor considera más adecuado, y que permite que las áreas vecinas compartan información.

Análisis espacial para el cáncer de mama en mujeres

En el caso del tumor maligno de mama en mujeres, el análisis espacial no pone de manifiesto diferencias significativas entre comarcas. La noticia sería, pues, que la incidencia es similar en todas las comarcas. Concretamente, para el periodo 1986-1990 todas las comarcas tienen una *RME* con valores entre 0,9 y 1,1; excepto algunas excepciones. En el periodo 1991-1995 sólo se separa de este resultado general la comarca del Maresme, con una *RME* que va de 0,7 a 0,9. En el periodo 1996-2000 el Baix Empordà y el Valle de Arán aparecen con un indicador ligeramente superior (entre 1,1 y 1,3) y el Vallès Occidental está entre 0,7 y 0,9. Pero estos resultados no permiten afirmar la existencia de patrones de comportamiento ni distribuciones de incidencia consolidadas.

Análisis espacial para el cáncer de estómago

El cáncer de estómago, a pesar de la tendencia decreciente de la mortalidad, sigue siendo una causa de muerte muy frecuente. El número de defunciones anuales por cáncer de estómago en mujeres es de casi 200 casos menos que las observadas en hombres y las tasas estandarizadas de mortalidad en mujeres son aproximadamente la mitad que las observadas en hombres. En la Figura 7 se representa para cada sexo la razón de mortalidad estandarizada por edad (*RME*) de cada comarca suavizada con el modelo bayesiano. Aquellas comarcas con valores de la *RME* superiores a la unidad presentan riesgos de mortalidad superiores a las del conjunto de Cataluña, considerado como estándar.

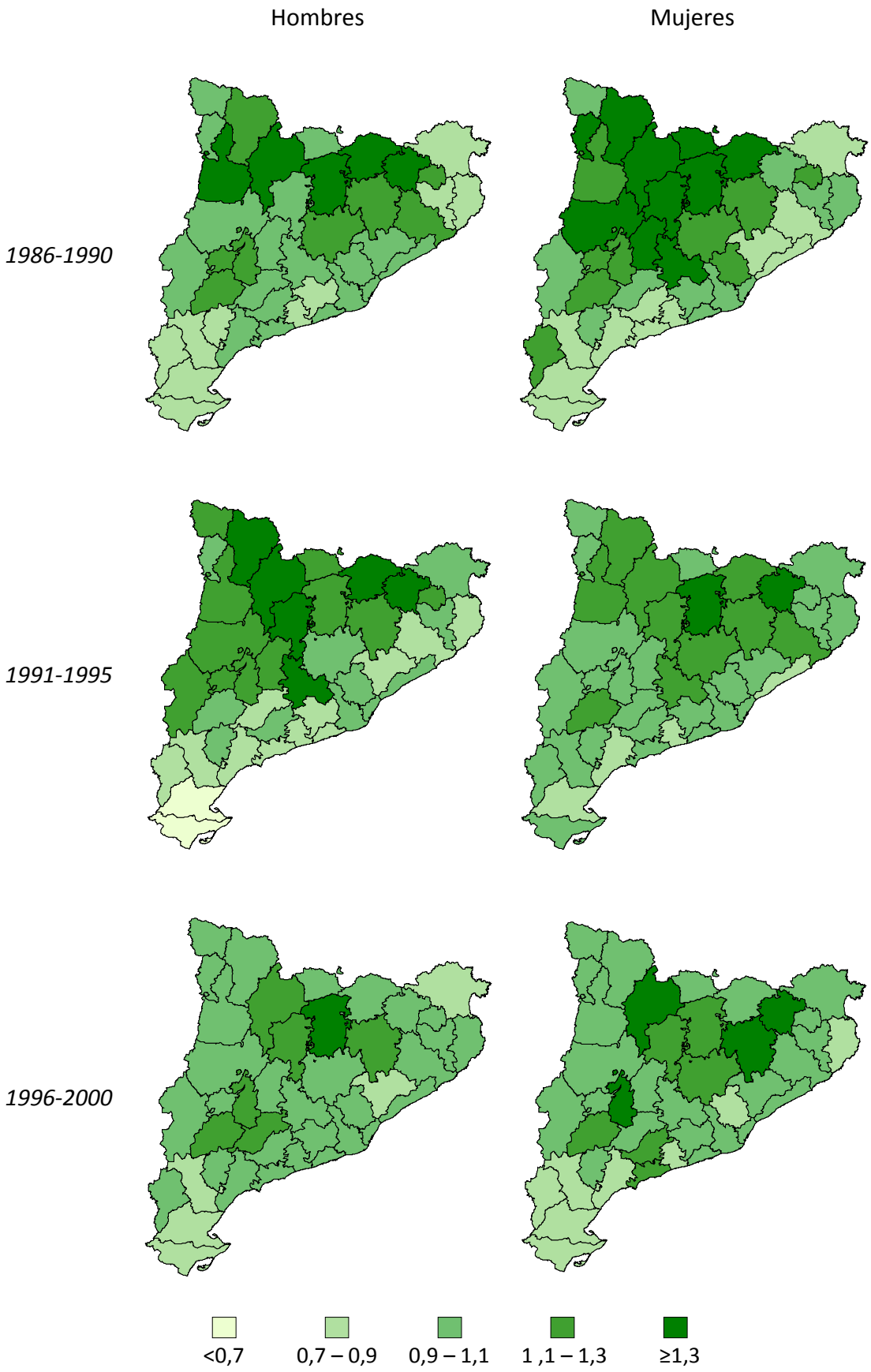


Figura 7: *RME por cáncer de estómago por comarcas de Cataluña durante el periodo 1986-2000*

En el análisis espacial se observa un claro patrón de distribución con un gradiente "costa-interior" para los dos sexos. En los dos primeros periodos estudiados las zonas con riesgo más bajo se encuentran en el litoral, y el riesgo va aumentando hacia el interior y hacia las comarcas pirenaicas. El número de zonas con un riesgo más elevado que el conjunto (las más oscuras) descienden en el último periodo, sobre todo en el caso de los hombres, lo cual provoca que el patrón de distribución espacial se vuelva más homogéneo.

Conclusiones

Los resultados de la evolución de la mortalidad por cáncer de mama muestran un incremento importante en la década de los años ochenta así como una reducción en los últimos años, compatible con las tendencias observadas en otros estudios de nuestro entorno. El año de cambio ha sido diferente para los diversos grupos de edad. No es extraño pensar que el efecto beneficioso de las intervenciones sanitarias, terapéuticas y asistenciales afecte de forma diferente a los diversos grupos de edad y, en particular, que los primeros beneficiados sean los grupos más jóvenes.

La distribución geográfica "costa-interior" observada en la mortalidad por cáncer de estómago es muy peculiar y no se presenta para otras causas de mortalidad. Este patrón característico también ha sido descrito para el conjunto de España y se repite al estudiar zonas más pequeñas, como en nuestro caso. En este estudio, la mortalidad en la zona litoral es inferior a la de las zonas del interior y norte de Cataluña, de forma muy consistente para ambos sexos. Destaca, además, que este patrón se ha ido difuminando con el paso del tiempo, y aunque persisten algunas zonas con una mortalidad superior a la del conjunto, se muestra claramente una tendencia hacia la uniformidad. Esta diferencia puede ser debida a variaciones en los hábitos alimenticios entre la población de la zona costera (más urbana) y la de la zona interior (más rural) asociados al cáncer de estómago (o a la protección contra él), así como una menor accesibilidad a las técnicas que facilitan el diagnóstico (endoscopia digestiva alta para procesos sintomáticos) y al tratamiento, en las zonas más alejadas de los hospitales con mayor tecnología. Por otra parte, la generalización de los hábitos dietéticos así como un mejor acceso a la asistencia sanitaria en todo el territorio, muy probablemente han contribuido a la reducción y homogeneización observada en la mortalidad.

Que la mortalidad por cáncer de estómago haya disminuido de forma importante en casi todo el territorio de Cataluña es una muy buena noticia. No obstante, la

persistencia todavía de algunas zonas con una mortalidad superior debe ser una llamada de atención para los servicios asistenciales y de planificación. La modificación del patrón de mortalidad en otras áreas geográficas significa que en las zonas de mortalidad alta existen espacios donde es posible la intervención, tanto desde el punto de vista asistencial como desde el punto de vista de la prevención.

(Proyecto de la Licenciatura de Estadística presentado en junio de 2003 con el título "Anàlisi de la mortalitat a Catalunya, 1986-2000. Evolució temporal i distribució espacial")

Es indudable el interés de este tipo de estudios. Muy a menudo es la estadística la que va abriendo el camino a los avances en el mundo de medicina y la salud pública, la que dice hacia donde hay que mirar, donde puede estar la causa...

A partir de los resultados de este trabajo también se han publicado artículos en revistas especializadas:

- *Xavier Puig; Rosa Gispert; Josep Ginebra; Josep Bisbe. (2006): "Mortalidad por cáncer de estómago en Cataluña: distribución geográfica y evolución temporal entre 1986 y 2000". Medicina Clínica , 126 (13). Páginas 481-484*
- *Xavier Puig; Rosa Gispert; Josep Ginebra. (2005). "Estudio de la evolución temporal de la mortalidad mediante modelos lineales generalizados". Gaceta Sanitaria , 19 (6). Páginas 481-485.*

Xavier Puig cursó la Diplomatura de Estadística en la Universidad Autónoma de Barcelona y después realizó la Licenciatura en la UPC. Mientras estudiaba la licenciatura trabajó como becario en el Instituto Catalán de Oncología y después se incorporó al Departamento de Salud de la Generalitat de Cataluña. Dice que llegó a la estadística por casualidad, y que hasta el tercer año de la diplomatura no se dio cuenta de que le gustaba. Ahora es su profesión, enseña a otros y continua estudiando. Es profesor de estadística en la UPC y está realizando su tesis doctoral.

13

Expectativas de complicaciones postoperatorias en función de las características del paciente

Proyecto realizado por**: Zahara Marina Briones Carrió**
Director**: Jaume Canet i Capeta**; Tutor**: Erik Cobo Valeri**

Las complicaciones postoperatorias son uno de los principales problemas que pueden presentar las personas sometidas a intervenciones quirúrgicas. Su incidencia es elevada y son la causa de estancias hospitalarias prolongadas y de un aumento de la mortalidad.

Evidentemente, es muy importante conocer a priori cuál es la probabilidad de que se presenten estas complicaciones. Hay evidencia de que esta probabilidad está relacionada con el estado de salud del paciente y con el procedimiento quirúrgico y anestésico realizado. Con respecto al estado de salud, diferentes estudios muestran que pacientes con la llamada "enfermedad pulmonar obstructiva crónica" tienen un riesgo más elevado de sufrir complicaciones postoperatorias. Pero los estudios realizados hasta ahora tienen algunas limitaciones (grupos reducidos, sólo para determinados tipos de cirugía, datos de un solo hospital...) y la mayoría se han realizado en áreas anglosajonas. Este trabajo presenta un estudio a gran escala realizado en el marco del proyecto ARISCAT, patrocinado por la Sociedad Catalana de Anestesiología, que pretende dar respuesta a cuáles son los factores de riesgo en la población de nuestro entorno.

Introducción al problema médico

La enfermedad pulmonar obstructiva crónica (EPOC) consiste en un trastorno pulmonar crónico que produce un bloqueo del flujo de aire a los pulmones. Generalmente, la lesión que causa esta enfermedad es permanente e irreversible. Los síntomas que presenta un enfermo de EPOC son: dificultad respiratoria que persiste meses o años, sibilancias, disminución de la tolerancia al ejercicio y tos. Además, si se combina con un consumo habitual y prolongado de tabaco ocasiona la inflamación del pulmón y la destrucción de los alveolos pulmonares (en los alveolos es donde tiene lugar el intercambio gaseoso entre el aire inspirado y la sangre).

En Cataluña el hábito de fumar ha disminuido entre los hombres pero ha aumentado entre las mujeres y la gente joven. En el año 1998 se atribuyeron 55.613 muertes al hábito de fumar en toda España, y 2.205 solo en Barcelona, representando el 13,8% de muertes en la población mayor de 35 años. Respecto a la población quirúrgica, el hábito de fumar incrementa el riesgo de complicaciones postoperatorias respiratorias (CPR), incluso en aquellos pacientes que no sufren EPOC. El tabaquismo activo está asociado con un riesgo de CPR aproximadamente 6 veces mayor.

Figura 1: *Paciente sometido a anestesia con respiración asistida*

Cada año aumenta el número de pacientes con EPOC y/o hábito tabáquico que se someten a cirugía. La EPOC incrementa las complicaciones respiratorias, pero falta información sobre el perfil epidemiológico, los factores de riesgo, la calidad de vida y las complicaciones respiratorias en este tipo de pacientes.

Objetivo del estudio

El objetivo principal del estudio es identificar y cuantificar los factores de riesgo de las complicaciones postoperatorias respiratorias en pacientes que se someten a una intervención quirúrgica. Conocer cuáles son estos factores permitirá al equipo médico establecer estrategias específicas en función del tipo de paciente para reducir el riesgo de dichas complicaciones.

Este estudio ha sido realizado en el marco del proyecto ARISCAT, promovido por la Sociedad Catalana de Anestesiología, y ha contado con financiación de la Maratón de TV3, que dedicó su programa de 2003 a las enfermedades crónicas respiratorias.

Planificación y recogida de los datos

A través de los hospitales participantes en el estudio se han recogido datos de una muestra de los pacientes sometidos a intervenciones quirúrgicas, así como el tipo de intervención y el tipo de anestesia que se les aplicó. También se ha recogido información sobre su estado de salud a los 3 meses de la intervención.

Ámbito del estudio

El ámbito geográfico de estudio es toda Cataluña. Inicialmente se contó con la participación voluntaria de 63 centros hospitalarios, pero finalmente 4 de ellos no pudieron participar por problemas organizativos. El Hospital Clínico Universitario de Valencia mostró interés en el estudio y dado que pertenece también a la región Mediterránea, se supuso que su población no difería de la del resto de centros universitarios participantes, razón por la cual también fue incluido. En total, el número de hospitales participantes son los que se indican en la Tabla 1.

Tabla 1: *Número de hospitales participantes en el estudio*

Zona	Hospitales participantes en el estudio
Girona (provincia)	10
Lleida (provincia)	4
Tarragona (provincia)	7
Comarcas de Barcelona	22
Barcelona ciudad	15
Valencia ciudad	1
TOTAL	59

Selección de la muestra de pacientes

La población objetivo son los pacientes a los cuales se realizó algún tipo de intervención quirúrgica, bajo anestesia general o loco-regional[1] en los hospitales participantes y que no cumpliera ninguno de los siguientes criterios de exclusión:

- Edad inferior a 18 años.
- Intervención relacionada con embarazo o parto.
- Intervención realizada con bloqueo periférico o anestesia tópica y/o sedación; es decir, aquellos procedimientos realizados fuera de un quirófano.
- Reintervención en el ingreso (cuándo el tiempo entre el alta hospitalaria correspondiente a la intervención anterior es inferior a 30 días)
- Paciente intubado a su llegada al quirófano
- Trasplante de órganos.

El periodo de estudio se fijó en un año debido a que se creía que algunas variables podían presentar un patrón estacional (comportamiento específico en ciertas épocas del año, o en determinados días de la semana). Para seleccionar la muestra de pacientes de cada hospital se asignaron aleatoriamente 7 días en cada uno de ellos, distribuidos entre el 10/01/2006 y el 09/01/2007, uno por cada día de la semana (no consecutivos). La muestra aportada por cada hospital fueron los pacientes intervenidos en estos 7 días, que no cumplían los criterios de exclusión y que aceptaron participar en el estudio. Con este procedimiento se garantizaba una representación adecuada de todos los días de la semana y una representación equilibrada de procedimientos urgentes y programados.

En total, el número de pacientes intervenidos quirúrgicamente en los centros hospitalarios los días asignados fue 7.571. De este total, se excluyeron 4.221 por no cumplir alguno de los criterios de exclusión definidos. Finalmente, de los 3.350 pacientes de estudio potenciales un 10,7% rechazaron participar, por lo cual la muestra constó finalmente de 2.991 individuos, de los cuales se consiguió realizar el seguimiento completo a 2.954 (Figura 2).

Estudiando las características de los pacientes que rechazaron participar en el estudio se observa que el rechazo entre pacientes fumadores es un 6,12% superior al de no fumadores, mientras que la proporción de fumadores en la muestra es inferior a la de no fumadores. También se pone de manifiesto que existen diferencias con respecto a

[1] Inyección de anestésicos en la proximidad de un nervio o de la columna vertebral para evitar que se sienta dolor en la región donde se realiza la intervención quirúrgica.

la edad: en promedio los pacientes estudiados son 2,47 años más jóvenes que los que han rechazado participar. Este hecho se da probablemente porque los pacientes de edades más avanzadas presentan problemas de comunicación para firmar el consentimiento. Con respecto a las intervenciones, a los pacientes que aceptaron participar en el estudio se los realizaron un 19,34% menos de intervenciones urgentes y un 4,64% más de cirugía programada.

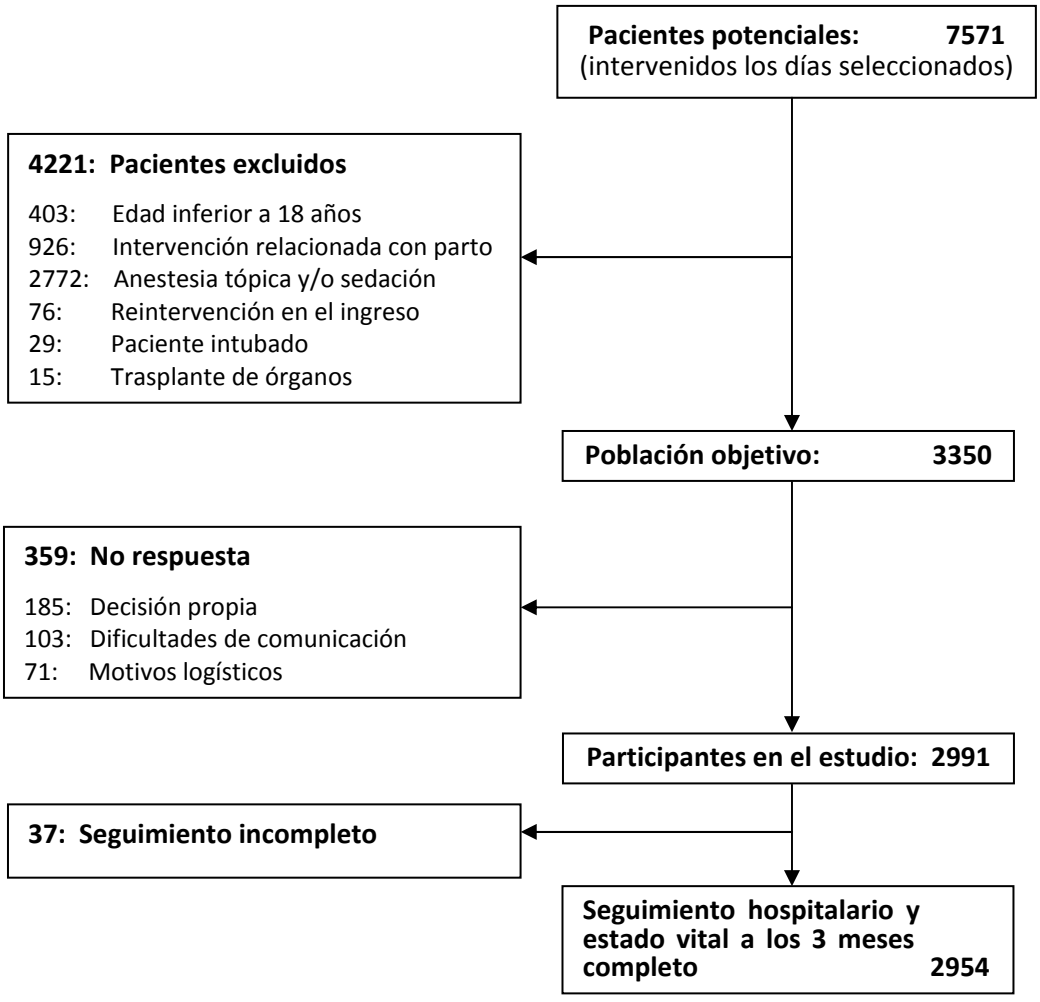


Figura 2*: Diagrama de flujo de la selección de la muestra, indicando los motivos de no aceptación*

Recogida de los datos

El seguimiento de los pacientes incluidos en el estudio tuvo una duración de tres meses desde la fecha de intervención, a excepción de los pacientes ingresados durante más de tres meses que fueron seguidos hasta el momento en que recibieron el alta hospitalaria.

La recogida de los datos durante la estancia en el hospital la realizaron los anestesistas de cada centro. Una vez obtenida el alta hospitalaria se continuó recogiendo información mediante un cuestionario realizado telefónicamente. En este aspecto fue muy importante la colaboración de los pacientes, que raramente rechazaron participar. Contactar con los pacientes comportó la realización de más de 6.500 llamadas telefónicas. En el caso de no poder contactar a través de esta vía se les envió un cuestionario en el cual se incluía un sobre franqueado. Por esta vía se consiguió localizar a bastantes pacientes.

Los datos recogidos se pueden clasificar en variables predictoras y variables respuesta. Las variables predictoras hacen referencia al estado de salud previo a la intervención así como a las características de la intervención y al tipo de anestesia utilizada. Los estudios similares ya realizados anteriormente y la experiencia médica de los investigadores permitieron seleccionar aquellas variables que tienen relevancia clínica en este estudio. La Tabla 2 contiene el listado de estas variables.

La principal variable respuesta es la presencia (1) o ausencia (0) de complicaciones postoperatorias respiratorias (CPR). Aunque no formaba parte de los objetivos de este estudio, para un posible análisis posterior se recogió también información sobre otros tipos de complicaciones además de las respiratorias (cardiovasculares, de herida quirúrgica, y por otras causas).

Control de calidad del proceso de recogida de datos

Los datos se recogieron a través de formularios estructurados en una aplicación vía web y mediante ciertas reglas de validación se intentó evitar la presencia de datos incoherentes o faltantes.

