PECVNIA

Revista de la Facultad de Ciencias Económicas y Empresariales
Universidad de León

María Jesús Mures Quintana (Coord.)

Monográfico 2012

Estadística Aplicada a la Investigación Cuantitativa
Applied Statistics to Quantitative Research

PECVNIA

Revista de la Facultad de Ciencias Económicas y Empresariales
Universidad de León

María Jesús Mures Quintana (Coord.)

Monográfico 2012

Estadística Aplicada a la Investigación Cuantitativa
Applied Statistics to Quantitative Research

Redacción y correspondencia:
PECVNIA
Facultad de Ciencias Económicas y Empresariales
UNIVERSIDAD DE LEÓN
Campus de Vegazana, s/n
24071 León (España)
u l e p e c @ u n i l e o n . e s
http://www3.unileon.es/pecvnia/pecvnia.htm

ISSN: 1699-9495 (Ed. impresa)
ISSN 2340-4272 (Internet)
Depósito Legal: LE-1514-2005
Maquetación: Pilar Fernández Cañón

PECVNIA
Revista de la Facultad de Ciencias Económicas y Empresariales
Universidad de León

ISSN 1699-9495

Estadística Aplicada a la Investigación Cuantitativa
Applied Statistics to Quantitative Research Monográfico 2012

SUMARIO

Prólogo

En el año 2013 se ha celebrado el Año Internacional de la Estadística (*Statistics2013*), que diversas Asociaciones Estadísticas marcaron para conmemorar el segundo centenario de la publicación póstuma de la obra de J. Bernouilli, *Ars Conjectandi*, en 1713. La edición apareció ocho años después de su muerte y estuvo a cargo de su sobrino, Niklaus Bernoulli.

Previamente, se había celebrado el Día Mundial de la Estadística, el 20 de octubre de 2010, promovido por la Comisión de Estadística de las Naciones Unidas y aprobado por la Asamblea General de las Naciones Unidas el 3 de junio de 2010.

Ambas celebraciones han tenido como objetivo fundamental destacar la importancia de la Estadística en el avance del conocimiento y en el desarrollo económico y social. Concretamente, en el caso de *Statistics2013*, ha supuesto un reconocimiento a quienes han contribuido al desarrollo de la Estadística, siendo los principales objetivos promulgados por la organización los de aumentar la conciencia pública sobre el poder y el impacto de la Estadística en todos los ámbitos de la sociedad; fomentar la Estadística como profesión, especialmente entre los jóvenes; y promover la creatividad y el desarrollo en las ciencias de la Probabilidad y la Estadística.

El monográfico que se presenta constituye la aportación del Área de Estadística e Investigación Operativa de la Universidad de León al desarrollo de la Estadística, en concordancia con los objetivos planteados en *Statistics2013*. Con los artículos seleccionados se pretende divulgar el interés que presenta esta ciencia, de la que tanto se habla, desde una perspectiva teórica, pero fundamentalmente sobre sus múltiples aplicaciones. Al lado de las aportaciones de los investigadores de la Universidad de León figuran otras pertenecientes a autores que desarrollan sus pesquisas en la misma dirección que las nuestras.

Conscientes de que uno de los fines primordiales de la Estadística, como ya es sabido, es dar a conocer su carácter de instrumento para la práctica totalidad de todas las ciencias, no se puede obviar el hecho de que la base teórica es necesaria e imprescindible. Por esta razón, se presentan trabajos en esta publicación acordes con esa línea. En relación a las aplicaciones, éstas se concretan especialmente en el ámbito de la empresa, uno de los campos científicos en los que la Estadística tiene una especial relevancia. Con carácter divulgativo, se ha incluido también un trabajo relativo a los aspectos técnicos de las estadísticas oficiales más relevantes que se elaboran en España por parte del Instituto Nacional de Estadística; este trabajo se presenta exclusivamente en el idioma castellano dadas sus peculiaridades y su carácter estrictamente relativo al estado español. El funcionamiento concreto del mencionado Instituto lo convierte en un caso concreto y preciso cuyo transporte a la lengua inglesa hubiera acarreado errores que hubieran hecho de este artículo más una rémora que un activo.

El Área de Estadística e Investigación Operativa agradece a la Facultad de Ciencias Económicas y Empresariales de la Universidad de León la publicación de este número monográfico, así como a todos los autores por sus contribuciones, que indudablemente abren nuevos y constantes retos científicos y técnicos.

Prologue

The International Year of Statistics *(Statistics2013)* has been celebrated in 2013, organized by various statistical associations in remembrance of the second centenary of J. Bernouilli's posthumous publication *Ars Conjectandi* (1713). The edition appeared eight years after his death and was the responsibility of his nephew, Niklaus Bernouilli.

Previously, the World Day of Statistics had been celebrated on 20 October 2010, this time it was organized by the United Nations Statistical Commission and passed by the General Assembly on 3 June 2010.

Both celebrations have had the same aim: Highlighting the importance of Statistics in the progress of knowledge and in the economic and social development. In particular, *Statistics2013* has implied a sincere homage to those who have contributed to the development of Statistics; the main targets made public by the organization were enlarging public consciousness about the power and impact of Statistics in every field of the society; encouraging Statistics as a profession, mainly among the younger generations; and promoting creativity and development in the sciences of Probability and Statistics.

The Monograph here presented is the contribution of those researchers working in the field of Statistics and Operations Research at the University of León according to the outlined targets in *Statistics2013*. With the selected papers we are trying to spread the interest this particular science poses, so much talked about, both theoretically and mainly in its multiple applications. Next to the contributions by the researchers from the University of León, there are some others belonging to authors who also work in the same fields and orientations.

Conscious as we are about the main targets of Statistics which, as it is very well known, is letting people know it is a good tool for almost every science, it cannot be taken for granted that the theoretical basis is both necessary and essential. That is why some of the papers presented in here go in that direction. As far as applications are concerned, they are very centered upon the business world, one of the leading scientific fields in which Statistics is of an uttermost relevance. The volume also contains a paper referred to technical aspects of the most relevant official statistics worked out in Spain by the Spanish Institute of Statistics. This paper is only presented in Spanish due to its characteristics and because it is only devoted to the Kingdom of Spain. The particular way of working of this Institute recommended to present it only in Spanish; otherwise, a translation into English might be misleading in its orientation and misguiding in its findings.

The researchers in Statistics and Operations Research want to take this opportunity to thank the Faculty of Economics and Business Studies at the University of León