Aun así, es inevitable que algún dato contenga un valor anómalo. Por ello se evaluó la calidad de los datos realizando una auditoría a 150 pacientes, que equivalen aproximadamente al 5% del total. Se encontraron 379 errores en algunas de las 130 variables auditadas (la mayoría relacionados con la duración de la intervención), los cuales representan un 1,94% de errores en el total de variables auditadas:

$$\frac{379\ errores}{130\ variables \cdot 150\ individuos} \cdot 100 = 1{,}94\%$$

También se auditaron los centros hospitalarios: para cada uno de ellos se comprobaron dos hojas de programación de quirófano para verificar que la selección de pacientes se había realizado de la forma correcta. Se concluyó que en los centros auditados los criterios de elegibilidad se aplicaron de la forma establecida.

Tabla 2: *Variables predictoras incluidas en el estudio*

Variable	Tipo variable	Valores
Sexo	Cualitativa	H: Hombre, M: Mujer
Edad	Cuantitativa	18, 19, 20, ...
Dependencia funcional	Cualitativa	1: Independiente, 2: Dependiente
Tipo de fumador	Cualitativa	1: No fumador, 2: Ex fumador, 3: Fumador
Paquetes/año	Cuantitativa	1, 2, 3, ...
Test de la tos	Cualitativa	1: Positivo, 2: Negativo
Síntomas respiratorios	Cuantitativa	A partir de cuestionario del *British Medical Council*
MPOC	Cualitativa	0: Ausencia, 1: Presencia
Asma	Cualitativa	0: Ausencia, 1: Presencia
Otras enfermedades respiratorias	Cualitativa	0: Ausencia, 1: Presencia
PRAUM	Cualitativa	0: Ausencia, 1: Presencia
Historia de ronquidos	Cualitativa	0: Ausencia, 1: Presencia
Historia de somnolencia	Cualitativa	0: Ausencia, 1: Presencia
Diagnóstico oncológico	Cualitativa	0: Ausencia, 1: Presencia
Obesidad	Cualitativa	0: Ausencia, 1: Presencia
Patología cardiovascular	Cualitativa	0: Ausencia, 1: Presencia
Patología neurológica	Cualitativa	0: Ausencia, 1: Presencia
Hepatopatía	Cualitativa	0: Ausencia, 1: Presencia
Insuficiencia renal	Cualitativa	0: Ausencia, 1: Presencia
Diabetes	Cualitativa	0: Ausencia, 1: Presencia
Inmunodepresión	Cualitativa	0: Ausencia, 1: Presencia
Sepsis	Cualitativa	0: Ausencia, 1: Presencia
Alcoholismo	Cualitativa	0: Ausencia, 1: Presencia
Anemia	Cualitativa	0: Ausencia, 1: Presencia
Sonda nasogástrica	Cualitativa	0: Ausencia, 1: Presencia
Intervención urgente	Cualitativa	U: Urgente, P: Programada
Anestesia general-combinada	Cualitativa	1: Neuroaxial, 2: General – Combinada
Incisión	Cualitativa	1: Periférica, 2: Abdominal sup. 3: Intratorácica
Tipo de incisión	Cualitativa	1: Cerrada, 2: Abierta
Escala agresividad quirúrgica	Cuantitativa	1 (menor) - 5 (mayor)

Análisis exploratorio de los datos

Un primer análisis exploratorio de los datos ya permite observar que el porcentaje de pacientes con complicaciones, tanto durante la operación (intraoperatorias) como posteriores (postoperatorias), es mayor para el grupo de pacientes que sufren EPOC. En particular, para las complicaciones postoperatorias respiratorias, la diferencia entre los dos tipos de pacientes es muy clara (Figura 3).

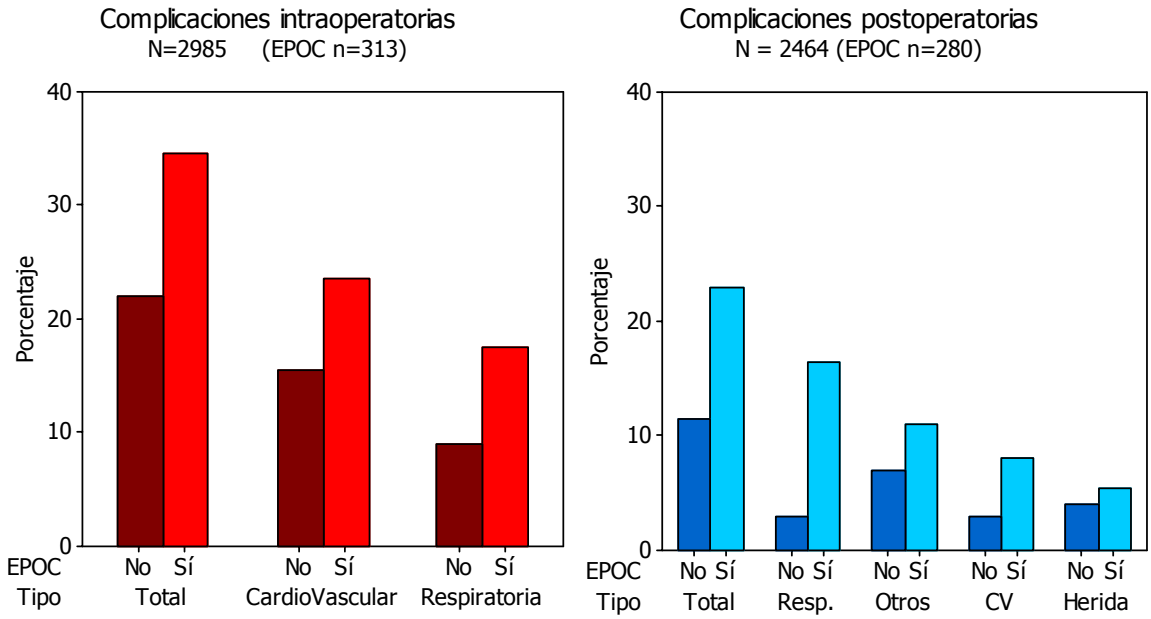


Figura 3: *Porcentaje de casos con complicaciones intraoperatorias y postoperatorias según se padezca o no EPOC*

Construcción del modelo de previsión

La identificación y cuantificación de los factores de riesgo para las complicaciones postoperatorias de tipo respiratorio se realiza mediante un modelo (una expresión matemática) que da la probabilidad de sufrir este tipo de complicación en función del valor que tienen las variables que le afectan.

Cuando la respuesta es una probabilidad, la forma más habitual de modelarla es a través de la llamada regresión logística. Se trata de ajustar un modelo de regresión lineal en el cual la respuesta es lo que se llama un *log-odd*. Un *odd* es el cociente entre la probabilidad de que suceda un acontecimiento y la probabilidad de que no suceda. Una vez se saca el logaritmo a un *odd*, se obtiene un *log-odd*.

Este modelo se escribe como:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Donde p es la probabilidad de tener complicaciones postoperatorias; $x_1, x_2, \cdots, x_n$ representan los valores que toman los factores de riesgo, y $\beta_1, \beta_2, \cdots, \beta_n$ son los coeficientes que acompañan a estos factores y cuantifican su influencia. β_0 es un término independiente que permite calcular la respuesta cuando todos el factores de riesgo toman un valor igual a cero.

A partir de la expresión anterior se puede aislar fácilmente el valor de p, obteniéndose:

$$\left(\frac{p}{1-p}\right) = exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)$$

$$p = \frac{exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}{1 + exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}$$

A partir de los datos obtenidos y aplicando las técnicas de selección de modelos de regresión, se llega a un modelo para el *log-odd* en el que aparecen las variables regresoras y los coeficientes que se presentan en la Tabla 3 (para construir el modelo sólo hemos utilizado una parte de los datos, como se comenta más adelante).

Tabla 3: *Variables que entran en el modelo y coeficientes estimados para cada una de ellas*

Variables que entran en el modelo	Designación en el modelo	Valores que puede tomar la variable	Coeficientes estimados
EPOC	EPOC	0: No; 1: Sí	1,4130
Diagnóstico oncológico	DO	0: No; 1: Sí	0,8961
Sexo	Sexo	0: Mujer; 1: Hombre	0,8954
Edad	Edad	Edad en años	0,0262
PRAUM*	PRAUM	0: No; 1: Sí	1,3299
Intervención urgente	IU	0: No; 1: Sí	1,4185
Incisión abdominal superior	IAS	0: No; 1: Sí	0,9148
Incisión intratorácica	IIN	0: No; 1: Sí	1,7172
EAQ**	EAQ	De 1 (menor) a 5 (mayor)	0,6692

* Proceso respiratorio agudo durante el último mes
** Escala de agresividad quirúrgica

El término independiente es $\beta_0 = -8{,}2905$, y el modelo puede escribirse de la forma:

$$\ln\left(\frac{p}{1-p}\right) = -8{,}2905 + 1{,}413\ EPOC + 0{,}8961\ DO + 0{,}8954\ Sexo + 0{,}0262\ Edad + 1{,}3299\ PRAUM + 1{,}4185\ IU + 0{,}9148\ IAS + 1{,}1772\ IIN + 0{,}6692\ EAQ$$

Podemos considerar que con EPOC = 0 y con unos valores determinados de las variables regresoras el *log-odd* toma un determinado valor k, es decir:

$$\ln\left(\frac{p}{1-p}\right) = k$$

Si EPOC =1, manteniendo los valores del resto de variables regresoras, tendremos:

$$\ln\frac{p}{1-p} = k + 1{,}413$$

O también:

$$\frac{p}{1-p} = e^{k} \cdot e^{1{,}413}$$

Si con EPOC = 0 tenemos p = 0,5, es fácil comprobar que con EPOC = 1 el valor de p pasará a ser 0,8.

Respecto a la edad, aumentarla en un año manteniendo constantes el resto de variables, multiplica los *odds* por $e^{0{,}0262} = 1{,}0265$; es decir, la probabilidad de sufrir una complicación postoperatoria respiratoria (CPR) aumenta un 2,65% por cada año de diferencia respecto al nivel de partida (18 años).

Validación del modelo

A partir de la modelización lo que interesa es conocer si un paciente con determinados valores de los factores de riesgo sufrirá o no complicaciones respiratorias. La previsión se hace en función del valor que da el modelo para la variable respuesta. Por debajo de un cierto umbral se considera que el paciente no tendrá complicaciones, y por encima se considera que sí las tendrá.

Muestra de aprendizaje y muestra de validación

Como se ha dicho anteriormente, el modelo se construye sólo con una parte de los datos disponibles. Normalmente se utilizan dos tercios de los datos para construir el modelo (muestra de aprendizaje) y el otro tercio se reserva para comprobar si da buenos resultados (muestra de validación). En nuestro caso, la muestra de aprendizaje contiene 1.628 pacientes (66%) y la de validación 836 (34%).

Sensibilidad y especificidad del modelo

La muestra de validación se puede utilizar para construir una tabla (Figura 4) con las categorías cruzadas de las respuestas observadas en los pacientes (ha tenido o no ha tenido complicaciones) y de las respuestas previstas por el modelo en función del valores de los factores de riesgo.

		El modelo prevé que…	
		Tendrá complicaciones	No tendrá complicaciones
Realmente…	Ha tenido complicaciones	Verdadero positivo VP	Falso negativo FN
	No ha tenido complicaciones	Falso positivo FP	Verdadero negativo VN

Figura 4: *Clasificación de los valores previstos en función de los observados*

A partir de los valores de esta tabla se pueden definir los siguientes indicadores, referidos a la capacidad de predicción del modelo:

- *Sensibilidad*: Proporción de verdaderos positivos que son estimados positivos

$$Se = \frac{VP}{VP + FN}$$

- *Especificidad*: Proporción de verdaderos negativos que son estimados negativos

$$Sp = \frac{VN}{VN + FP}$$

Hay que tener en cuenta que el valor obtenido para cada uno de estos indicadores dependerá del umbral escogido para la asignación de la respuesta como positiva (tendrá complicaciones) o negativa (no las tendrá). Por ejemplo, si se utiliza un umbral muy bajo la respuesta casi siempre será positiva y, por lo tanto, la sensibilidad será alta (cosa que interesa) pero la especificidad será baja. Por otra parte, si el umbral es alto tendremos una sensibilidad baja y una especificidad alta.

La relación entre sensibilidad y especificidad en función del umbral se representa mediante la llamada curva ROC (del inglés *Receiver Operating Characteristic*). En la Figura 5 podemos encontrar tres situaciones posibles de curva ROC:

a) La curva ROC está formada por dos rectas, una siguiendo el eje de las x y la otra el de las y. Esta curva contiene un punto (extremo superior izquierdo) que da una sensibilidad del 100% y una especificidad también del 100% (1 - especificidad = 0) representado, por lo tanto, una situación de discriminación perfecta.

b) La curva representa una situación realista en la que se puede escoger un umbral que combine un buen valor tanto para la sensibilidad como para la especificidad.

c) Tenemos una situación en que no se puede encontrar un umbral que combine buenos valores de ambas características. Esta situación corresponde a un modelo que no tiene ningún poder de discriminación.

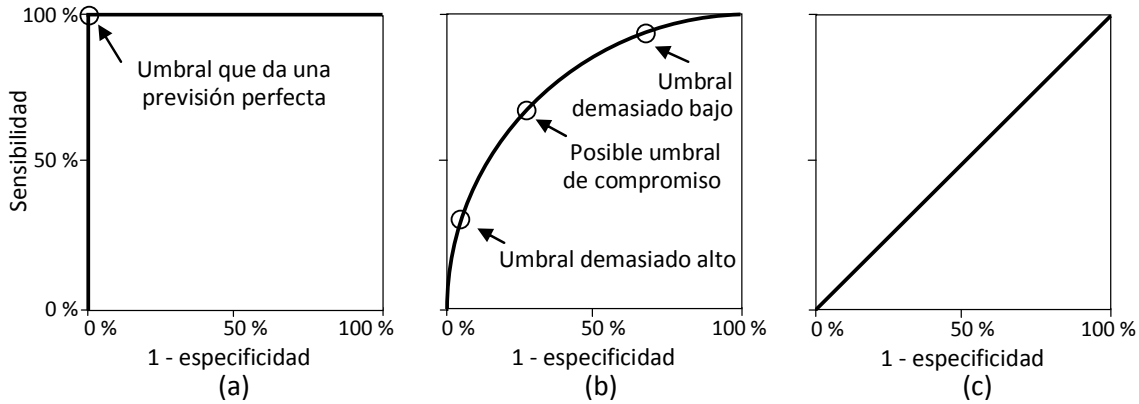


Figura 5: *Diferentes formas de la curva ROC: (a) Situación ideal que permite escoger un umbral que discrimina perfectamente, (b) situación realista y (c) situación en que el modelo no es útil*

El área que queda bajo la curva ROC es una medida de la capacidad predictiva del modelo. El valor máximo es 1, caso de la Figura 5 a) y el mínimo es 0,5, Figura 5 c). Si el área tiene un valor superior a 0,8 se considera que el modelo tiene una buena capacidad de predicción.

El modelo seleccionado da un área bajo la curva de 0,89. Se han probado también otros tipos de modelo pero tienen una capacidad predictiva menor, o la tienen ligeramente superior pero a costa de complicar mucho el modelo y se ha considerado que no vale la pena introducir esas complicaciones.

Conclusiones

Los resultados de la modelización han permitido encontrar un conjunto de parámetros útiles para caracterizar a los pacientes con complicaciones postoperatorias respiratorias (CPR) y poder predecir su aparición a partir de ellos. La capacidad predictiva del modelo se ha contrastado satisfactoriamente en una muestra independiente (muestra de validación).

El modelo resultante tiene como predictores tanto variables referentes a la intervención, como al estado de salud y a otras características del paciente. Las variables que influyen en las CPR referentes al paciente son el sexo y la edad; con respecto al estado de salud previo influyen la EPOC, el diagnóstico oncológico y haber sufrido un proceso respiratorio agudo durante el último mes. Finalmente, sobre las variables relacionadas con la intervención los predictores incluidos en el modelo son la urgencia de la intervención, el tipo de incisión y la escala de agresividad quirúrgica.

Hay que tener presente que la población de estudio es muy diferente a la de estudios previos, por lo cual es difícil comparar los resultados. Aun así, algunos de los factores predictivos identificados ya estaban descritos previamente en la literatura científica. La principal novedad que aporta este modelo es la aparición del proceso respiratorio agudo durante el último mes como factor predictivo de complicaciones postoperatorias respiratorias en la población adulta catalana.

(Proyecto de la Licenciatura de Estadística presentado en febrero de 2008 con el título: "Factores predictivos de complicaciones postoperatorias respiratorias en la población quirúrgica: un estudio de Cohortes (ARISCAT 2006)")

Diseñar la recogida y el análisis de los datos de forma que permita obtener la información que interesa siempre es importante. Y todavía lo es más cuando a partir de esta información se tomarán decisiones de las que puede depender la vida de una persona.

Zahara Briones obtuvo la Licenciatura en Matemáticas en la Universidad de Barcelona y la Licenciatura en Estadística en la UPC. Mientras estudiaba matemáticas realizó una estancia como estudiante Erasmus en la Universidad de Bielefeld (Alemania) y mientras estudiaba en la UPC trabajó con un convenio de cooperación educativa en Editorial Planeta. Su interés por la estadística recae principalmente en la bioestadística, motivo por el cual trabajó en el estudio ARISCAT. Actualmente está trabajando en Viena en una gran multinacional dedicada a los seguros. Entre muchas otras habilidades, Zahara habla alemán perfectamente.

Prensa y televisión

14

Análisis de patrones y tendencias en las votaciones del festival de Eurovisión

Proyecto realizado por: **Laura Marí Tomàs**
Dirigido por: **Lluís Marco Almagro**

Cada mes de mayo se celebra el festival de la canción de Eurovisión. Aunque hay gente que lo considera un programa kitsch y con poco interés, pocos acontecimientos no deportivos reúnen a tantos millones de personas cada año delante de la televisión.

El festival es un filón de frases tópicas que se pueden ir repitiendo mientras se mira el programa: "ganará Suecia, porque los países nórdicos se votan entre ellos", "la canción de Francia es buena, pero no ganará porque ha actuado en tercer lugar, y la gente no vota las primeras canciones porque no las recuerda", "para ganar es mejor que te voten muchos países, aunque te den pocos puntos"... Sin embargo, ¿qué hay de cierto en todo eso?

La estadística se ocupa de responder preguntas, pero basándose en datos, y no en opiniones poco fundamentadas. Por eso nos hacía ilusión mirar las puntuaciones de unos cuantos años del festival de Eurovisión (o sea, utilizar los datos) y, mediante herramientas estadísticas descriptivas (básicamente gráficos que estuvieran bien pensados), comprobar si algunos de estos tópicos tan reiterados son ciertos o no. También se nos ocurrió alguna cosa un poco más sofisticada, como simular lo que llamamos 'festival aleatorio' para compararlo con el real... Pero vayamos por partes.

Un poco de historia y nuestra materia prima: ¡los datos!

El festival de Eurovisión nació inspirado en el festival italiano de la canción de San Remo y su primera edición se celebró en Lugano, Suiza, en 1956. Durante las primeras dos décadas el número de países participantes se fue incrementando (España hizo su debut en el año 1961) y el sistema de puntuaciones se modificó unas cuantas veces. En el año 1975 se instauró el sistema de votaciones que todavía se utiliza: cada país distribuye puntos del 1 al 8, 10 y 12, de manera que vota a 10 canciones diferentes. Dado que el número de países participantes crecía mucho y hacía el festival demasiado largo, a partir del año 2004 se incorporó una semifinal en la cual actúan unos 20 países. De éstos, aproximadamente la mitad pasan a la final, donde actúan entre 22 y 24. De hecho, en el año 2008 se añadió una segunda semifinal.

Este estudio se ha centrado en el periodo comprendido entre los años 1975 y 2003: en todo este periodo el sistema de puntuaciones y la mecánica del concurso se ha mantenido estable. Son 29 años, suficientes para comprobar si son ciertos o no algunos de los tópicos, y para detectar patrones de comportamiento entre países. Ciertamente, ha habido algunas innovaciones, como la sustitución progresiva del jurado por el televoto de los espectadores (vía llamadas o mensajes de móvil) a partir de 1996. Se tendrá en cuenta este hecho en el análisis de los datos.

Tabla 1: *Parte de una tabla con las puntuaciones del año 1975*

		País que da los votos						
		Bélgica	*Finlandia*	*Francia*	*Alemania*	*Irlanda*	*Israel*	*Italia*
País que recibe los votos	*Bélgica*	0	0	0	7	0	0	0
	Finlandia	0	0	0	12	5	8	0
	Francia	2	0	0	0	12	7	8
	Alemania	0	0	0	0	0	0	0
	Irlanda	12	1	6	0	0	0	4
	Israel	1	3	1	1	1	0	2
	Italia	6	12	4	4	6	5	0
	Malta	7	0	8	0	0	1	0
	Mónaco	0	2	0	3	0	2	5
	Noruega	0	0	0	0	0	0	7

Alemania da 4 votos a Italia

La obtención de los datos fue más fácil de lo que inicialmente se pensaba. ¡Hay muchas páginas web dedicadas al festival, y es que hay muchos eurofans! En concreto se

extrajeron de la página *esctoday.com*, que tiene tablas como la Tabla 1 para cada año de festival. A partir de estos datos creamos un documento de Excel donde cada pestaña correspondía a un año, desde 1975 hasta el 2003. Había que pensar cómo se podían juntar los datos de todos los años. No es una buena idea sumar simplemente los puntos totales que un país A ha dado a un país B en todos los años, porque no todos los años han participado los mismos países. Si se hace así, hay que esperar que Francia haya dado en total bastantes más puntos al Reino Unido que a Marruecos, porque Francia y Reino Unido han coincidido prácticamente siempre, pero Marruecos sólo participó una vez (en 1980, y sin demasiado éxito).

Por tanto, creamos una nueva tabla que incluía, para cada pareja de países posibles, cuántos años habían coincidido. Finalmente, se dividió el número de puntos totales que un país A ha dado a otro país B entre el número a veces que los dos países han coincidido en el festival. Así tenemos la media de puntos que el país A ha dado al B todos los años que eso ha sido posible. En el caso de países que no han coincidido nunca se ha puesto un asterisco (*). La Tabla 2 muestra una parte de esta tabla.

Tabla 2: *Parte de la tabla con las puntuaciones promedio*

		País que da los votos				
		Croacia	*Chipre*	*Dinamarca*	*Estonia*	*Finlandia*
País que recibe los votos	*Dinamarca*	1,71	1,94	0	5,20	2,56
	Estonia	1,44	1,75	6,40	0	8,40
	Finlandia	0	1,06	0,50	4,40	0
	Francia	1,54	3,1	2,48	3,78	3,87
	Macedonia	6,67	1,67	0	0	0,33
	Alemania	2	2,40	5,18	1,88	2,87
	Grecia	2,22	10,75	1,06	0,86	1,89
	Hungría	1,75	0	0	3	6
	Islandia	1,22	1,64	4,38	3,86	0,73
	Irlanda	3,50	4,74	5,55	4,50	3,86
	Israel	1,50	2,59	2,32	1,83	5,10
	Italia	5	4,60	1	0	7,19
	Letonia	0,50	1,67	5	9,25	9
	Lituania	0,50	1,67	0	1	0
	Luxemburgo	0	2,67	2	*	2,89

Chipre da a Italia,
4,6 votos en promedio

Análisis descriptivo de los patrones de votación

El objetivo del análisis descriptivo de estos datos es obtener información sobre las puntuaciones utilizando herramientas gráficas. Básicamente, se han utilizado diagramas bivariantes ya que permiten averiguar cuál es la relación que existe entre dos variables.

Relaciones entre países

En la Figura 1 cada punto representa la puntuación que, en promedio, un país A ha dado a un país B, en función de las veces que han coincidido[1]. A medida que aumentan los años coincididos la dispersión se va reduciendo. Destacan los dos puntos rodeados con un círculo. Se trata de la pareja de países Grecia y Chipre. Lo que más sorprende de esta pareja es que la media de los puntos que se han dado en los 16 años que han coincidido supera los 10 puntos. O sea, que se votan muchísimo entre ellos.

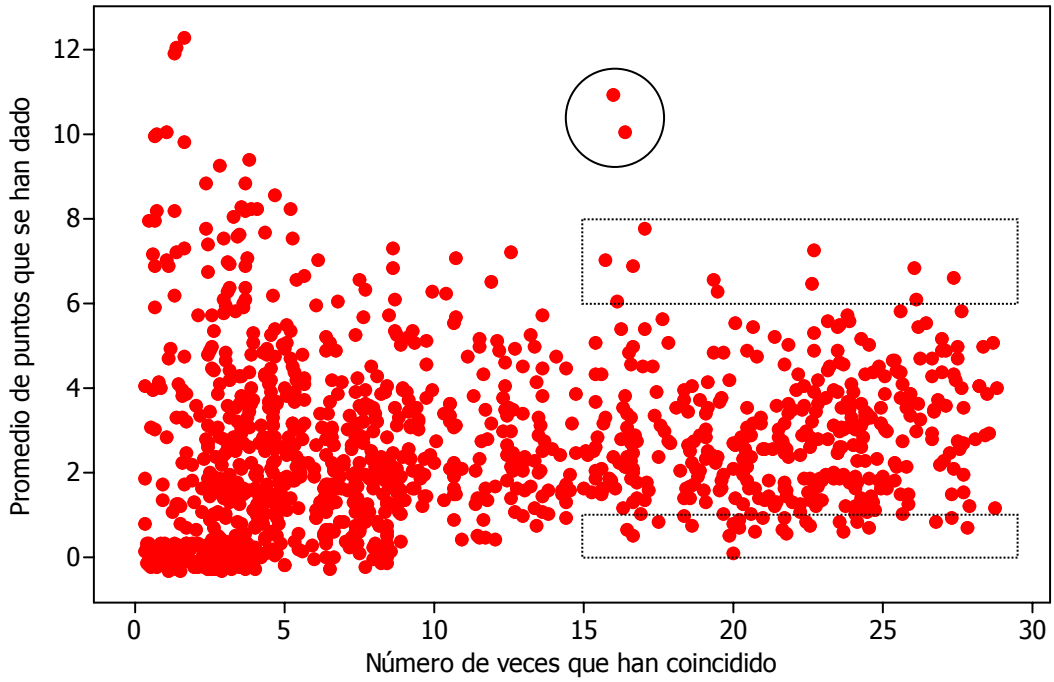


Figura 1: *Diagrama bivariante del promedio de puntos que un país ha dado a otro respecto al número de veces que han coincidido. Los recuadros corresponden a las zonas que se muestran ampliadas en figuras posteriores*

[1] Para evitar que los puntos que tienen las mismas coordenadas queden superpuestos y se pierda la visión de la densidad (cosa que pasa a menudo cuando en los dos ejes las variables son discretas), el programa con que se han realizado estos gráficos (MINITAB) permite activar una opción que mueve ligeramente los puntos para que no queden superpuestos. De esta forma, aunque se pierde precisión en las coordenadas de los puntos, se gana en claridad.

Si nos concentramos sólo en algunas zonas especialmente relevantes del gráfico de la Figura 1 como por ejemplo los países que han coincidido más de 15 veces y que se han dado un promedio de puntos por encima de 6, se puede ver que con promedios por encima de 7 están los puntos que Portugal ha dado a Italia, Finlandia a Italia, España a Italia y Dinamarca a Suecia (ver ampliación, Figura 2). En este gráfico no aparece la relación de amistad entre Grecia y Chipre, dado que la escala vertical de la gráfica sólo llega hasta 8).

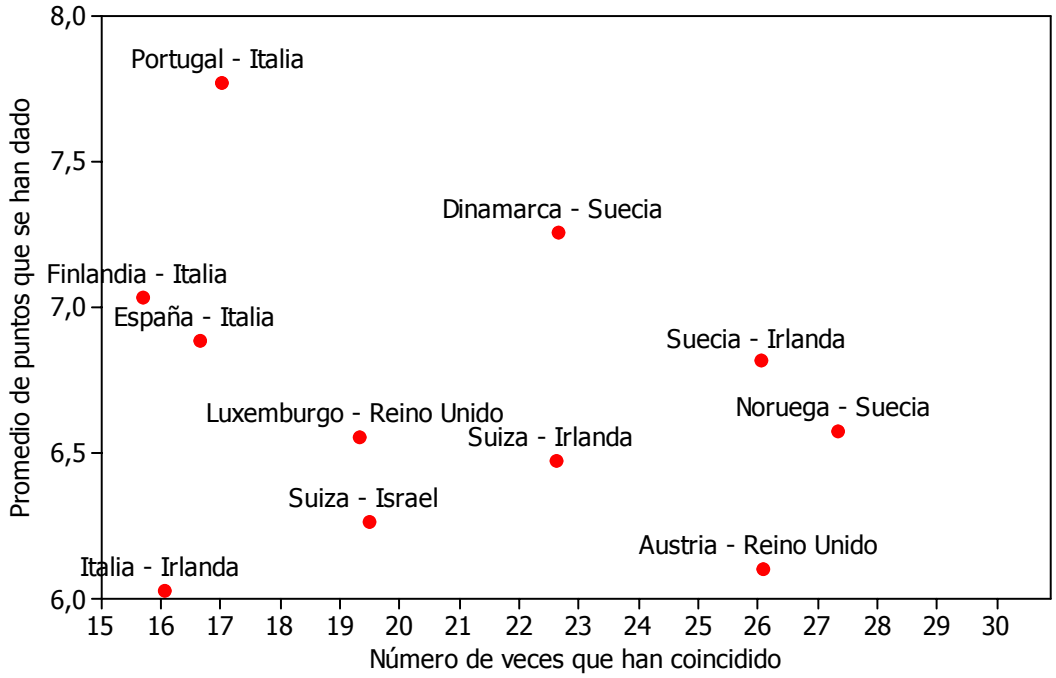


Figura 2: *Diagrama bivariante de la "zona amiga" (zona donde están los países que han coincidido muchas veces y que se han votado también muchas veces)*

Otra forma de mostrar relaciones de amistad entre países es comprobando de qué forma se reparten los votos. Hay 10 tipos diferentes de votos (del 1 al 12, excluyendo el 9 y el 11). Más adelante se verá que los países ganadores obtienen muchas valoraciones de 10 y 12 puntos. Centrémonos, pues, en cómo se reparten los países las puntuaciones de 12. El gráfico de la Figura 3 está construido de la siguiente forma: el grueso de las líneas es proporcional al número a veces que un país ha dado 12 puntos a otro; las líneas negras muestran relaciones en un solo sentido (y el sentido se indica con una flecha), mientras que las líneas rojas denotan relaciones recíprocas. La relación de amistad entre Grecia y Chipre se hace muy evidente. Hay relaciones recíprocas entre Dinamarca y Suecia, Suecia y Noruega, Suecia e Irlanda, Irlanda e Italia, Francia y Portugal y, finalmente, entre Alemania y Turquía. Se podría intentar buscar justificaciones a estas relaciones. La relación entre Alemania y Turquía se podría

justificar por la gran cantidad de inmigración turca que existe en Alemania. Las relaciones entre Suecia y Noruega y entre Suecia y Dinamarca podrían justificarse por la proximidad cultural.

En esta red de la Figura 3 también podemos observar cómo Inglaterra ha sido el país que ha recibido más valoraciones de 12 puntos, ya que la mayoría de flechas le apuntan. En cambio, sólo dos flechas salen de Inglaterra hacia otros dos países (es decir, Inglaterra da de una manera muy repartida sus 12 votos). También Irlanda ha recibido muchas veces los 12 puntos. Irlanda e Inglaterra son los dos países que han ganado más veces el festival, 6 y 4 veces, respectivamente.

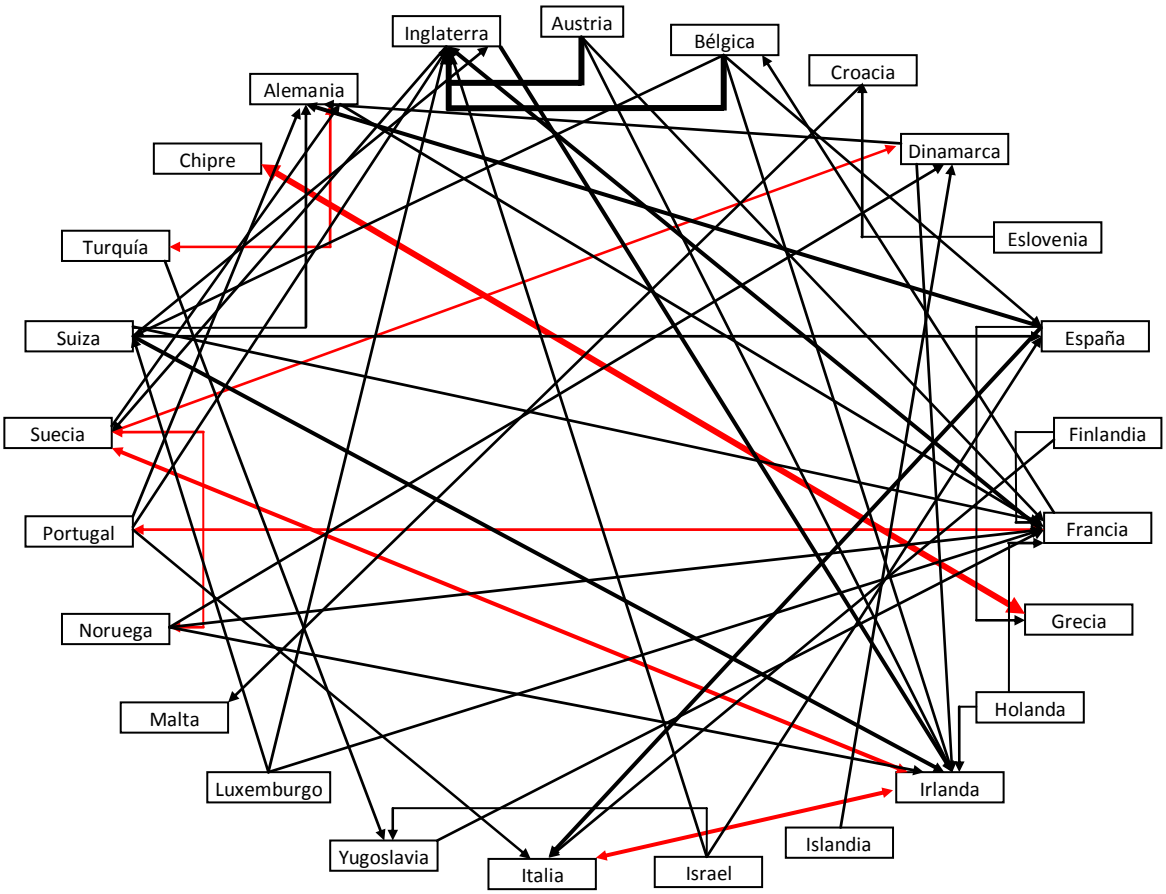


Figura 3: *Número de veces que un país ha dado 12 puntos a otro*

Retomamos de nuevo el gráfico de la Figura 1, pero ahora para destacar las relaciones de enemistad. En la Figura 4 se ha ampliado la parte baja de la Figura 1 mostrando sólo los países que han coincidido más de 15 veces. Se ha rodeado la relación entre Chipre y Turquía, que han participado 20 años conjuntamente, y en cambio no se han votado prácticamente nunca (en realidad ¡Turquía no ha dado nunca ni un solo voto a Chipre!).

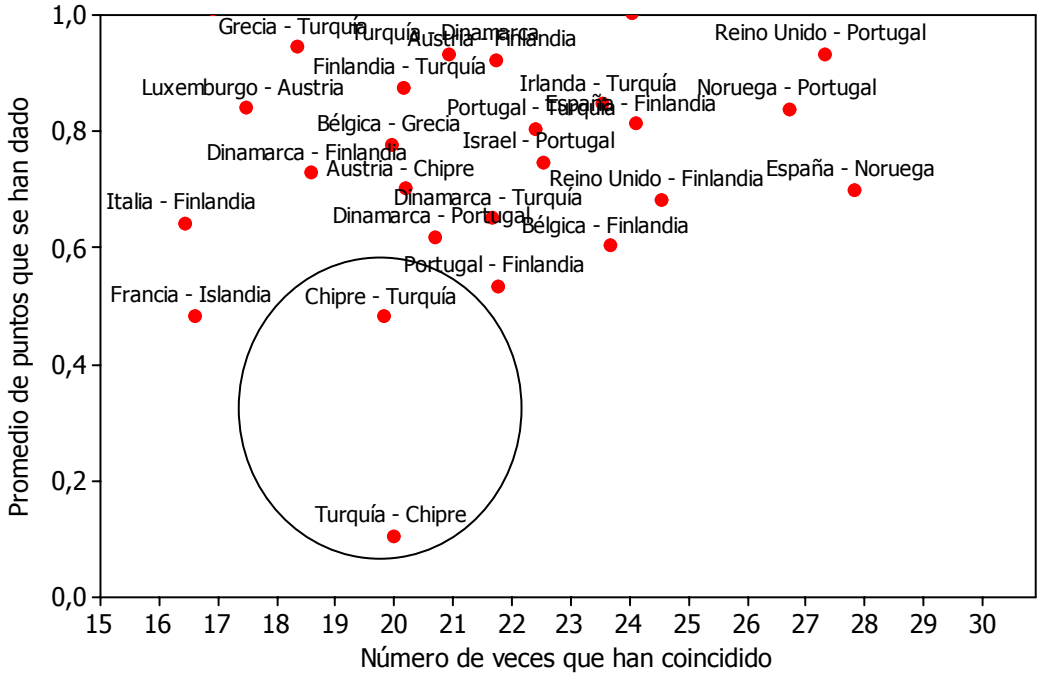


Figura 4: *Chipre y Turquía se dan muy pocos puntos*

La Figura 5 pone de manifiesto como el Reino Unido tiene tendencia a dar pocos votos a España y Portugal. También se observa que España y Reino Unido, Suecia y Noruega, aunque han participado muchas veces juntas (28 o 29 años), se han dado muy pocos puntos. Tampoco se han votado demasiado entre ellos España y Portugal o España y Francia (no parece que España tenga vecinos que la estimen demasiado…).

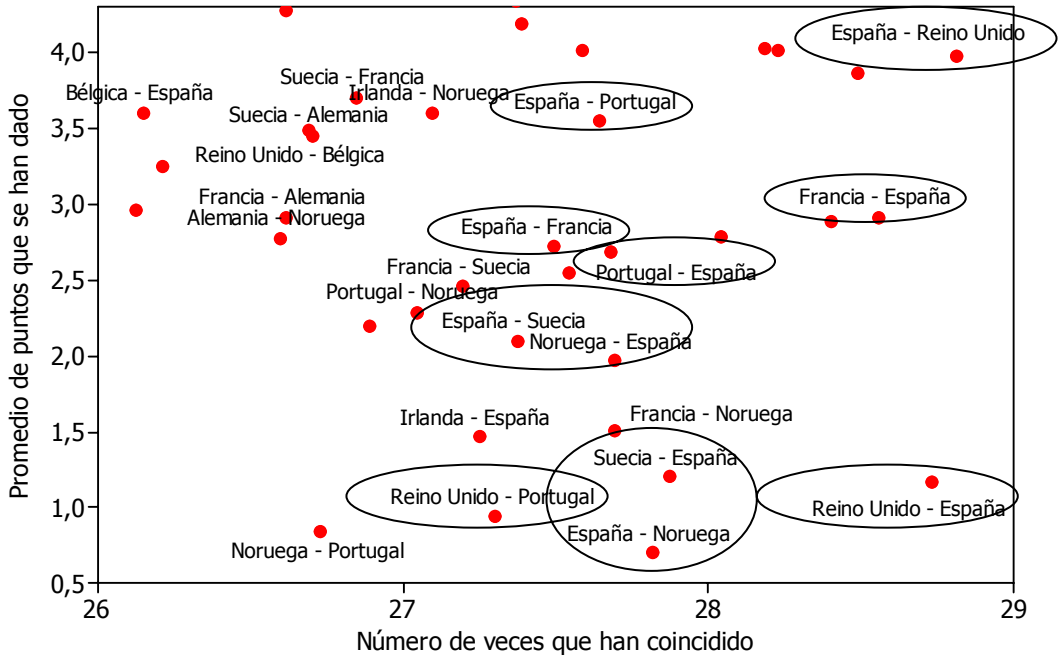


Figura 5: *Países que no se votan mucho entre ellos*

Y para acabar con este tema de la relación entre países, la Figura 6 muestra un diagrama bivariante donde en el eje horizontal se representa el promedio de votos que un país da a otro, y en el eje vertical el promedio de votos que un país recibe de otro. El degradado de colores indica el número de años que cada pareja de países ha coincidido en el Festival.

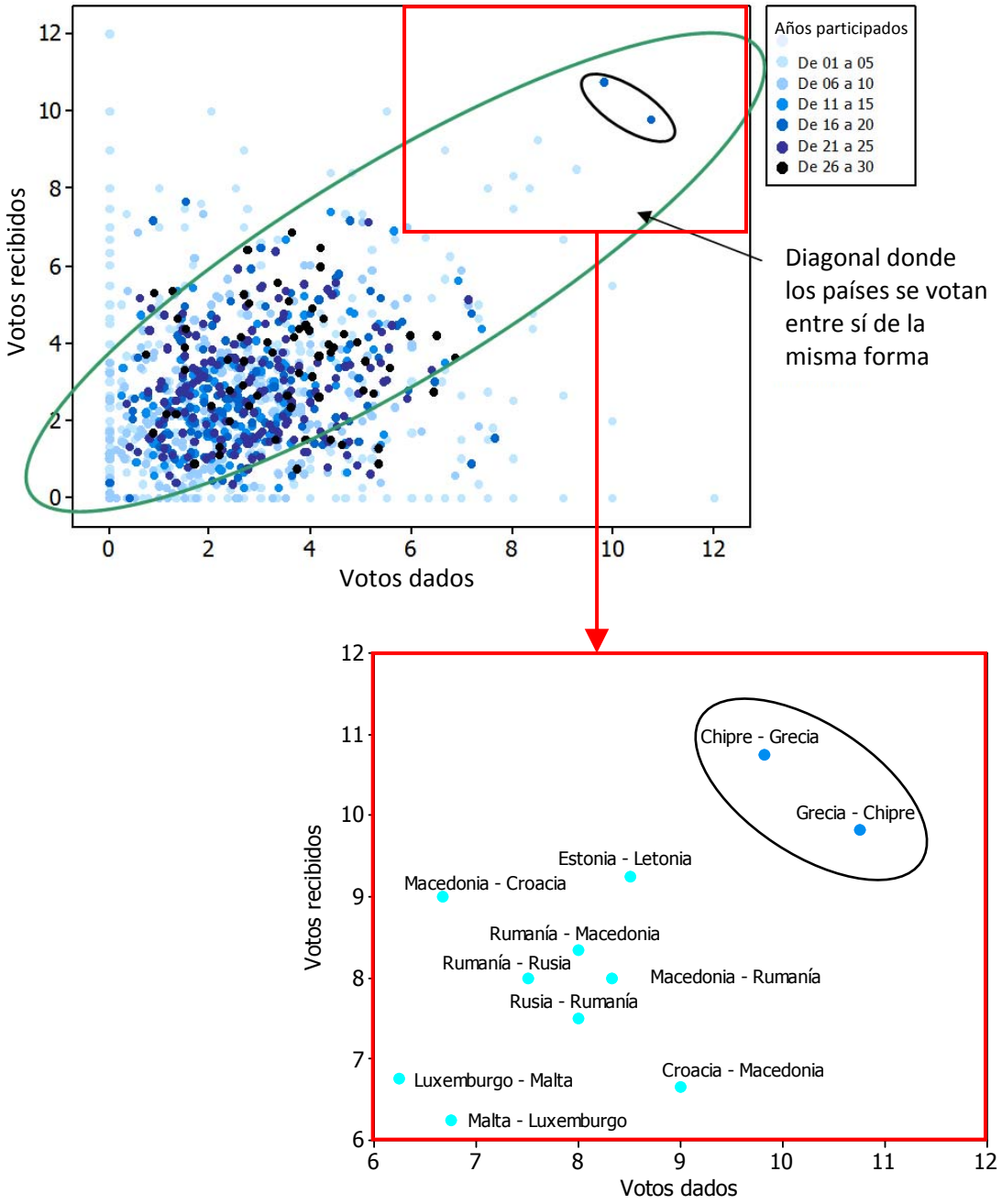


Figura 6: *Diagrama bivariante con los votos dados y recibidos por cada país*

El gráfico es, como se puede ver, simétrico. La pareja enmarcada en negro es la de Grecia y Chipre. La zona de la diagonal son parejas de países que se votan entre ellos y, además, lo hacen de una manera similar. Hemos ampliado y puesto nombres a los países de la zona enmarcada en rojo, donde aparecen los que se han dado más de 6 votos en promedio.

La creación de un festival aleatorio

Recordemos que el objetivo era comprobar si en el festival hay patrones de comportamiento no aleatorio en las puntuaciones que se dan los países. Es decir, ¿hay países que se votan entre ellos más de lo que cabría esperar si las puntuaciones se distribuyeran aleatoriamente? Con el fin de verlo se creó una distribución de referencia de puntuaciones entre países, simulando lo que podrían haber sido las puntuaciones de los festivales analizados si los puntos se hubieran distribuido al azar. Este festival donde los puntos se distribuyen al azar, que se llamó festival aleatorio, se basa en la suposición de que todas las canciones del festival son igual de buenas (o de malas). En este festival aleatorio, en algunas ediciones un país votará mucho a otro, pero no lo hará sistemáticamente. Hasta hace unos años este supuesto no era tan descabellado: los países acostumbraban a presentar canciones muy similares, todas "festivaleras". Actualmente esto ya no es tan cierto, y más adelante se analizará con más profundidad.

Esta idea de simular un festival aleatorio surgió de un artículo sobre Eurovisión publicado en el 2005 por un grupo de físicos en la revista "Physica A: Statistical Mechanics and its Applications". El artículo se titula *How does Europe Make Its Mind Up? Connections, cliques, and compatibility between countries in the Eurovision Song Contest*, y es muy interesante (de hecho, es sorprendente hacer una busca de la palabra "Eurovisión" en bases de datos científicas, y comprobar que se han escrito artículos sobre el festival en revistas de áreas muy diferentes: sociología, psicología, física, matemáticas...).

Cómo el festival se va haciendo cada vez menos aleatorio

Antes de empezar con la simulación del festival aleatorio había que crear un índice que resumiera el grado de aleatoriedad de las puntuaciones de un festival. Así, para cada año se creó una tabla que contenía el número de países que habían dado puntos a cada uno de los países participantes (Tabla 3).

País	Número de países que le han votado
Austria	17
Bélgica	22
Bosnia & Herzegovina	3
Croacia	6
Chipre	3
Estonia	4
Francia	4
Alemania	13
Grecia	6
Islandia	16
Irlanda	12
Israel	4

Tabla 3: *Porción de la tabla para el año 2003 con el número de países que ha dado votos a cada uno de los países que ha participado*

En un festival que fuera totalmente aleatorio, cada país recibiría, en promedio, puntuaciones de otros 10 países, independientemente del número de países que participaran. Efectivamente, imaginemos que un año cualquiera participan N países. Un país A podrá ser votado, como mucho, por N-1 países (un país no se puede votar a si mismo). Cada país le puede dar al país A: 1, 2, 3, 4, 5, 6, 7, 8, 10 o 12 puntos, o bien no votarlo y darle por lo tanto 0 puntos. Por lo tanto, la probabilidad de ser votado (de recibir puntos) es la siguiente:

$$P(\text{ser votado}) = \underbrace{\frac{1}{N-1} + \cdots + \frac{1}{N-1}}_{10 \text{ sumandos}} = \frac{10}{N-1}$$

La esperanza matemática (la media) del número de países que votarán a un determinado país es la siguiente:

$$E[\text{número de países que votarán}] =$$
$$= P(\text{ser votado}) \cdot \text{Número de países que pueden votar} =$$
$$= \frac{10}{N-1} \cdot (N-1) = 10$$

Por lo tanto, el valor esperado del número de países que votarán a cada uno de los países participantes es 10. Si el festival es muy aleatorio, el número de países que han dado puntos a cada uno de los países no se alejaría mucho de 10, y la desviación tipo (una medida de la dispersión de los datos) de todos estos valores será pequeña. En cambio, si imaginamos que el festival es muy poco aleatorio y que un país tiene mucho éxito y recibe puntos de todos los demás, y otro tiene un gran fracaso y no lo vota nadie, entonces necesariamente la desviación tipo será mayor.

La Figura 7 muestra cómo esta desviación tipo ha ido aumentando a lo largo de los años. Este es un resultado sorprendente: a partir de mediados de los 90, cuando se empieza a introducir el televoto y son los espectadores los que escogen sus canciones preferidas y no un jurado, el festival empieza a hacerse menos aleatorio, y cada vez se detectan más patrones de votación entre países.

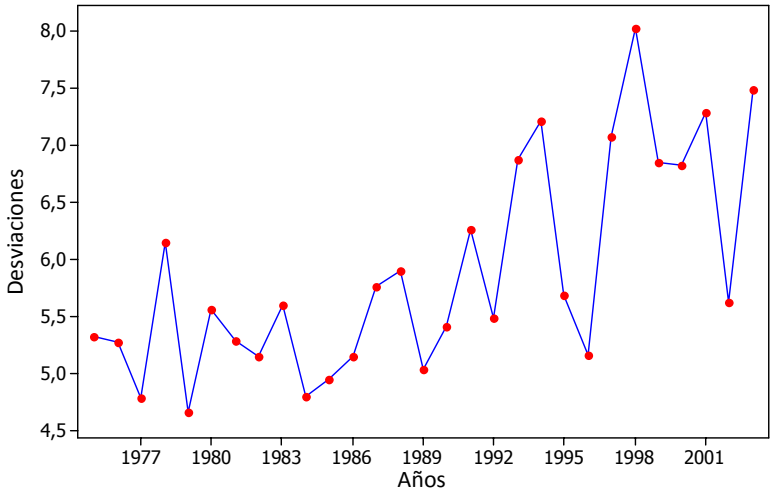


Figura 7: *Gráfico que muestra la evolución de la desviación tipo del número de países que votan a cada uno de los participantes.*

Simulación del festival aleatorio

Como antes hemos dicho, para simular un festival aleatorio se tiene que suponer que la calidad de las canciones no influye en las votaciones. En una edición del festival, la probabilidad de ser votado por un país (sin tener en cuenta cuántos puntos te dan) es $10/(N-1)$, dónde *N* es el número de países que participan en aquella edición. En este festival aleatorio, igual que en el real, es más probable no recibir ningún punto de un país que el hecho de que ese país te dé puntos.

La probabilidad de ser puntuado con cualquiera de los 10 posibles valores (1, 2, 3, 4, 5, 6, 7, 8, 10 y 12) es $1/(N_i-1)$, dónde N_i es el número de países participantes en el año *i*-ésimo. Por ejemplo, si un año participan 24 países la probabilidad de obtener votos de un determinado país es 0,0435. La probabilidad de recibir 0 puntos es $1-[10/(N_i-1)]=0{,}565$. Para cada edición del festival (desde 1975 hasta 2003) hemos creado una tabla con las probabilidades de recibir cada una de las puntuaciones (estas probabilidades varían según el número de países que han participado cada año).

A continuación, para cada pareja de países A y B se han simulado 10.000 posibles valores de votaciones del país A hacia el B en las ediciones de los festivales de 1975 a 2003 (es decir, 10.000 conjuntos de 29 valores, correspondientes a las que podrían

haber sido 10.000 puntuaciones de los 29 años de festival). Estas votaciones simuladas se generaron a partir de las probabilidades de cada año. Evidentemente, en algunas ediciones los dos países A y B pueden no haber participado simultáneamente. En esas ediciones no habrá votaciones entre ellos. Finalmente, lo que se obtiene son valores posibles de la media de puntos que el país A ha dado al B durante períodos de 5 años. Asimismo, se ha calculado para cada pareja de países el promedio de puntos que el país A ha dado realmente al país B en cada uno de esos períodos.

Todo este proceso ha sido bastante laborioso. Utilizamos macros de Excel para simular las puntuaciones y calcular los promedios anuales. Pero había tantos años y tantas parejas de países que tuvimos problemas de capacidad de memoria, y las macros de Excel eran tan lentas de ejecutar que había que dejar ordenadores en marcha durante la noche para que fueran calculando. Finalmente lo tuvimos todo calculado y ya estábamos en condiciones de comprobar qué países se votaban entre ellos exageradamente, mucho más de lo que les "tocaría".

Detección de patrones de votación

¿Cómo decidir si el país A ha votado al país B mucho más de lo que sería razonable esperar durante un periodo de tiempo? Con los promedios de los puntos que el país A ha dado a B en la simulación de ese periodo se ha dibujado un histograma. A este histograma le podemos llamar distribución de referencia. Por otra parte, también se dispone del valor real del promedio de puntos que el país A ha dado al B en ese periodo de tiempo. A este valor real le llamaremos estadístico de prueba. Lo que hay que hacer ahora es comparar el estadístico de prueba con la distribución de referencia. Es decir, ver por dónde cae el valor real de puntos que se han dado sobre el histograma de puntuaciones posibles.

Si el valor real cae por el medio del histograma (cómo pasa a la Figura 8 con la pareja Francia-Italia), podemos creer que éste es un valor posible de ese histograma. Como el histograma muestra valores de puntuaciones cuando el festival es aleatorio, parece que Francia no ha votado de manera exagerada Italia durante estos años.

En cambio, cuando el valor real queda muy a la derecha del histograma (cómo pasa en la Figura 9 con la pareja Chipre-Grecia), no parece razonable pensar que este valor tan alto sea fruto del azar. Lo que pasa es que Chipre ha dado puntuaciones mucho más altas en Grecia de lo que se podría esperar si el festival fuera aleatorio. Realmente, Chipre da a Grecia más votos de los que le "tocan".

Este proceso se repitió para todas las parejas de países (y en los dos sentidos, es decir, los votos de A hacia B y los de B hacia A), y para todos los periodos. Se han seleccionado aquellas parejas de países (A-B) que en un periodo de 5 años tenían un

valor medio de puntuación que dejaba sólo un 5% o menos de valores simulados por encima del valor real (es decir, que la cola hacia la derecha de la distribución de referencia era inferior al 5%). A esto se llama trabajar con un nivel de significación del 5%: valores como nuestro valor real o mayores, pero que sean fruto del azar, sólo los encontramos un 5% de las veces. Nuestra apuesta al decir que no es fruto del azar, sino que A ha votado a B más de lo que "tocaba" es, por lo tanto, bastante razonable.

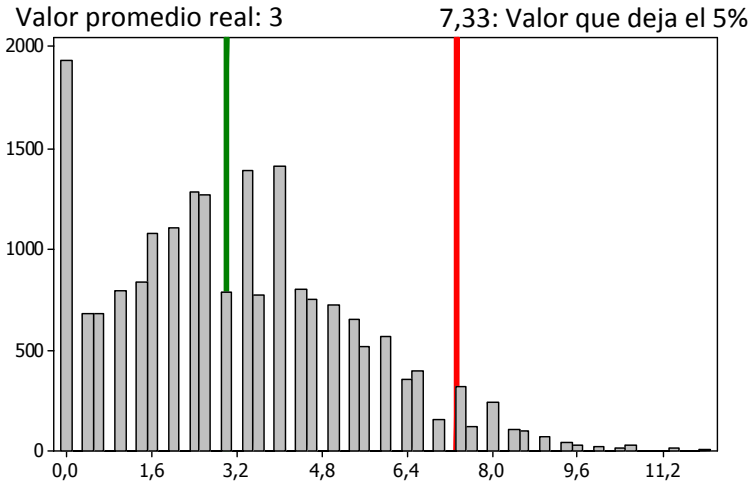


Figura 8: *Histograma de los valores promedio simulados para la pareja Francia-Italia (años 1980-1984)*

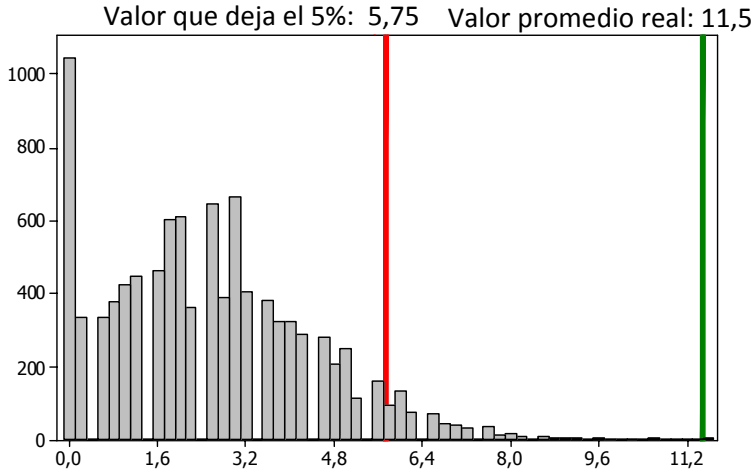


Figura 9: *Histograma de los valores promedio simulados para la pareja Chipre-Grecia (años 1995-1999)*

La Figura 10 muestra las parejas de países que recíprocamente se han votado de forma exagerada (con un nivel de significación del 5%) en cada período de años. Algunas parejas son muy fieles, como la de Grecia y Chipre; otros son más esporádicas. Lo que sí se puede ver con claridad es que, a medida que pasan los años, cada vez aparecen más países, y más relaciones de amistad entre ellos. El festival es cada vez menos aleatorio, un fenómeno que ya se había detectado al observar la desviación tipo de los países que habían dado votos a cada uno de los participantes.

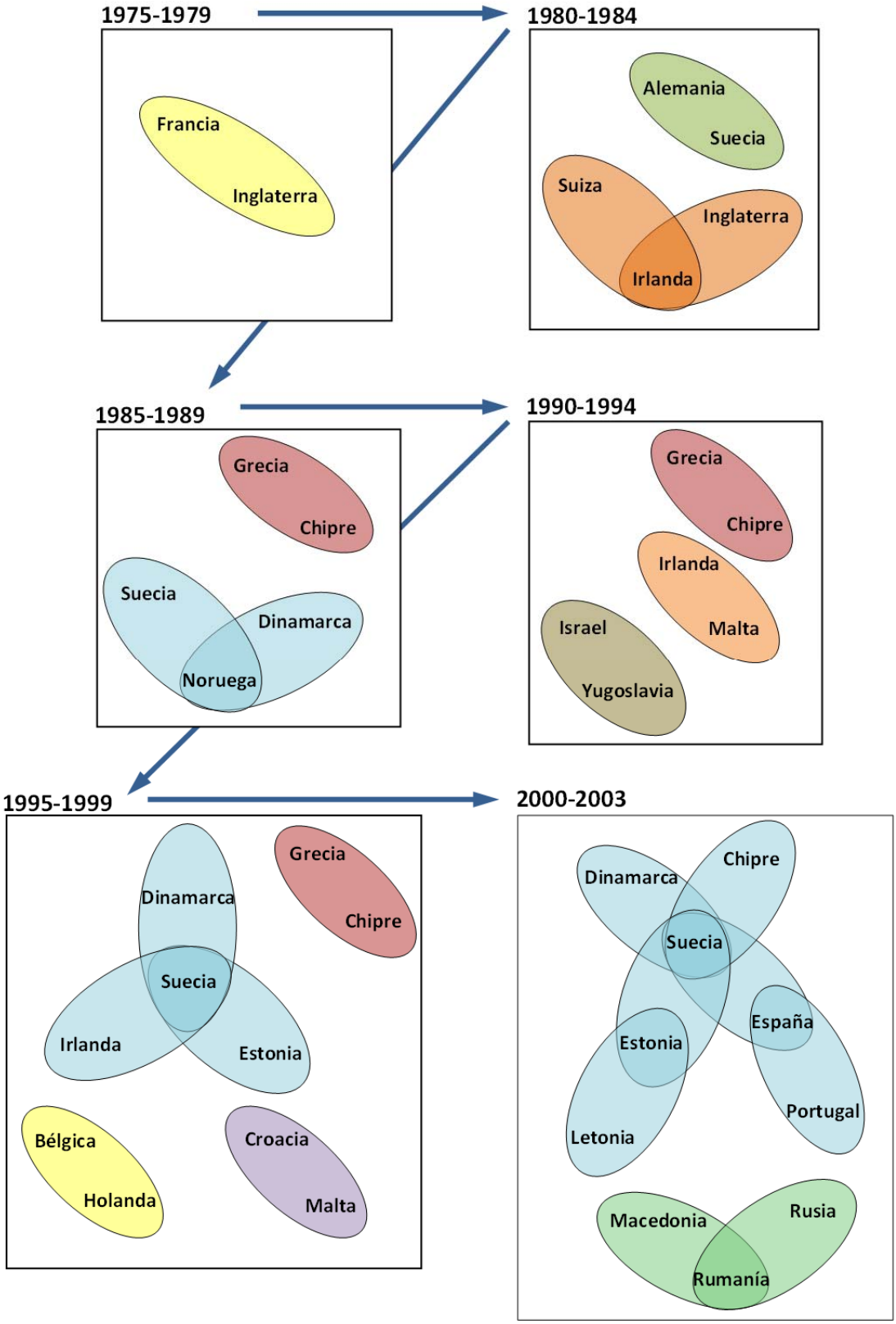


Figura 10: *Diagramas con las parejas de países que se votan recíprocamente más de lo normal de una manera significativa durante períodos de 5 años entre 1975 y 2003*

Algunos tópicos del festival de Eurovisión: ¿verdadero o falso?

La motivación que animó a la autora del proyecto y a su director a realizar este proyecto fue descubrir si son verdaderos o falsos algunos tópicos que año tras año se van repitiendo sobre el festival de Eurovisión. Éstos son:

¿Los países nórdicos se votan entre ellos?

Antes de responder a esta pregunta hay que definir cuáles son los países nórdicos. Se han considerado países nórdicos: Noruega, Finlandia, Suecia, Dinamarca e Islandia. Se han utilizado los resultados de la simulación del festival aleatorio para comprobar si las parejas de países nórdicos se han votado más de lo que había que esperar, con una significación del 5%.

En la Tabla 4 se pueden ver los periodos de años en que estos países han participado y las parejas más relevantes que han salido. El primer periodo (de 1975 en 1979) no aparece debido a que en este no ha salido ninguna pareja. Finlandia no ha aparecido en ningún periodo relacionada con ningún otro país nórdico.

Tabla 4: *Tabla con las parejas nórdicas más significativas dándose votos*

1980 - 1984	1985 - 1989	1990 - 1994	1995 - 1999	2000 - 2003
	Dina > Nor			
Dina > Sue	Dina > Sue		Dina > Sue	Dina > Sue
	Isla> Dina		Isla > Dina	Isla > Dina
			Isla > Nor	
	Nor > Dina			Nor > Dina
				Nor > Isla
Nor > Sue	Nor > Sue		Nor > Sue	Nor > Sue
	Sue > Dina		Sue > Dina	Sue > Dina
		Sue > Isla	Sue > Isla	
	Sue > Nor			Sue > Nor

Nor: Noruega, Sue: Suecia, Dina: Dinamarca, Isla: Islandia

Aquellos países que han tenido una relación recíproca de votaciones (aunque sea en periodos diferentes) aparecen dibujados con diagramas de Venn en la Figura 11.

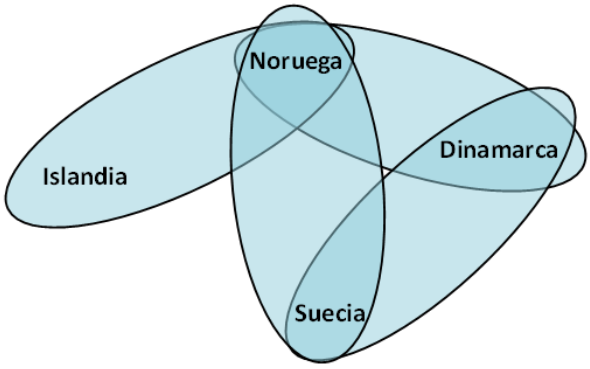


Figura 11: *Países nórdicos que se votan entre ellos (relaciones recíprocas)*

Si Islandia no apareciera (y de hecho la relación entre Islandia y Noruega es débil, sólo se presenta en un periodo de años en cada sentido), se podría concluir que hay como un "pacto" de tres países que se votan entre ellos de una manera exagerada. Estos tres países son Dinamarca, Suecia y Noruega. Por lo tanto, se puede concluir que es cierto que algunos países nórdicos se votan entre ellos.

¿Grecia y Chipre se votan entre ellos?

Ya se ha visto en otros apartados que estos dos países se votan entre ellos de manera muy descarada. La Tabla 5 nos lo muestra de forma clara: podemos ver los votos que se han dado Grecia y Chipre y al revés a lo largo de los 16 años que han participado conjuntamente. Con un color más oscuro se destacan las veces que los dos países se han dado las puntuaciones máximas (un total de 6). Con un color más claro se destacan las ediciones en que se han intercambiado 10 y 12 puntos. En el resto de años (color blanco) las puntuaciones que se han dado no son nada malas, aunque en el año 1983 Grecia no le diera ningún voto en Chipre.

Tabla 5: *Tabla con los votos que Grecia y Chipre se han dado*

Año	Grecia - Chipre	Chipre - Grecia
1981	12	6
1983	0	12
1985	8	8
1987	12	12
1989	7	12
1990	6	6
1991	12	10
1992	10	12

Año	Grecia - Chipre	Chipre - Grecia
1993	10	12
1994	12	12
1995	8	12
1996	12	10
1997	12	12
1998	12	12
2002	12	12
2003	12	12

Si se analizaran las canciones presentadas por estos dos países, el intercambio de votos se rebelaría todavía más escandaloso. El año 2002, por ejemplo, Grecia presentó la canción "S.A.G.A.P.O." una canción demencial interpretada por Mihalis Rakintzis (que por otra parte es un cantante famoso en su país). Aunque podía recibir votos de 23 países, 19 de ellos le dieron 0 puntos, pero Chipre lo votó con 12 puntos. Recordemos, en cambio, que Chipre no vota casi nunca a Turquía, como hemos podido ver en la Figura 4 (y todavía menos vota Turquía a Chipre).

Las votaciones más altas se consiguen los últimos años, desde que está implantado el sistema del televoto.

¿Actuar hacia el final del festival aumenta las posibilidades de ganar?

Ésta es una afirmación hecha a menudo por los comentaristas españoles del festival (especialmente por Beatriz Pecker, que presentó el festival del 2004 al 2007 sustituyendo al mítico José Luis Uribarri quien, por cierto, lo volvió a comentar en el 2008).

La Figura 12 muestra un diagrama de puntos de la posición en que han actuado los países ganadores. Se puede ver cómo no tiene ninguna relación la posición de salida con el hecho de ser ganador del festival o no.

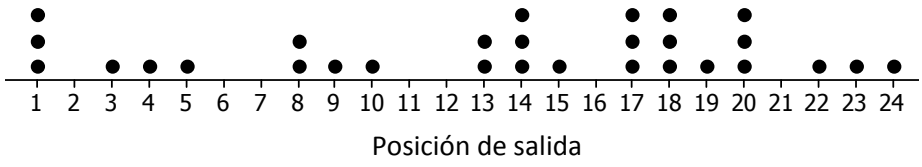


Figura 8: *Diagrama de puntos con las posiciones de salida de los ganadores*

Conclusiones

La afición tanto de Laura (que hizo este proyecto) como de Lluís (que lo dirigió) por el festival de Eurovisión es lo que los animó a realizar este estudio. La idea era analizar los datos (o sea, utilizar la estadística) para confirmar o desmentir algunas creencias en torno al festival.

Para comprobar cómo se votan los países entre ellos crearon este artificio del festival aleatorio, que les permitió ver qué países se dan puntos de forma exageradamente alta. Esta parte del proyecto es interesante, porque utiliza conceptos utilizados a

menudo en estadística (la distribución de referencia y el nivel de significación), pero de una forma un poco diferente a la habitual.

El estudio ha demostrado que sí es verdad que hay grupos de países que se votan entre ellos más de lo que cabría esperar, y que es mucho más fácil ganar el festival si perteneces a uno de estos grupos. Al utilizar datos hasta el 2003 no han aparecido "pactos" entre países del Este de Europa, pero seguramente si se actualizara el proyecto con los datos de los últimos años, estos países también aparecerían en las relaciones de amistad. Es curioso observar cómo la introducción del televoto a partir de comienzos del siglo XXI ha incrementado muchísimo los patrones de votación que se van repitiendo año tras año.

Si no perteneces a un grupo de países con "amigos" que te votan, probablemente la mejor estrategia para ganar sea presentar una canción que sorprenda, o que sea muy excéntrica. Esto es lo que hizo Finlandia en el año 2006 con Lordi, donde un grupo *heavy* con los intérpretes disfrazados de monstruos ganaron con la mayor puntuación de la historia del festival. Quizá la idea de España de presentar Rodolfo Chikilicuatre (año 2008) con su "Baila el chiki-chiki" no era, por tanto, tan mala. De hecho quedó en la posición 16 de 24 (por delante de D'Nash el año 2007, que quedaron en posición 20; de Las Ketchup que quedaron en posición 21 el año 2006, o de Son de Sol, también en posición 21 el año 2005).

(Proyecto de la Diplomatura de Estadística, presentado en julio de 2006 con el título "Anàlisi de patrons i tendències en les votacions del festival d'Eurovisió")

Ejemplo perfecto de análisis de datos utilizando técnicas fáciles de entender, con rigor, creatividad, ingenio y sentido del humor. ¿Qué más se puede pedir?

Laura Marí realizó la Diplomatura de Estadística y también la Licenciatura y el Máster en Estadística e Investigación Operativa en la UPC. Mientras estudiaba la licenciatura trabajó en prácticas en una empresa de estudios de mercado (TNS, Taylor Nelson Sofres). Le encanta la informática y es una programadora experta. Ahora está trabajando en el Departamento de Estadística e Investigación Operativa de la UPC en el 'Grupo de Optimización Numérica y Modelización' en el marco de un proyecto de investigación del *Ministerio de Ciencia y Tecnología* sobre modelización y optimización del mercado eléctrico.

15

La estadística en la prensa. Estudio crítico

Proyecto realizado por: **Sara Fontdecaba i Rigat**
Maria Montón i Domingo
Dirigido por**: Pere Grima Cintas**

Muchas veces se dice que en los medios de comunicación se usa la estadística para presentar la realidad de la forma que más interesa.

¿Es eso cierto? Este proyecto quería responder a esta pregunta. Pero naturalmente no se podían tratar los medios de comunicación en general, y el estudio se centró en el medio que era más fácil de analizar: la prensa.

Después de organizar cuáles eran los aspectos a considerar (qué había que mirar) y de hacer un seguimiento exhaustivo de los diarios estudiados durante unos tres meses, se detectaron cosas que se hacen muy bien, y también otras que se pueden mejorar.

Los ejemplos que ilustran los aspectos que se han considerado, y que han sido obtenidos de los propios diarios estudiados, son una de las partes más interesantes del proyecto.

Objetivo. Contenido

El objetivo de este proyecto era realizar un análisis crítico del uso que se hace de la estadística en la prensa. Se trataba de analizar un conjunto de diarios durante un cierto periodo de tiempo tomando nota de todos los aspectos relacionados con la estadística. Además, también se quería realizar una especie de estadística del uso de la estadística: qué técnicas se utilizan y con qué frecuencia.

La primera decisión fue escoger a que diarios se haría el seguimiento. Una primera idea era hacerlo a un conjunto amplio, incluyendo también algunos de difusión gratuita, pero como éste es un trabajo que exige mucho tiempo, finalmente se eligieron solo los 3 diarios con mayor tirada en Cataluña y que, además, se consideró que eran diarios serios: "El Periódico de Cataluña", "La Vanguardia" y "El País".

El periodo de tiempo durante el cual se hizo el análisis más detallado fue de unos 3 meses, desde mediados de diciembre de 2005 hasta finales de marzo de 2006. Se empezó a trabajar en el proyecto desde el mes de octubre, pero debido a los estudios preliminares y a las pruebas piloto que se hicieron, no se pudo iniciar el estudio sistemático hasta el mes de diciembre.

Desde el principio había que definir bien lo que se tenía que buscar y qué criterios aplicar para su valoración. Para estructurar las ideas se consultaron diversos libros, y los que resultaron más útiles fueron los de Darrell Huff: "How to Lie with Statistics" y el de Stephen K. Campbell: "Equívocos y falacias en la interpretación de estadísticas". Después de una prueba piloto que duró aproximadamente un mes, se estructuraron los aspectos a considerar y se creó una base de datos para ir anotándolos. Para este resumen se han escogido los aspectos más relevantes (Figura 1) y se ha seleccionado un ejemplo para ilustrar cada uno de ellos.

Resumen de datos	¿De qué estamos hablando? Precisión adecuada ¡Cuidado con los porcentajes! No ignorar la variabilidad
Representaciones gráficas	Originalidad: los pictogramas Claridad Proporcionalidad Comparaciones adecuadas Necesidad
Estudios basados en muestras	Representatividad Tamaño Margen de error

Figura 1: *Aspectos considerados en este resumen*

Resumen de datos

¿De qué estamos hablando?

Los datos pierden valor si están referidas a un concepto ambiguo. Es más grave cuando los términos utilizados sugieren un significado, pero en realidad tienen otro.

En el titular "El 22% de los jóvenes salen de noche sin supervisión de los padres" el término "supervisión" tiene un significado poco claro. Algunos de nosotros lo interpretaríamos como que los padres deben conocer los lugares y las personas con quienes irán sus hijos, para otros será que lo deben conocer y además tienen que dar el visto bueno, o quizá para alguien supervisar implica acompañar al hijo y asegurarse de que no hace nada malo. La interpretación de "supervisión" depende del nivel de exigencia de los padres. Y si son los jóvenes los que han contestado la encuesta, su interpretación de lo que significa supervisión seguramente será diferente a la interpretación de los padres.

Otro tema es qué se entiende por "joven" en el contexto de este titular. ¿Hasta 16 años, 17, 18? En el texto se especifica que el sondeo se llevó a cabo entre jóvenes de 15 a 29 años. Una persona de 28 o 29 años, quizás casado y con hijos, es realmente una persona joven, pero no el tipo de joven en el que pensamos al leer este titular.

COLECTIVOS

El 22% de los jóvenes salen de noche sin la supervisión de los padres

ocio nocturno. El sondeo, inédito hasta ahora, se realizó entre jóvenes de 15 a 29 años. Un 67% de los consultados salen normalmente las noches de los fines de semana, frente a un tercio (32,1%) que no lo hace.

Figura 2: *El Periódico, 30 de enero de 2006. Página 26*

Precisión adecuada

El grado de precisión debe estar en sintonía con los conocimientos que se tienen sobre la medida en cuestión.

El sábado 10 de junio de 2006 se realizó una manifestación en Madrid en la cual el número de manifestantes fue un valor discutido (como en la mayoría de manifestaciones). Según la Comunidad de Madrid eran un millón de personas. En

cambio, según el Gobierno, fueron exactamente 242.923. Estos cálculos suelen realizarse estimando la superficie del suelo ocupado por manifestantes y multiplicando cada metro cuadrado por el número de personas que se cree que hay. Pero es imposible realizar un cálculo como éste con gran exactitud, lo realmente importante en estos casos es el orden de magnitud. Hay que destacar que aquí la "falta" era del Gobierno, no de la Vanguardia que, por lo visto, sólo reprodujo esta información.

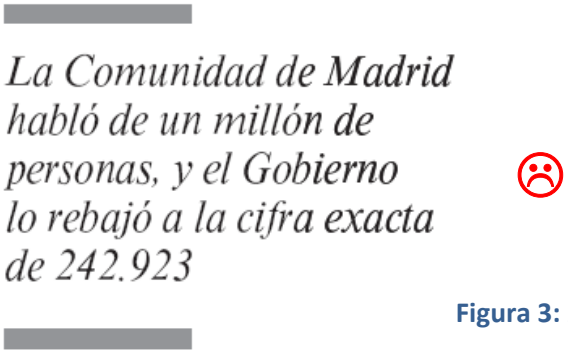


Figura 3: *La Vanguardia, 11 de junio de 2006. Pág. 22*

¡Cuidado con los porcentajes!

Debe estar clara la base sobre la que se calcula el porcentaje y, al realizar operaciones matemáticas, es necesario ponderar cada término de la forma adecuada.

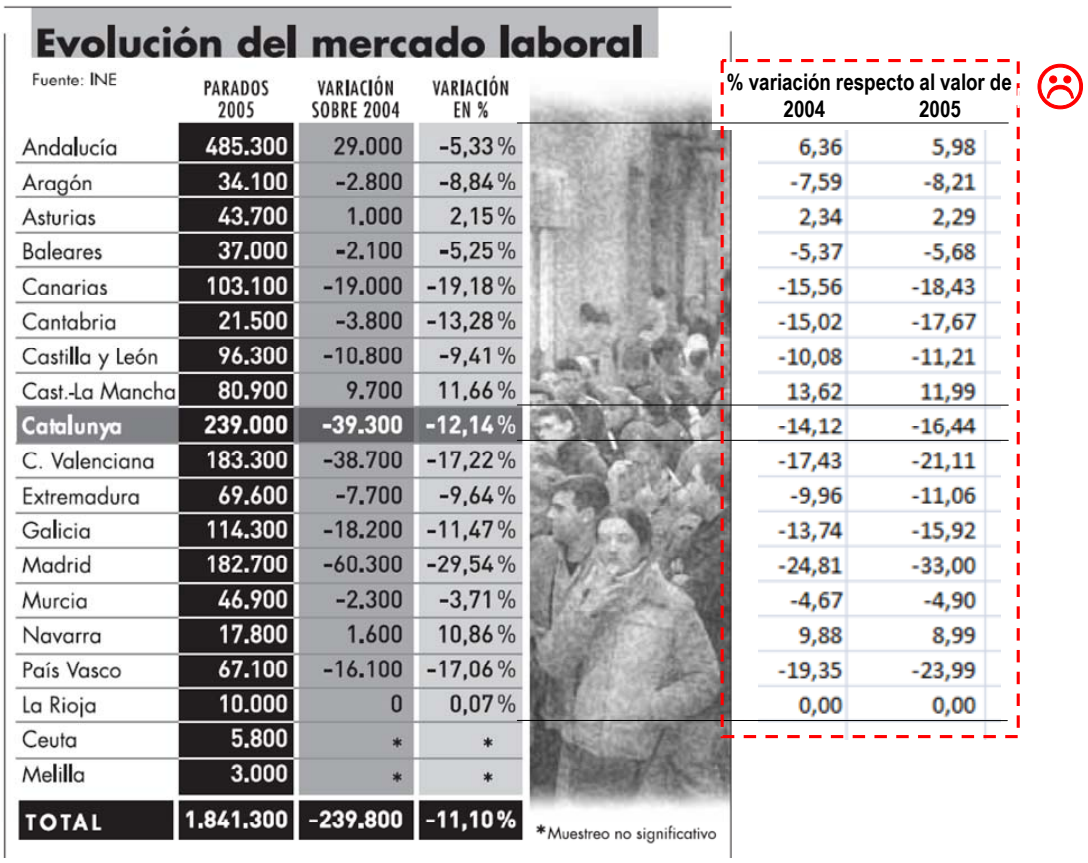
El 28 de enero de 2006 se publicaron unos datos sobre la evolución del mercado laboral en cada Comunidad Autónoma (Figura 4). El cuadro que acompaña la información indica el número de parados en el 2005, la variación sobre el 2004 (por lo tanto, pueden deducirse también los datos del 2004) y la variación, en porcentaje, de un año respecto del otro. Entre otros, hay dos aspectos a destacar:

1. No queda claro si la variación está calculada sobre los datos del 2005, que son los que aparecen, o sobre las del 2004, que seguramente sería el más razonable.

2. En cualquier caso, si calculamos estos porcentajes con los datos disponibles, los valores obtenidos no coinciden ni con un criterio ni con el otro. Aparte de un problema de signo en Andalucía, hay diferencias importantes entre el valor reproducido y el valor calculado con los datos proporcionados en comunidades como Cataluña o el País Vasco. En algún caso, como por ejemplo en Castilla-La Mancha, el porcentaje de variación es muy parecido a lo que sale respecto del 2005, pero en otros (Comunidad Valenciana) es similar con respecto al 2004.

Los datos de desempleo y de variación están redondeados mientras que los porcentajes se dan con 2 decimales. Quizá estos porcentajes se han calculado con los datos sin redondear, aunque las diferencias son demasiadas grandes. Se dice que la

fuente es el INE, pero no queda claro si de toda la tabla o solo de algunos de los datos (¿cuáles?, ¿ya redondeados?). Finalmente, y ya puestos a criticar, la expresión "Muestreo no significativo" no tiene sentido. El muestreo puede ser insuficiente o no representativo, lo que podría ser no significativo es la diferencia observada.



Evolución del mercado laboral

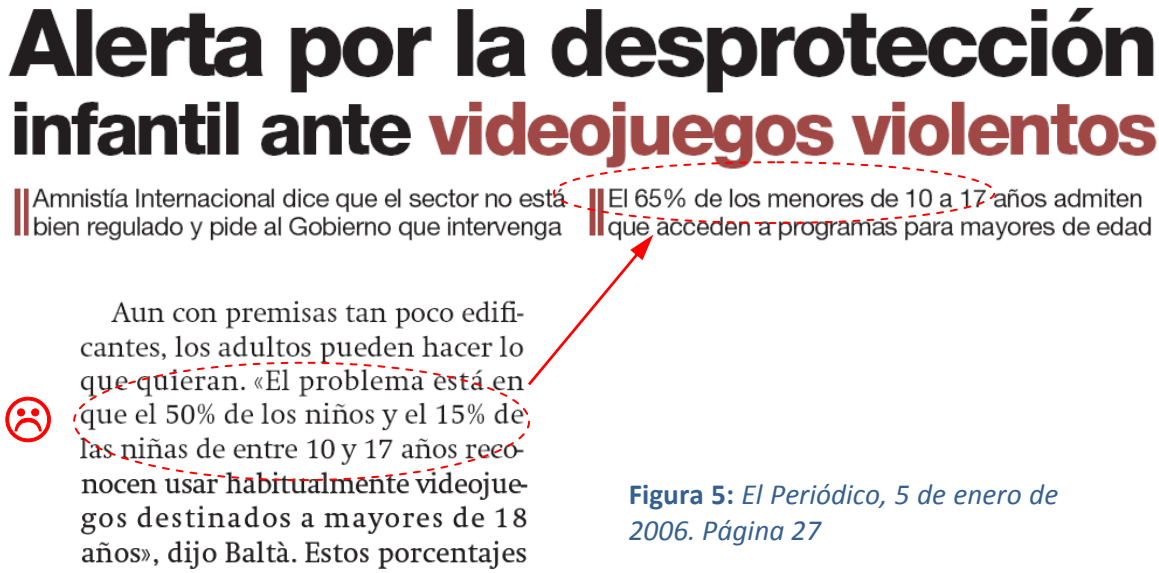
Fuente: INE

	PARADOS 2005	VARIACIÓN SOBRE 2004	VARIACIÓN EN %	% variación respecto al valor de 2004	% variación respecto al valor de 2005
Andalucía	485.300	29.000	-5,33 %	6,36	5,98
Aragón	34.100	-2.800	-8,84 %	-7,59	-8,21
Asturias	43.700	1.000	2,15 %	2,34	2,29
Baleares	37.000	-2.100	-5,25 %	-5,37	-5,68
Canarias	103.100	-19.000	-19,18 %	-15,56	-18,43
Cantabria	21.500	-3.800	-13,28 %	-15,02	-17,67
Castilla y León	96.300	-10.800	-9,41 %	-10,08	-11,21
Cast.-La Mancha	80.900	9.700	11,66 %	13,62	11,99
Catalunya	239.000	-39.300	-12,14 %	-14,12	-16,44
C. Valenciana	183.300	-38.700	-17,22 %	-17,43	-21,11
Extremadura	69.600	-7.700	-9,64 %	-9,96	-11,06
Galicia	114.300	-18.200	-11,47 %	-13,74	-15,92
Madrid	182.700	-60.300	-29,54 %	-24,81	-33,00
Murcia	46.900	-2.300	-3,71 %	-4,67	-4,90
Navarra	17.800	1.600	10,86 %	9,88	8,99
País Vasco	67.100	-16.100	-17,06 %	-19,35	-23,99
La Rioja	10.000	0	0,07 %	0,00	0,00
Ceuta	5.800	*	*		
Melilla	3.000	*	*		
TOTAL	1.841.300	-239.800	-11,10 %		

*Muestreo no significativo

Figura 4: *La Vanguardia, 28 Enero 2006. Pág. 55 (la parte del recuadra a trazos está añadida)*

Las operaciones con porcentajes hechas a la ligera pueden dar "perlas" como la publicada al Periódico el 5 de enero de 2006 (Figura 5). Evidentemente el porcentaje global tiene que ser del 32,5% suponiendo que hay tantos niños como niñas. Dos porcentajes no se pueden sumar si antes no se han ponderado por la proporción de individuos de cada tipo existentes en la población, de lo contrario nos podríamos encontrar con casos (como el del ejemplo) en qué el 100% de los niños y el 100% de las niñas resultan ser el 200% de los menores.



Alerta por la desprotección infantil ante videojuegos violentos

Amnistía Internacional dice que el sector no está bien regulado y pide al Gobierno que intervenga

El 65% de los menores de 10 a 17 años admiten que acceden a programas para mayores de edad

Aun con premisas tan poco edificantes, los adultos pueden hacer lo que quieran. «El problema está en que el 50% de los niños y el 15% de las niñas de entre 10 y 17 años reconocen usar habitualmente videojuegos destinados a mayores de 18 años», dijo Baltà. Estos porcentajes

Figura 5: *El Periódico, 5 de enero de 2006. Página 27*

No ignorar la variabilidad

Hay que evitar identificar a toda la población con el valor de su media. En muchos casos no es suficiente con la media para describir un conjunto de datos.

El día 25 de noviembre de 2005 aparecían dos informaciones casi idénticas en el Periódico y La Vanguardia sobre los gastos en el periodo de Navidad. El gasto por habitante se calcula dividiendo el gasto total estimado (no se explica cómo ni quién lo ha hecho) por el número de catalanes (tampoco sabemos cuáles). El Periódico especifica en su titular que se trata de una media, pero en la Vanguardia parece que cada catalán se gastará exactamente 719 €.

Las costumbres navideñas ▸ La economía doméstica

Cada catalán gastará en compras de Navidad una media de 719 €

VIERNES, 25 NOVIEMBRE 2005 VIVIR | 5

Cada catalán gastará 719 euros en las compras de Navidad

Figura 6: *El Periódico (arriba) y La Vanguardia (abajo). 25 de noviembre de 2005. Pág. 42 y 5 (suplemento Vivir) respectivamente*

Dar sólo la media muchas veces no es suficiente para describir una determinada situación. Por ejemplo, que los consejeros de sociedades ganen una media de 195.000 € al año, en rigor no informa sobre si ganan mucho o poco. Podría ser que unos pocos ganaran muchísimo, y la mayoría ganaran más bien poco. Sería interesante saber, además de la media, cuál es el rango de los salarios o, puestos a dar sólo un número, en este caso sería más adecuado dar la mediana, valor que deja por debajo, y también por encima, el 50% de los valores. Aunque también es verdad que ésta es una medida con la que no se está muy familiarizado.

INFORME SOBRE LA GESTIÓN EMPRESARIAL

Los consejeros de sociedades ganan 195.000 € por año de media ☹

Figura 7: *El Periódico, 23 de diciembre de 2005. Página 39*

Representaciones gráficas

Originalidad: los pictogramas

Es una buena idea introducir originalidad en las representaciones gráficas siempre y cuando la información se transmita de manera clara y correcta.

Los pictogramas combinan un tipo determinado de gráfico con una imagen relacionada con la temática de que trata la noticia. Éste es un recurso muy utilizado por El Periódico y, con menos frecuencia, por La Vanguardia y El País.

La representación de la Figura 8 recuerda un diagrama de barras y muestra de una manera muy clara el número de Oscar a que había sido nominada cada película, y el número finalmente conseguido.

Figura 8: *La Vanguardia, 7 de marzo de 2006. Página 3 – Especial Oscar*

En cambio, la Figura 9 muestra un pictograma que no sólo no ayuda a entender la información contenida sino que más bien complica la comprensión de los datos que, además, no mantienen ninguna proporcionalidad con la imagen representada.

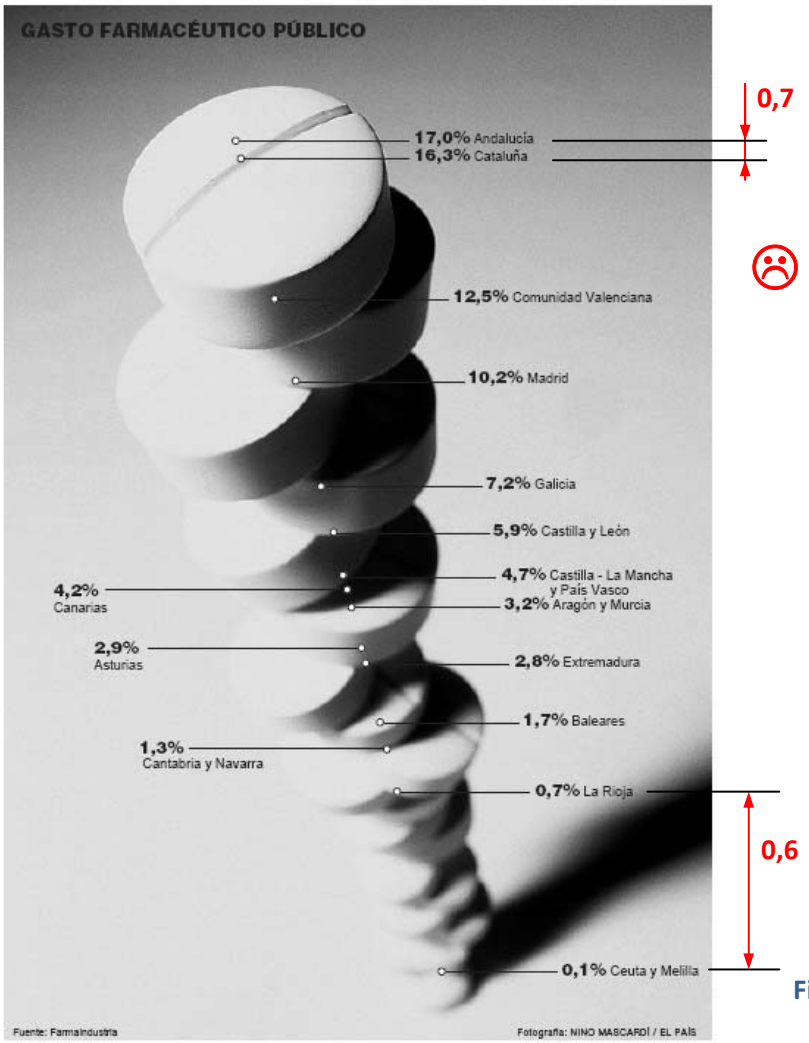


Figura 9: *El País, 22 de noviembre de 2005. Página 15*

Claridad

Todo gráfico tiene que perseguir que su lectura y su interpretación sean fáciles.

La Figura 10 compara actividades relacionadas con el "consumo cultural" de jóvenes y adultos. La parte de arriba "lo que no hacen los jóvenes" es realmente difícil de interpretar. ¿Quién va más al teatro, los jóvenes o los adultos? Estaría más claro si lo hicieran tal como lo presentan en la parte de abajo, explicando qué es lo que sí hacen.

El gráfico de la Figura 11 (si se puede considerar así) no ayuda a entender, ni a captar con más rapidez el significado de los datos, sino que el hecho de mezclar datos y dibujos sin ningún sentido complica la lectura y la comprensión de la información.

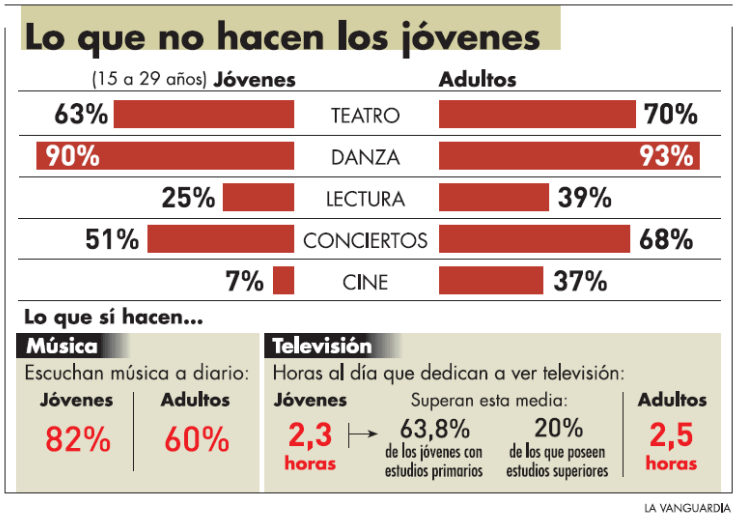


Figura 10: *La Vanguardia, 3 de marzo de 2006. Página 40*

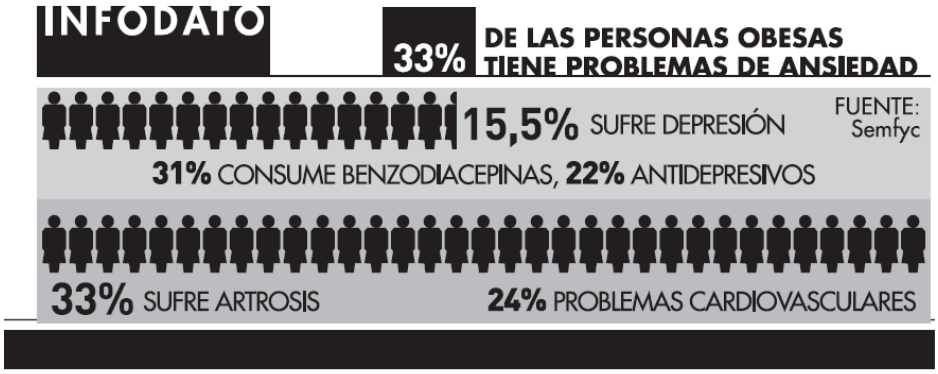


Figura 11: *La Vanguardia, 30 de noviembre de 2004. Página 34*

Proporcionalidad

Se debe mantener la proporcionalidad en los ejes de los gráficos.

Para que la representación gráfica dé una imagen fiel de la información que contienen los datos, es necesario que las escalas de los ejes mantengan su proporcionalidad. La falta de proporcionalidad es más frecuente en el eje vertical (la Figura 9 es un ejemplo) y puede provocar que un pequeño incremento parezca mayor de lo que realmente es, o disimular la importancia de otro mayor. En la prensa, seguramente esta práctica está más orientada a buscar imágenes creativas que a confundir al lector, pero en anuncios publicitarios esta deformación del gráfico puede tener interés más allá de la pura estética.

La Figura 12 muestra un ejemplo de gráfico con el eje horizontal no proporcional. Muestra la evolución del tipo oficial fijado por el Banco Central Europeo (BCE). Sólo se han representado aquellos meses en los cuales el BCE aumentó o disminuyó los tipos de interés. Una representación como la propuesta es más fiel a la realidad: se muestran las variaciones únicamente cuando se producen, se representa todo el periodo estudiado y los cambios, en vez de dibujarlos progresivos, se hacen puntuales (tal como ocurre con los tipos de interés).

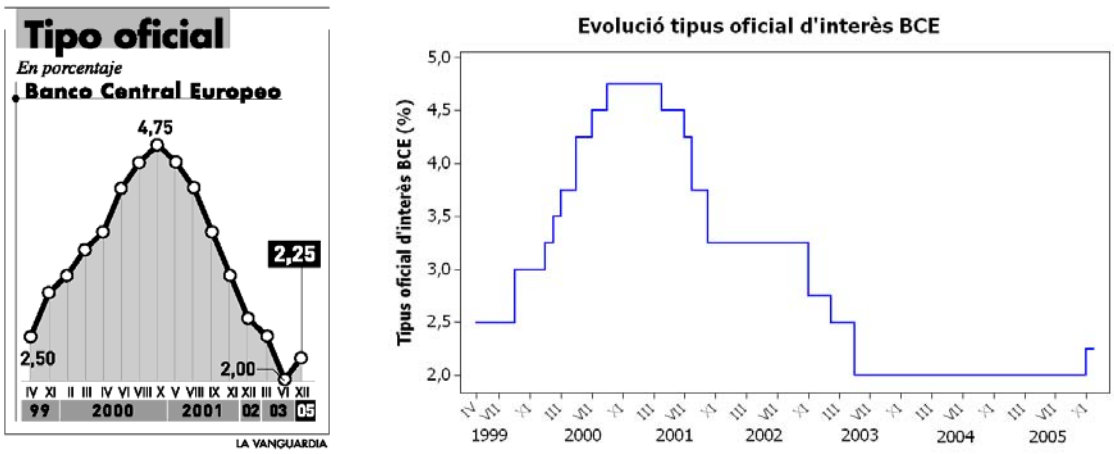


Figura 12: *Gráfico publicado en La Vanguardia, el 2 de diciembre de 2005, pág. 73 i propuesta de mejora (derecha)*

Comparaciones adecuadas

Todas las representaciones deben realizarse con los mismos criterios que se quieren comparar y estos criterios de comparación deben ser los adequados.

Muchas veces resulta de interés publicar datos que comparen el comportamiento de todo el Estado con el de una comunidad autónoma concreta. En la Figura 13 se hace la comparación de las audiencias de televisión en Cataluña y en España de un conjunto de partidos de fútbol. Aunque un partido despierte mucha más expectación en Cataluña, sería excepcional que el número de espectadores fuera más elevado que en el conjunto de toda España (y es imposible si en los datos de España se incluyen los de Cataluña) por el hecho de que el número total de habitantes en Cataluña es muy inferior al del conjunto del Estado.

De acuerdo con los datos que aparecen a la Figura 13, en porcentaje, las audiencias son mayores en Cataluña que en el resto de España cuando juega el Barcelona, y no es así cuando juega el Real Madrid. En valor absoluto (en total) siempre es mayor la audiencia del conjunto de toda España, como era de esperar. Pero lo que el gráfico compara son valores absolutos y, aunque también da el dato del porcentaje, del gráfico propiamente dicho no se extrae, de forma fácil, ninguna información relevante.

A la derecha se incluye el boceto de una representación gráfica que compara los porcentajes de audiencia donde se ve claramente que cuando juega el Barça la audiencia es mayor en Cataluña en términos relativos. Al hacer comparaciones entre poblaciones de diferentes tamaños es siempre mejor hacerlo en porcentajes.

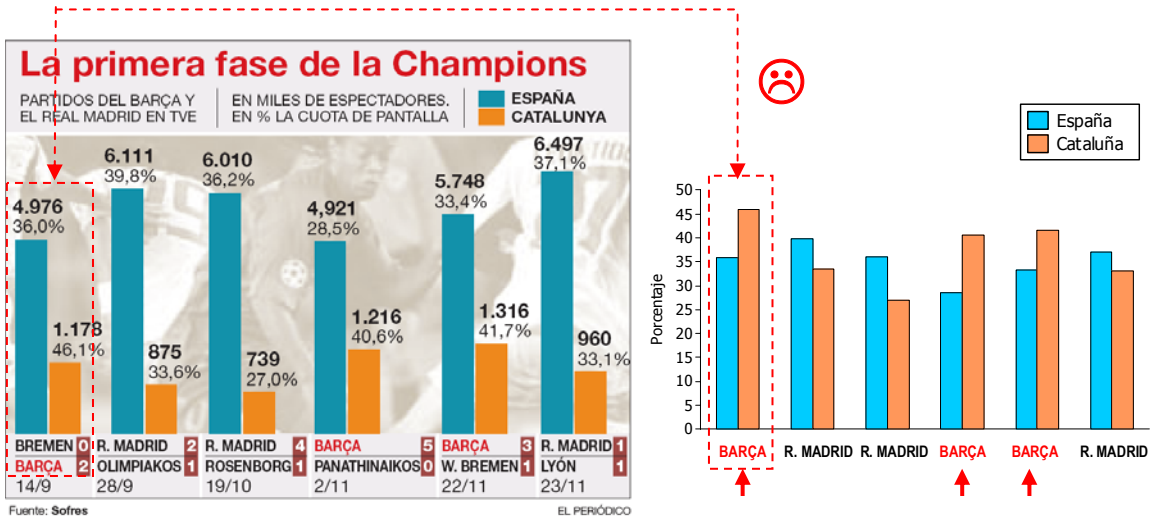


Figura 13: *El Periódico, 21 de diciembre de 2005, Pág. 80 y propuesta de mejora (esquemática)*

Necesidad

Conviene hacer una representación gráfica de los datos cuando esta ayuda a su lectura e interpretación, no cuando la dificulta.

Los gráficos de las Figuras 14 y 15 no ayudan a captar la información que contienen los datos, más bien complican la tarea. Sería mejor utilizar simplemente una tabla.

Figura 14: *La Vanguardia, 4 enero 2006, p. 6*

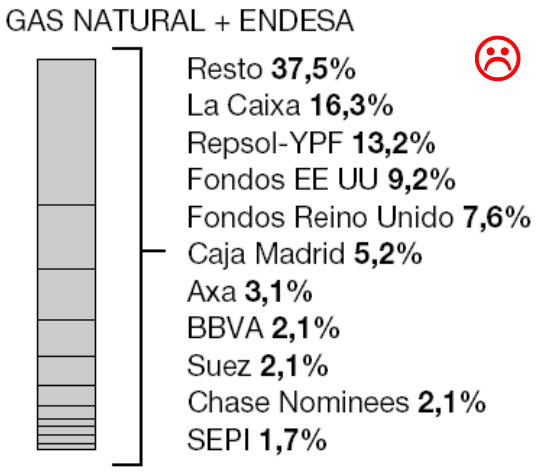


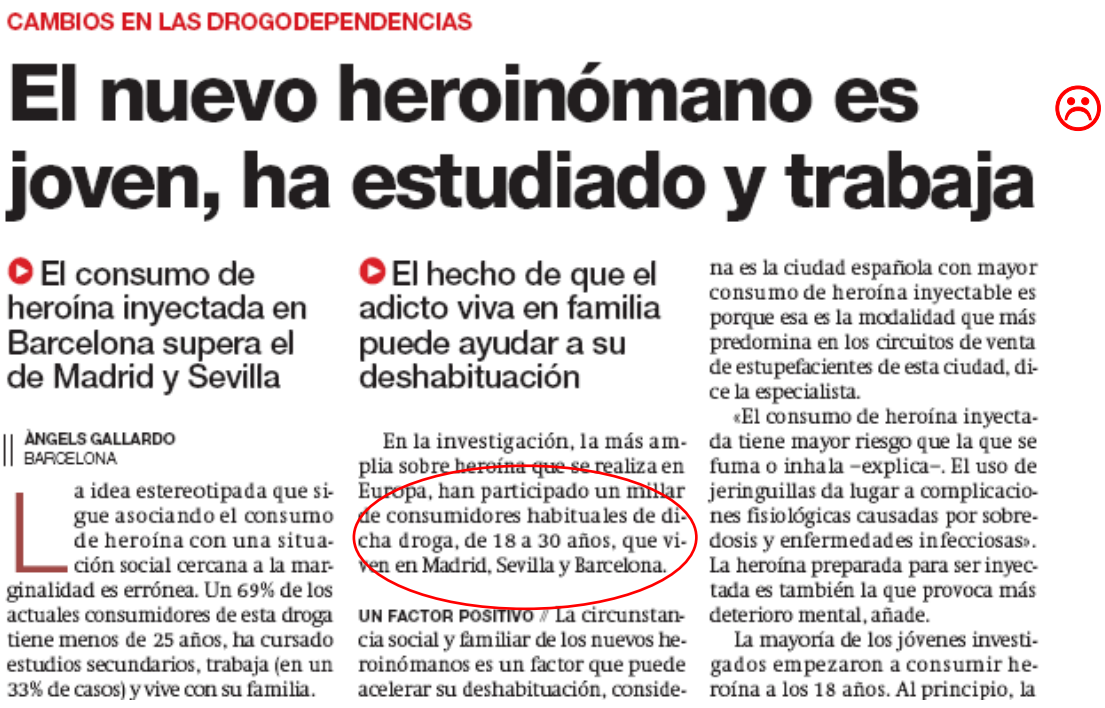
Figura 15: *El País, 18 diciembre 2005. Pág. 58*

Estudios basados en muestras

Representatividad

No se pueden sacar conclusiones si la muestra de que se dispone no representa adecuadamente a la población que se quiere estudiar.

La noticia de la Figura 16 hace referencia a un estudio sobre los cambios que se han producido en el perfil de los consumidores habituales de heroína. Dice el texto que en la investigación " ...han participado un millar de consumidores habituales de esta droga, de 18 a 30 años, que viven en Madrid, Sevilla y Barcelona." Lo peculiar de ésta noticia es el titular escogido: "El nuevo heroinómano es joven, ha estudiado y trabaja". Que es joven no puede ser una conclusión del estudio ya que en la muestra todos los individuos eran jóvenes, y no se podía concluir, por ejemplo, que el consumidor habitual tiene más de 40 años. Que ha estudiado es una afirmación ambigua, ¿quiere decir que tienen estudios universitarios? Esta muestra de personas jóvenes ha vivido en un periodo en que la enseñanza es obligatoria hasta los 16 años y no sería noticia haber estudiado hasta esta edad.



CAMBIOS EN LAS DROGODEPENDENCIAS

El nuevo heroinómano es joven, ha estudiado y trabaja

El consumo de heroína inyectada en Barcelona supera el de Madrid y Sevilla

El hecho de que el adicto viva en familia puede ayudar a su deshabituación

ÀNGELS GALLARDO
BARCELONA

La idea estereotipada que sigue asociando el consumo de heroína con una situación social cercana a la marginalidad es errónea. Un 69% de los actuales consumidores de esta droga tiene menos de 25 años, ha cursado estudios secundarios, trabaja (en un 33% de casos) y vive con su familia.

En la investigación, la más amplia sobre heroína que se realiza en Europa, han participado un millar de consumidores habituales de dicha droga, de 18 a 30 años, que viven en Madrid, Sevilla y Barcelona.

UN FACTOR POSITIVO // La circunstancia social y familiar de los nuevos heroinómanos es un factor que puede acelerar su deshabituación, considera es la ciudad española con mayor consumo de heroína inyectable es porque esa es la modalidad que más predomina en los circuitos de venta de estupefacientes de esta ciudad, dice la especialista.

«El consumo de heroína inyectada tiene mayor riesgo que la que se fuma o inhala –explica–. El uso de jeringuillas da lugar a complicaciones fisiológicas causadas por sobredosis y enfermedades infecciosas». La heroína preparada para ser inyectada es también la que provoca más deterioro mental, añade.

La mayoría de los jóvenes investigados empezaron a consumir heroína a los 18 años. Al principio, la

Figura 16: *El Periódico, 26 octubre 2005. Página 47*

También es destacable que en este estudio sólo han participado heroinómanos de tres de las ciudades con más habitantes de España, y es discutible que sus conclusiones se puedan generalizar, ya que es posible que las personas que residen en zonas con menos habitantes tengan unas características diferentes.

Tamaño

Los estudios basados en muestras deben tener un tamaño suficiente para que las conclusiones se puedan generalizar.

La Figura 17 muestra una pintoresca información sobre las conclusiones obtenidas en un estudio con solo 10 personas. Pero parece que el tamaño de la muestra no es el único aspecto poco serio de este estudio.

La sed acentúa la sensación de dolor

CIENTÍFICOS DE la Universidad de Melbourne analizaron la respuesta de 10 personas que fueron sometidas a sensaciones de dolor y de sed, inyectándoles sustancias salinas. Según sus resultados, la sed estimula el dolor, pero una sensación dolorosa no provoca ganas de beber.

Figura 17: *El Periódico, 31 enero 2006. Página 35*

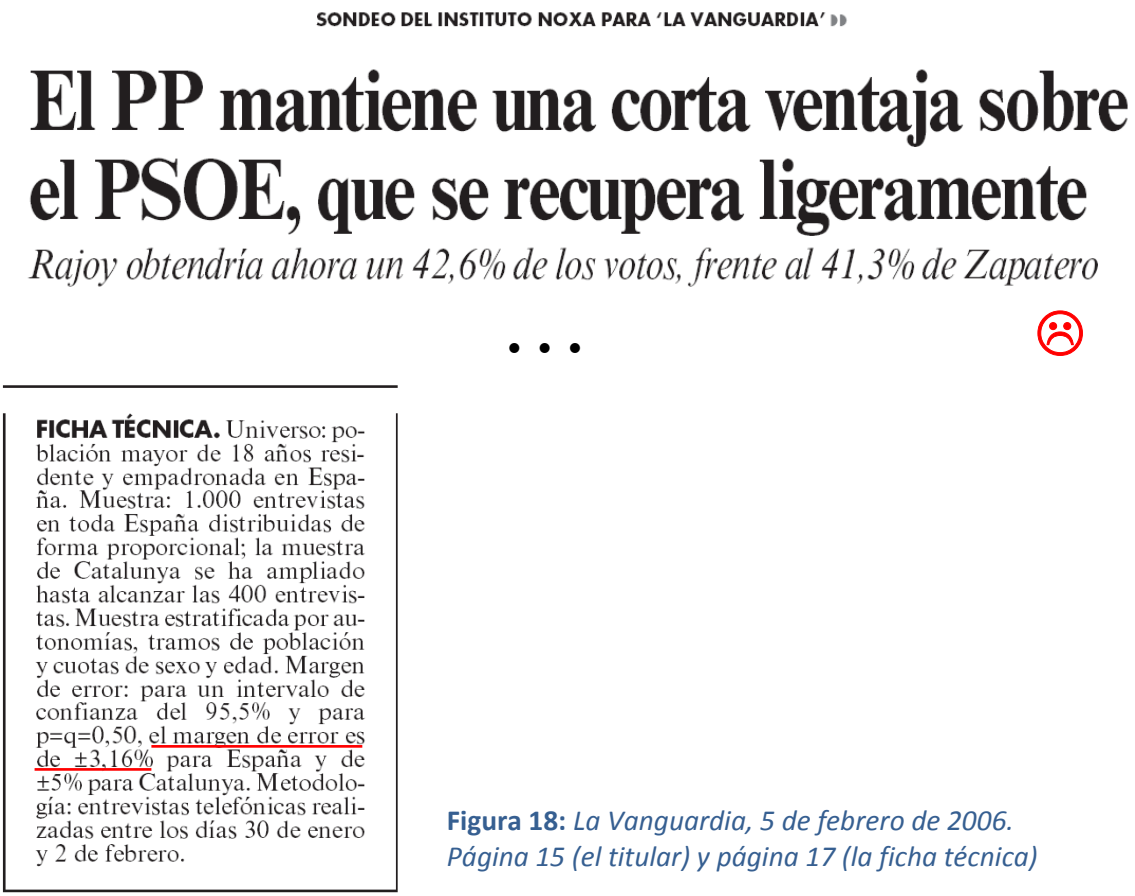
Margen de error

Es necesario tomar en consideración el margen de error en la interpretación de los resultados.

En general, cuando se estiman características de una población a través de una muestra y, en particular, cuando se estima el porcentaje de votos que tendrá un partido político, el valor obtenido no se puede considerar un valor exacto, sino que está sometido a un cierto margen de error (si se volviera a seleccionar al azar otra muestra de las mismas características el resultado no sería el mismo). Por ejemplo, si el resultado obtenido es de un 40% y el margen de error es del 5%, quiere decir que el porcentaje de votantes estará entre el 35 y el 45%. No se puede precisar si el porcentaje será del 38 o del 43% porque "el aparato de medida" no puede hilar más fino.

La noticia de la Figura 18 hace referencia a los resultados de un estudio que pretende, entre otras cosas, estimar los votos que tendrían PP y PSOE si se hubieran hecho las elecciones en aquel momento. El estudio incluye la ficha técnica (habitual en los estudios serios) donde informa de que el margen de error es del ±3,16%. Si el resultado del PP es del 42,6% y el del PSOE del 41,3%, no se puede afirmar que el PP mantiene una ventaja (ni corta ni larga) sobre el PSOE. Que el margen de error sea del ±3,16% quiere decir que si se volviera a hacer el estudio, exactamente de la misma forma y con una muestra de las mismas características, el resultado sería el que se ha obtenido ±3,16%. El titular tendría que transmitir la idea de que del estudio realizado se desprende que no hay diferencias entre la intención de voto a ambos partidos.

A menudo nos encontramos con que se adjunta y se comenta la ficha técnica del muestreo realizado, pero se ignora por completo en las conclusiones que se extraen.



SONDEO DEL INSTITUTO NOXA PARA 'LA VANGUARDIA'

El PP mantiene una corta ventaja sobre el PSOE, que se recupera ligeramente

Rajoy obtendría ahora un 42,6% de los votos, frente al 41,3% de Zapatero

• • •

FICHA TÉCNICA. Universo: población mayor de 18 años residente y empadronada en España. Muestra: 1.000 entrevistas en toda España distribuidas de forma proporcional; la muestra de Catalunya se ha ampliado hasta alcanzar las 400 entrevistas. Muestra estratificada por autonomías, tramos de población y cuotas de sexo y edad. Margen de error: para un intervalo de confianza del 95,5% y para p=q=0,50, el margen de error es de ±3,16% para España y de ±5% para Catalunya. Metodología: entrevistas telefónicas realizadas entre los días 30 de enero y 2 de febrero.

Figura 18: *La Vanguardia, 5 de febrero de 2006. Página 15 (el titular) y página 17 (la ficha técnica)*

Estadística del uso de la estadística

Los gráficos más utilizados en los tres diarios estudiados son los diagramas de barras, los diagramas de sectores y las series temporales, dejando muy atrás el número de otros tipos de representaciones como los diagramas de barras adosadas, las barras apiladas o los pictogramas. Concretamente, el tipo más frecuente en El País y La Vanguardia son los gráficos de barras y en El Periódico los diagramas de sectores.

Según las diferentes características evaluadas, aproximadamente un 60% de los gráficos encontrados en los tres diarios estudiados son correctos. Concretamente el uso de diagramas de sectores y series temporales acostumbran a ser correctos en todos los diarios. El 13% de los gráficos los podemos catalogar como mejorables, y el 28% restante están mal hechos. Dentro de la categoría de "Mal hechos" se han incluido los gráficos que contienen errores evidentes, los que son engañosos (pueden conducir a diferentes interpretaciones) y los absurdos (no contienen información). A rasgos generales destaca que La Vanguardia es el único de los tres diarios en que –en el periodo estudiado– se han detectado más usos incorrectos que correctos.

A pesar de que las referencias clasificadas como "resumen de datos" (medias, margen de error ...) no son tan abundantes como las representaciones gráficas, son muchos los aspectos que se pueden destacar. El error más frecuente reside en que muchas veces se proporciona el margen de error, pero casi nunca se tiene en cuenta al extraer las conclusiones. En segundo lugar se puede situar la representatividad de la muestra como otro de los puntos débiles. Sobre esta representatividad muchas veces ni se puede opinar, ya que no se especifica cuál ha sido la muestra ni de dónde se han sacado los datos ni el procedimiento seguido para obtenerlos.

Otro problema bastante general es el de interpretar la media como un todo, ignorando por completo la variabilidad de los datos. Inferir que todo el mundo es igual a la media es algo frecuente en todos los diarios.

(Proyecto de la Diplomatura de Estadística, presentado en julio de 2006 con el título "Ús de l'estadística a la premsa. Estudi crític")

El proyecto cumpleto ilustra los errores más frecuentes, y también las cosas que se hacen bien, a través de 103 ejemplos seleccionados durante el periodo estudiado. Alguien había recomendado que se divulgara más esta información, que se enviara a los diarios, que seguro que lo agradecerían. Pero siempre hay cosas más urgentes y ya no se había vuelto a tocar este tema. Hasta ahora.

Maria Sara

Sara Fontdecaba y Maria Montón han cursado la Diplomatura, la Licenciatura y también el Máster en Estadística e Investigación Operativa en la UPC. Entre sus muchas cualidades está la de ser buenas estudiantes (fueron las dos primeras de su promoción). En el último curso de la diplomatura, Sara realizó prácticas en una empresa de estudios de mercado y María en el departamento de calidad de una empresa industrial. Las dos han colaborado como becarias para trabajos de consultoría en el Departamento de Estadística de la UPC. María también ha sido una pieza clave en la edición de este libro (es verdad, trabajan mucho).

16

Previsión de la duración de las etapas de la *Vuelta Ciclista a España*

Proyecto realizado por: **Román Peñas Cambray**
Dirigido por: **Alexandre Riba Civil**

Aunque últimamente se están produciendo algunos cambios, las grandes vueltas ciclistas son uno de los acontecimientos deportivos más populares en todo el mundo. El Tour de Francia, la prueba más importante de estas características, tiene una audiencia televisiva mundial sólo superada por dos acontecimientos deportivos: las Olimpiadas y los campeonatos del mundo de fútbol.

El Tour, sin embargo, no es la única prueba de estas características. Por ejemplo, TVE estimó que en el año 1997 (año en que se realizó este proyecto) la Vuelta en España era seguida diariamente por millones de personas en 142 países de los 5 continentes.

Las vueltas ciclistas por etapas tienen un gran problema: el tiempo de duración de cada etapa no es fijo, depende de la velocidad a que circulan los ciclistas. Malas previsiones en el tiempo de llegada dificultan la organización y perjudican su difusión. Algunos de los afectados son los medios de comunicación que tienen que alterar sus programaciones, los patrocinadores que ven disminuido su efecto publicitario, los organizadores a los que se complica el trabajo o los espectadores que no pueden ver aquello que resulta más interesante: el final de la etapa.

Las vueltas ciclistas

Las vueltas ciclistas son competiciones que se disputan por etapas diarias. Podríamos decir que es la suma de una serie de competiciones seguidas, pero con la peculiaridad que para poder tomar la salida un día has debido acabar la carrera del día anterior.

En el mundo del ciclismo profesional hay tres vueltas que destacan por encima del resto: el Tour de Francia, el Giro de Italia y la Vuelta a España. Cada una de ellas transcurre a lo largo de tres semanas, o para ser más concretos, de 23 días, en los cuales se disputan un total de 22 etapas y se recorren unos 3.800 kilómetros. En estos 22 días de competición (hay uno de descanso) hay etapas con características muy diferentes: las de montaña con orografía accidentada, las llanas donde se consiguen grandes velocidades medias o las contrarreloj que se realizan de forma individual.

Figura 1: *Las grandes vueltas ciclistas generan gran expectación*

Desde el punto de vista de la tipología de las etapas se diferencian dos tipos: las etapas en línea y las contrarreloj. En las etapas en línea (llanas o montañosas) todos los ciclistas salen al mismo tiempo, se realizan entre 100 y 270 kilómetros y se permite la colaboración entre diferentes competidores. En cambio, en las contrarreloj, sólo se realizan de 5 a 60 kilómetros, los ciclistas salen separados por intervalos de tiempo y no se permite la colaboración entre ciclistas, sino que se trata de un esfuerzo puramente individual.

Una gran vuelta moviliza diariamente a unas 2.500 personas, entre las cuales se incluyen periodistas, policías, organizadores y, naturalmente, participantes. En una competición de este estilo toman parte un máximo de 25 equipos profesionales, lo cual implica unos 225 corredores (9 por equipo).

El problema

Las previsiones del tiempo de llegada en las tres grandes competiciones ciclistas por etapas (Tour de Francia, Giro de Italia y Vuelta a España) se realizan a ojo. El director técnico de la prueba, una vez decididos los recorridos, prevé la velocidad media a qué circularán los ciclistas en base a su propia experiencia en etapas de características similares. Una vez fijada esta media se calcula el tiempo que tardarán en recorrer la etapa y la hora a la cual debe empezar para que finalice a una hora pactada con la televisión.

También se calculan los tiempos previstos de paso por todos aquellos puntos que la organización considera importantes (poblaciones, cruces, metas volantes, puertos de montaña...). En este aspecto hay organizaciones que se decantan por hacerlo de forma totalmente lineal (como la Vuelta) y otros que aplican una corrección según la orografía (como el Tour). De la misma forma se calculan tres tiempos de paso más: el horario rápido (con una velocidad media 2 km/h superior a la prevista), el horario lento (con una media 2 km/h inferior a la prevista) y el horario de la caravana publicitaria (normalmente una hora antes que el horario rápido).

El objetivo que se planteaba en esta parte del proyecto era obtener la máxima precisión en la previsión de los horarios de llegada en vueltas ciclista por etapas usando técnicas de modelado estadístico. Más concretamente se pretendía relacionar el tiempo del ganador de cada etapa con las características de la propia etapa (los kilómetros que tienen que realizar, los desniveles que tienen que superar, la categoría de los puertos que tienen que pasar ...), así como otras variables que la relacionan con el resto de la competición (la dureza de la etapa anterior y posterior, la importancia que a priori tiene la etapa para la clasificación general, los kilómetros acumulados desde el inicio de la competición...).

Recogida de datos. Selección de posibles variables explicativas

Para modelizar el comportamiento de los ciclistas en la Vuelta a España se han tenido en cuenta los resultados de los 6 años anteriores a la realización del proyecto, o sea, de las vueltas correspondientes a los años 1991 a 1996. Se ha escogido este número de años porque es un buen compromiso entre la cantidad de datos que se obtienen, 137, y poder garantizar que no se han producido grandes cambios en el comportamiento de los ciclistas. No es preocupante el hecho de que en los últimos años se corra más que antes, cosa que sería fácilmente cuantificable, lo realmente preocupante es cómo modelizar cambios de comportamiento más complejos, como por ejemplo la mayor especialización de los ciclistas actuales con respecto a los de décadas anteriores, por eso se ha escogido un periodo de tiempo no demasiado largo.

Los datos se han obtenido, básicamente, de las siguientes fuentes:

- Los libros de ruta de La Vuelta: Son los documentos que los organizadores reparten entre los directores de equipo, periodistas, policías locales y otras personas que intervienen en el acontecimiento. En ellos se detallan los horarios, recorridos, cotas, situación de zonas peligrosas, reglamentos, etc... (ver Figura 2).
- Revistas especializadas donde se publican los resultados y se comentan los incidentes más importantes de la carrera.

Debido a que los dos tipos de etapas existentes –las etapas en línea y las contrarreloj– tienen características muy diferentes no parece adecuado modelizarlas con el mismo modelo estadístico y, por lo tanto, se construirá uno para cada tipo de etapa. Las etapas más importantes, sin embargo, son las etapas en línea ya que son mucho más variables (con respecto a relieve, distancia...) y dependen mucho más del estado de los corredores. Por tanto, son también las más difíciles de modelizar y de prever. En este resumen se presentará sólo el resultado para este tipo de etapas.

La variable respuesta que se tiene que predecir es el tiempo, en minutos, que tarda en recorrer la etapa el ganador de la misma. Se tienen 105 observaciones del tiempo que varían entre los 150 y los 450 minutos aproximadamente, con una media de unos 300 minutos.

Como posibles variables explicativas se han considerado:

- Distancia de la etapa en kilómetros (km): Seguramente será una de las variables más influyentes sobre el tiempo que se tarda.
- Metros de diferencia (mdif): Diferencia entre la altitud del punto de llegada y la del punto inicial, medida en metros.
- Metros subidos (msub): Metros de desnivel de la etapa. El desnivel más pronunciado es de unos 4.000 metros mientras que hay etapas que se pueden considerar totalmente planas.
- Número de puertos de categoría especial (prtsE), de primera categoría (prts1), de segunda categoría (prts2) y tercera categoría (prts3).
- Última etapa (última): indica si la etapa es la última de la competición (la 22ª).
- Antes de contrarreloj (a_contra): indica si la etapa en cuestión es la anterior a una contrarreloj.

- Después de contrarreloj (d_contra): indica si la etapa precede a una contrarreloj.
- Antes de etapa de montaña (a_mon): indica si la etapa es la anterior a una etapa de montaña (msub > 2.000m).
- Después de montaña (d_mon): indica si la etapa es la siguiente a una de montaña.
- Semana: indica a qué semana corresponde la etapa (primera, segunda o tercera).

La Tabla 1 contiene el listado de estas variables candidatas a entrar en el modelo indicando el tipo de cada una de ellas. Las variables cuantitativas se pueden utilizar sin ninguna precaución especial, las dicotómicas pueden tomar sólo 2 valores diferentes que codificamos como 0 y 1. La variable semana, como es cualitativa, no se puede utilizar "tal como está" porque el modelo entendería que la semana 3 vale 3 veces la semana 1, y eso no es cierto. Para utilizar esta variable hay que introducir unas nuevas variables auxiliares que en la literatura estadística se acostumbran a denominar "variables dummy".

Tabla 1: *Lista de variables candidatas a entrar en el modelo*

Variable	Abreviatura	Tipos
Distancia de la etapa en kilómetros	km	Cuantitativa
Diferencia de altitud	mdif	
Metros subidos	msub	
Número de puertos de categoría especial	prtsE	
Número de puertos de primera categoría	prts1	
Número de puertos de segunda categoría	prts2	
Número de puertos de tercera categoría	prts3	
Última etapa	ultima	Dicotómica
Antes de la contrarreloj	a_contra	
Después de la contrarreloj	d_contra	
Antes de etapa de montaña	a_mon	
Después de la etapa de montaña	d_mon	
Semana	semana	Cualitativa

ALTITUD	ITINERARIO	Km. REC.	POR REC.	HORARIO PREVISTO 32 Km/h.	34 Km/h.	36 Km/h.
	COMUNIDAD DE CANTABRIA					
10	Salida lanzada 200 m. después gasolinera, dirección Burgos por N-623.	0,0	173,4	12:24	12:24	12:24
30	MURIEDAS.	2,7	170,7	12:29	12:28	12:28
20	Cruce. Giro a la izquierda dirección Bilbao-Astillero por S-436.	4,1	169,3	12:31	12:31	12:30
10	Cruce. Giro a la derecha dirección Sarón por S-432.	8,3	165,1	12:39	12:38	12:37
70	Parque de la Naturaleza de Cabarceno. **SPRINT ESPECIAL**	15,5	157,9	12:53	12:51	12:49
100	SARON. **META VOLANTE** Cruce. Giro a la izquierda dirección Solares-Bilbao por N-634.	18,9	154,5	12:59	12:57	12:55
80	PAMANES. Cruce. Giro a la derecha dirección Liérganes por S-553.	26,5	146,9	13:13	13:10	13:08
110	LIERGANES.	29,4	144,0	13:19	13:15	13:13
120	RUBALCABA.	32,0	141,4	13:24	13:20	13:17
190	MIRONES.	37,7	135,7	13:34	13:30	13:26
430	SAN ROQUE DE RIO MIERA.	46,3	127,1	13:50	13:45	13:41
1.350	**PUERTO DE LUNADA. PREMIO MONTAÑA 1.ª CATEGORIA**	62,0	111,4	14:20	14:13	14:07
	PROVINCIA DE BURGOS					
880	Cruce. Giro a la derecha dirección Vega de Pas por BU-570.	70,6	102,8	14:36	14:28	14:21
1.170	**PUERTO DE LAS ESTACAS DE TRUEBA PREMIO MONTAÑA 3.ª CATEGORIA**	80,0	93,4	14:54	14:45	14:37
	COMUNIDAD DE CANTABRIA					
370	VEGA DE PAS.	94,3	79,1	15:20	15:10	15:01
230	ENTRAMBASMESTAS. Cruce. Giro a la izquierda dirección Burgos por N-623. **AVITUALLAMIENTO.**	104,8	68,6	15:40	15:28	15:18
360	SAN ANDRES DE LUENA.	112,5	60,9	15:54	15:42	15:31
390	LOS PANDOS.	113,5	59,9	15:56	15:44	15:33
560	BOLLACIN.	115,5	57,9	16:00	15:47	15:36
1.010	**PUERTO DEL ESCUDO. PREMIO MONTAÑA 1.ª CATEGORIA**	120,8	52,6	16:10	15:57	15:45
	PROVINCIA DE BURGOS					
880	Cruce. Giro a la derecha dirección Reinosa por C-6318.	123,5	49,9	16:15	16:01	15:49
880	CORCONTE.	125,0	48,4	16:18	16:04	15:52
	COMUNIDAD DE CANTABRIA					
900	LA POBLACION DE YUSO. **SPRINT ESPECIAL.**	132,5	40,9	16:32	16:17	16:04
890	LA COSTANA.	137,5	35,9	16:41	16:26	16:13
880	MONEGRO. **META VOLANTE**	140,0	33,4	16:46	16:31	16:17
890	ORZALES.	142,5	30,9	16:51	16:35	16:21
880	REQUEJO.	147,5	25,9	17:00	16:44	16:29
880	Cruce. Giro a la izquierda dirección Reinosa-Palencia por N-611.	148,6	24,8	17:02	16:46	16:31
880	REINOSA. Cruce. Dirección Campoo por C-625.	149,4	24,0	17:04	16:47	16:33
890	NESTARES.	150,5	22,9	17:06	16:49	16:34
900	SALCES.	151,9	21,5	17:08	16:52	16:37
920	FONTIBRE.	154,2	19,2	17:13	16:56	16:41
950	ESPINILLA. Cruce. Dirección Brañavieja por C-628.	157,3	16,1	17:18	17:01	16:46
1.020	ENTRAMBASAGUAS.	162,2	11,2	17:28	17:10	16:54
1.640	**ALTO CAMPOO. (BRAÑAVIEJA)** META. **PREMIO MONTAÑA CATEGORIA ESPECIAL**	173,4	0,0	17:49	17:30	17:13

Sala de prensa y oficina permanente: Oficinas de estación de esquí.

LLEGADA: Ultimo kilómetro ligeras curvas y recta de 600 m. en subida del 6%.

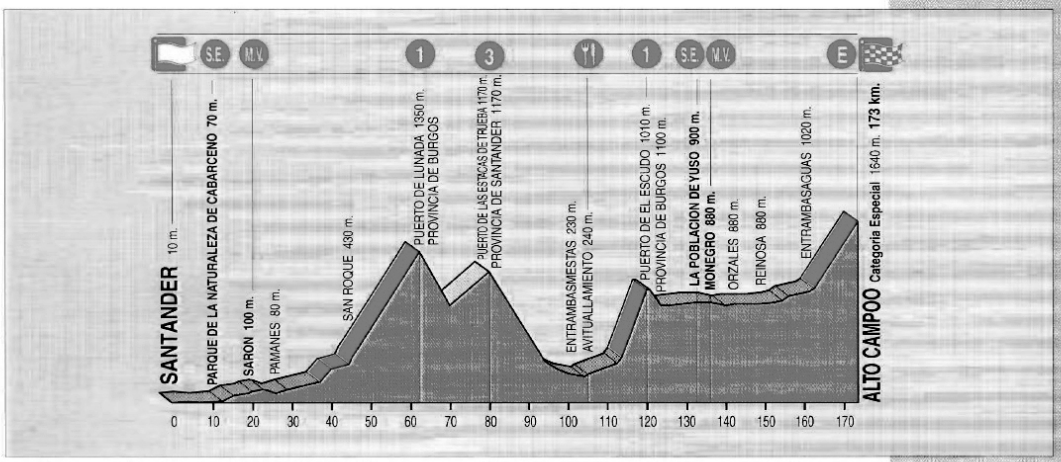
ATENCION: PASOS PELIGROSOS/BANDERA AMARILLA KMS.: 39,2 - 70,6 - 94,3 - 105,2.

MARTES
11 DE MAYO
DECIMOSEXTA ETAPA
16
(173,4 Kms.)
SANTANDER
ALTO CAMPOO

CONCENTRACION Y FIRMA
De 11:15 a 12:00 h. en la Plaza Porticada.

LLAMADA
A las 12:10 h.

SALIDA NEUTRALIZADA
A las 12:16 h., por Jesús de Monasterio, San Luis, San Fernando, Cuatro Caminos, Avda. Valdecilla, Avda. de Cajo y N - 623.
Total: 4,4 kms.



127

Figura 2: *Hoja del libro de ruta de la "Vuelta Ciclista a España 1993"*

Análisis exploratorio de los datos

Se empieza con un análisis exploratorio de los datos para identificar posibles valores anómalos (valores muy discrepantes, errores en la introducción de los datos...) y también para tener una primera impresión del comportamiento de los datos y de las relaciones entre ellos. Por ejemplo, la Figura 3 muestra la relación entre la longitud de la etapa (en km) y el tiempo que se tarda recorrerla (en minutos). Evidentemente existe una correlación muy clara entre ambas variables y también se puede ver que las etapas que tienen uno y, especialmente, dos puertos de categoría especial, tardan un tiempo que tiende a ser superior al de las etapas que recorren la misma distancia pero sin este tipo de puertos (Figura 4).

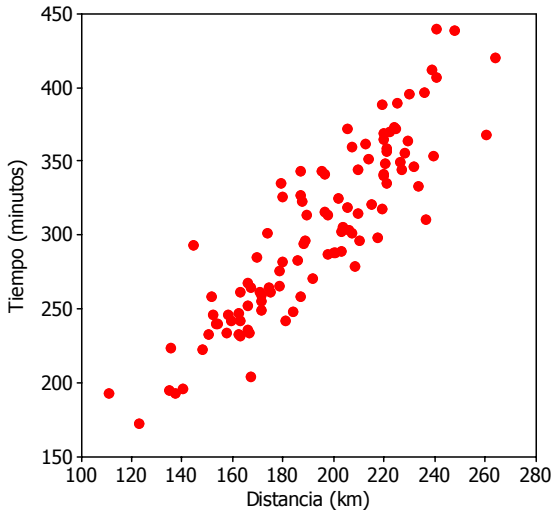


Figura 3: *Relación entre el tiempo que se tarda en recorrer la etapa y su número de km*

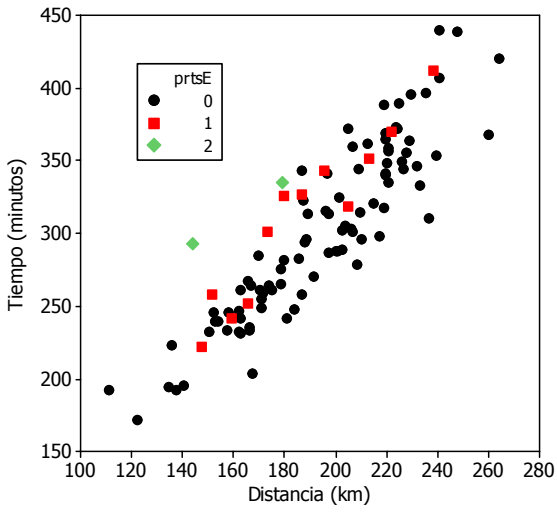


Figura 4: *Similar a la Figura 3 pero identificando las etapas por su número de puertos especiales*

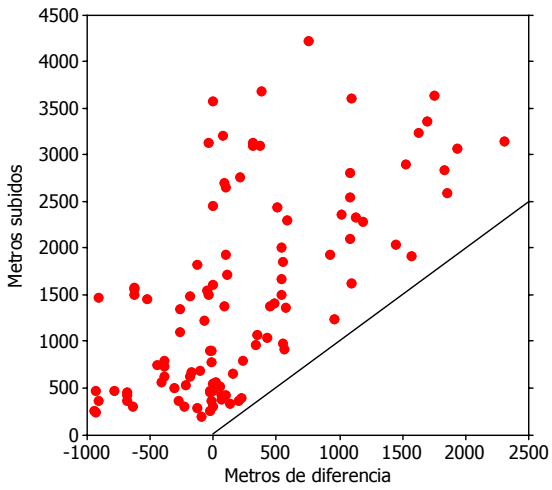


Figura 5: *Metros subidos en función de los metros de diferencia (cota de llegada menos cota de salida)*

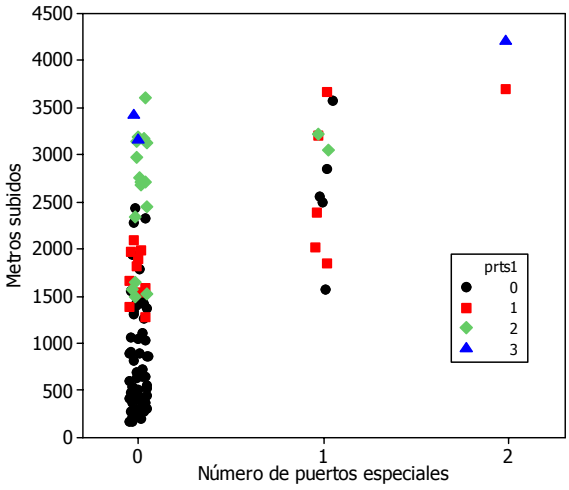


Figura 6: *Metros subidos en función del número de puertos especiales. Para cada grupo se indica también el número de puertos de 1ª categoría*

La Figura 5 muestra la relación entre los metros subidos y la diferencia entras las alturas de la cota de llegada y la de salida. Naturalmente, en una etapa no pueden ser menor el número de metros subidos que la diferencia de alturas, por eso no hay puntos por debajo de la línea trazada en el gráfico. En la Figura 6 se puede ser cómo los metros subidos están relacionados con el número de puertos especiales y también con el número de puertos de primera categoría.

Búsqueda de un buen modelo

Nos interesa un modelo con la máxima capacidad predictiva y lo más sencillo y fácil de utilizar que sea posible. Las variables que hemos seleccionado como posibles variables explicativas son eso: posibles, y habrá que estudiar cuáles conviene incluir y cuáles no, ya sea porque no aportan ninguna información sobre el comportamiento de la respuesta o porque son redundantes (lo que puede explicar una de ellas ya lo explican las otras).

También puede ser conveniente colocar como candidatas a formar parte del modelo transformaciones de las variables originales. Si en el análisis exploratorio de los datos se observa una relación cuadrática entre una variable y la respuesta, puede ser conveniente colocar esta variable elevada al cuadrado también como candidata, o también puede ser conveniente crear nuevas variables haciendo productos de las variables originales. Una vez identificado un primer modelo, ciertos tipos de comportamientos de los errores de predicción pueden aconsejar introducir también transformaciones logarítmicas de algunas variables, como se hizo en este caso.

Casi nunca el modelo más conveniente es aquél que incluye todas las variables en las que se ha pensado. Hace falta seleccionar qué subconjunto es el más conveniente de entre las variables originales y las transformadas. Para hacer esta selección es muy útil la utilización de programas de software estadístico. Hay dos grandes estrategias:

- Cálculo de todas las ecuaciones posibles: Es la estrategia de la fuerza bruta. Se calculan todas las ecuaciones posibles, se ordenan por criterios de calidad de ajuste y se escogen las que parecen más prometedoras para analizarlas a fondo y, si hace falta, introducir las modificaciones que convenga. El problema es que el número de ecuaciones posibles crece muy rápidamente: si k es el número de variables candidatas, el número de ecuaciones posibles es $2^k - 1$, y si k es grande (del orden de 30 o más) el tiempo de computación requerido hace inviable la aplicación de este método.

- Regresión paso a paso: Es una estrategia en que las variables entran (y también salen) del modelo una a una. La primera en entrar es la que está más correlacionada con la respuesta, la que al hacer una diagrama bivariante muestra una relación más estrecha (en nuestro caso sería la longitud de la etapa). Pero la segunda en entrar no es necesariamente la segunda que mejor explica el comportamiento de la respuesta (la segunda más correlacionada) sino la que mejor explica lo que falta por explicar, y de esta forma van entrando hasta que las que quedan fuera de ya no tienen una aportación significativa. Las variables van entrando pero también pueden salir si algunas de las que entran a continuación ya explican lo que explicaba una anterior.

  Un símil que ayuda a entender cómo funciona esta estrategia es pensar que uno se tiene que presentar a un examen en el que entran, por ejemplo, 100 temas, y no se sabe ninguno pero se puede llevar a los asesores que quiera de entre los compañeros de su clase y, puestos a suponer, se supone también que sabe cuales son los temas que conoce cada uno de ellos. Si sólo se pudiera llevar a uno ¿a cuál escogería? Naturalmente, al que más temas sepa (la variable que mejor explica la respuesta) ¿y si se pudiera llevar a dos? No necesariamente se llevaría a los dos que más sepan, porque podría pasar que el que más sabe conozca 80 temas y el segundo conozca 70, pero que éstos 70 ya están incluidos entre los que sabe el primero y, por lo tanto, el segundo no aporta nada. Será mejor uno que sólo conozca 15 pero que sean complementarios a los que conoce el primero. También puede pasar que uno conozca 50 y otro 40 diferentes, de forma que entre los dos conocen 90 y si entre éstos están los 80 que conoce el primero, éste primero no aportará nada y no hará falta que venga. Otra consideración es que si tenemos 100 compañeros, la mejor estrategia no será llevárnoslos todos, porque muchos no aportarán nada y sólo generarán ruido y confusión. Es mejor identificar a los pocos que mejor se complementan y pueden abarcar el mayor número de temas, y dejar a los que no saben nada o a los que saben lo que ya aportan los seleccionados.

  Existen diferentes variantes de este método. Su principal ventaja es que funciona rápidamente aunque el número de variables candidatas sea grande y orienta muy bien sobre qué tipo de ecuaciones tienen que ser exploradas con más detalle.

Realizando una primera exploración del tipo de modelos que parecían más prometedores, se llegó a una ecuación muy sencilla y con un notable poder de predicción:

$$\text{Tiempo} = -39{,}9 + 1{,}70\ \text{km} + 0{,}00756\ \text{msub} + 0{,}00545\ \text{msub} \cdot \text{prts\_E}$$

En este modelo el tiempo se explica en función del número de kilómetros de la etapa, los metros subidos y una nueva variable que es el producto entre los metros subidos y el número de puertos de categoría especial.

Pero el modelo que se consideró que tiene las mejores capacidades incluye las transformaciones logarítmicas del tiempo y de la longitud de la etapa:

$$\ln(\text{tiempo}) = 1{,}08\ln(\text{km}) + 0{,}000021\ \text{msub}\cdot\text{prts\_E} + \\ + 0{,}000007\ \text{msub}\cdot\text{prts1} + 0{,}00361\ \text{m\_dif/km}$$

Aun así, antes de dar este modelo como definitivo hay que mirarlo a fondo para descubrir posibles imperfecciones que habrá que intentar solucionar.

Verificación del modelo: Análisis de los residuos

El modelo sirve para estimar el valor de la respuesta en función de los valores de las variables explicativas. La diferencia entre el valor estimado y el valor real es lo que se llama residuo. Por ejemplo, en la primera etapa del año 1991 (la primera de la que tenemos datos) los valores de las variables que aparecen en el modelo son:

Variable	Abreviatura	Valor
Recorrido de la etapa en kilómetros	km	134,5
Diferencia de altitud	mdif	220
Metros subidos	msub	390
Número de puertos de categoría especial	prtsE	0
Número de puertos de primera categoría	prts1	0

La distancia recorrida se introduce en el modelo a través de su logaritmo y lo que se obtiene, junto con las otras variables, es el logaritmo del tiempo, en este caso: 5,308. Para tener el tiempo en minutos hay que deshacer la transformación: $e^{5{,}308} = 201{,}90$ minutos.

Como el valor real de la duración de esta etapa fue de 193,93 minutos, su residuo es:

Residuo: Valor real – Valor previsto = 193,93 – 201,90 = -7,97 minutos

Los residuos son la parte de la respuesta que el modelo no es capaz de explicar y hay que asegurarse de que toman valores totalmente aleatorios. Si siguen algún patrón de comportamiento que permita sacar alguna información, ésta tendrá que ser aprovechada e incorporada al modelo. La Figura 7 muestra un gráfico en el que cada

punto corresponde a una etapa y relaciona los residuos con los valores previstos estratificando por año (cada punto se representa de forma diferente según el año).

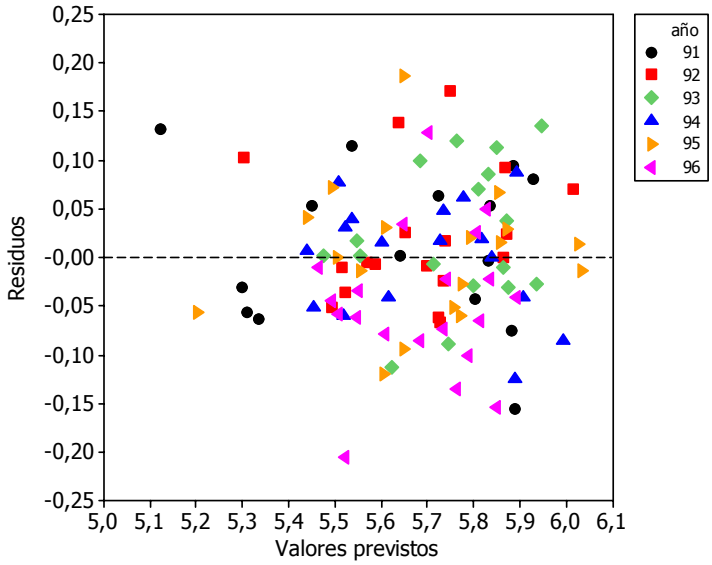


Figura 7: *Gráfico de los residuos frente a los valores previstos estratificados por año*

Se observa que los residuos correspondientes al año 1996 aparecen mayoritariamente por debajo de cero. Exactamente, de las 20 etapas de aquel año, sólo en 4 de ellas el residuo es positivo y la probabilidad de que eso ocurra por azar (sólo 4 residuos o menos por encima de cero) es del orden del 6 por mil. Si el modelo es adecuado, los residuos tienen la misma probabilidad de ser mayores o menores que cero (es igualmente probable que el error se cometa por exceso o por defecto). Si sistemáticamente el modelo da errores por defecto quiere decir que no apunta bien, y habrá que reajustarlo.

Este reajuste se ha realizado creando una nueva variable, año_96, que vale 0 para todas las etapas excepto para las del año 1996, que valdrá 1. El nuevo modelo obtenido es:

$$\begin{aligned}\ln(\text{temps}) = {} & 1{,}08\ln(\text{km}) + 0{,}000020\ \text{msub} * \text{prtsE} + \\ & + 0{,}000007\ \text{msub} * \text{prts1} + 0{,}00364\ \text{mdif/km} \\ & - 0{,}0597\ \text{año\_96}\end{aligned}$$

Obsérvese que lo que hace el último término es introducir una corrección de -0,0597 al logaritmo del tiempo en las etapas del año 1996. Para las etapas de otros años, este término es como si no estuviera (la variable vale 0) y, en consecuencia, está

disminuyendo sistemáticamente el tiempo de duración de la etapa en caso de que se trate de una etapa del año 1996.

¿Hemos encontrado un buen modelo?

En primer lugar hay que verificar que el modelo da resultados coherentes de acuerdo con lo que se podía esperar. Por ejemplo, es razonable que la velocidad media sea menor en las etapas largas que en las cortas.

Utilizando el modelo se ha construido la Figura 8, que muestra la velocidad media por etapa en función de su longitud, para etapas sin puertos y con 0 metros de diferencia. Efectivamente, la velocidad disminuye con la distancia, y el año 1996 es unos 2,5 km/h superior a los años anteriores.

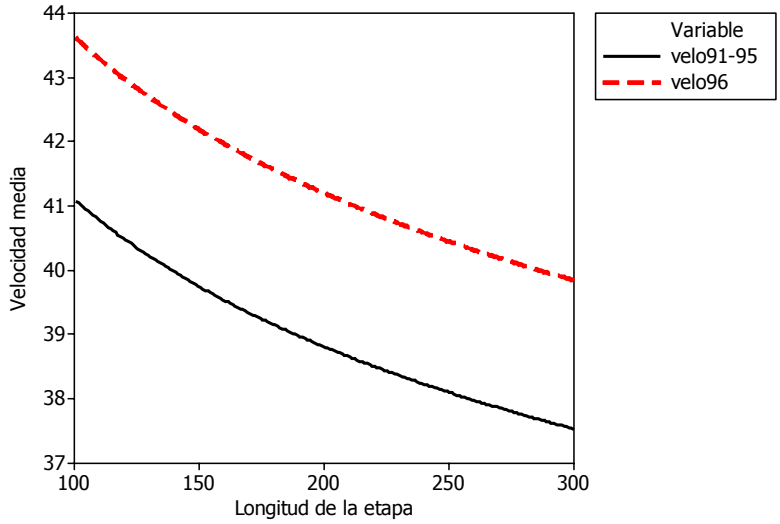


Figura 8: *Variación de la velocidad media en función de la longitud de la etapa, según el modelo encontrado para etapas sin puertos y con 0 metros de diferencia.*

Con respecto a las medidas que valoran la calidad de un modelo estadístico, hay diferentes tipos, pero en este caso una buena forma de valorarlo será comparando sus previsiones con las que da el experto. Si las previsiones del modelo son mejores, éste será útil, de lo contrario no habremos sabido superar al "modelo" del experto, porque no se ha sabido sacar toda la información que tienen los datos, o porque no se tiene suficiente cantidad (el experto utiliza datos o información que no se han considerado para crear el modelo).

Si comparamos las previsiones del modelo con las que hace el experto para las 105 etapas de las que tenemos datos, gana el modelo. La suma de los errores (o de los

residuos, en valor absoluto) del modelo es 1.747,88, mientras que la suma de errores del experto es de 2.119,30. Pero aquí hay trampa. Estas previsiones se realizado para los mismos valores utilizados para calcular el modelo, y no tiene mérito ser capaces de explicar lo qué ya se sabe. El modelo tiene que demostrar su utilidad haciendo previsiones en etapas nuevas, que no hayan servido para calcularlo.

Por ejemplo, el modelo obtenido utilizando solo las etapas de los años 1991 a 1994 es:

$$\begin{aligned}\ln(\text{temps}) = {} & 1{,}08\,224\,\ln(\text{km}) + 0{,}00001682\ \text{msub}\cdot\text{prtsE} + \\ & + 0{,}00000801\ \text{msub}\cdot\text{prts1} + 0{,}004681\ \text{m\_dif/km}\end{aligned}$$

Utilizando este modelo para hacer las previsiones del año 1995 (ahora ya no se hace trampa), la suma de los errores que da el modelo es 276,54 mientras que la suma de errores del experto es 307,71. Por lo tanto, si se hubiera utilizado este modelo para el año 1995 los errores de previsión habrían sido menores que los que realmente existieron utilizando las previsiones oficiales. Pero también se tiene que decir que las diferencias son pequeñas y un test de comparación de las medias de los residuos pone de manifiesto que estas diferencias no son estadísticamente significativas.

Finalmente, utilizando los datos de las etapas de 1991 a 1995 se obtiene el modelo:

$$\begin{aligned}\ln(\text{temps}) = {} & 1{,}08183\ \ln(\text{km}) + 0{,}00001919\ \text{msub}\cdot\text{prts\_E} + \\ & + 0{,}00000764\ \text{msub}\cdot\text{prts\_1} + 0{,}004149\ \text{m\_dif/km}\end{aligned}$$

En el que no se ha incorporado el término de corrección del año 1996 porque sólo se puede conocer cuando ya se tienen los datos de 1996. Utilizando este modelo para hacer la predicción de la duración de las etapas de aquel año, la suma de los errores que provoca el modelo es de 457,98 mientras que la suma de los errores de las previsiones oficiales fue 574,13. Y haciendo un test de comparación de las medias de los errores, ahora la diferencia sí sale estadísticamente significativa. Para hacer estas predicciones sí que podemos decir que el modelo es mejor.

Más cosas que se podrían hacer…

Una variable explicativa que resultaría muy interesante incluir en el modelo es la meteorología, ya que es evidente que el viento o la lluvia influyen en el desarrollo de una etapa. El problema, sin embargo, es que no se dispone de datos fiables para poder cuantificar esta influencia. Aunque la meteorología no se pueda prever, sabiendo el efecto que ésta ha tenido en el pasado, podríamos descubrir los efectos de otras variables que sí son predictibles. Poner en el mismo saco etapas que se han disputado en condiciones pésimas y otras que se han disputado en condiciones óptimas nos

puede impedir ver el efecto de otras variables obvias, como por ejemplo que el último día de una gran vuelta los ciclistas siempre circulan con más lentitud de la habitual.

Puede ser interesante aplicar herramientas estadísticas a las predicciones de los tiempos de paso. En la actualidad hay organizaciones que calculan los tiempos previstos de paso de forma totalmente lineal (cómo es el caso de la Vuelta) y otros que aplican una corrección según la orografía (como el Tour). El objetivo sería intentar hacer las predicciones de los tiempos de paso con dos correcciones diferentes: la orográfica y la distancia a la meta. De esta forma se facilitaría la organización de la carrera (cortes de carreteras, caravanas publicitarias...) pero sobre todo se podría facilitar a los medios de comunicación, a determinados kilómetros antes de finalizar la etapa, una estimación muy precisa de la hora de llegada. Es decir, se podrían ir actualizando las previsiones a medida que se tuviera nueva información.

También sería interesante caracterizar las etapas que aportan más variabilidad al tiempo de llegada; es decir, aquéllas que aportan más error, de esta manera obtendríamos una radiografía de las etapas con un tiempo de llegada más impredecible y se podrían tomar las decisiones oportunas.

(Proyecto de la Diplomatura de Estadística, presentado en septiembre de 1997 con el título "Resolució de problemes reals mitjançant el modelatge estadístic")

Se han disputado muchas Vueltas, Tours y Giros desde que se hizo este proyecto, y también han pasado algunas cosas en torno al mundo del ciclismo, pero la afición y el seguimiento de los medios de comunicación continúa siendo muy importante.

A pesar de su antigüedad, y de que los datos ya están obsoletos (sería interesante ver cómo cambian las cosas con los datos actuales) nos ha gustado incorporar este proyecto porque es un buen ejemplo del proceso que se sigue en la construcción de un modelo de regresión. Estos datos se utilizan todavía para que los estudiantes de la Diplomatura de Estadística los analicen y discutan sus posibilidades para la construcción de modelos de regresión.

Román Peñas cursó la Diplomatura de Estadística en la UPC y una Licenciatura en *Estadística e Investigação Operacional* en la Universidad de Lisboa. Empezó a trabajar en el Departamento de Marketing Estratégico del Deutche Bank, explotando la base de datos de clientes mediante técnicas de modelado estadístico. Después se incorporó a la empresa Masquelack, realizó un MBA (*Master in Bussines Administration*) en el IESE y actualmente es su gerente. En su época de estudiante no solo realizaba trabajos académicos sobre temas deportivos, también practicaba deporte a nivel competitivo (pentatlón moderno, atletismo de fondo y ciclismo amateur), pero nos explica que ahora sus responsabilidades profesionales y familiares le obligan a tener un poco abandonada esa faceta deportiva.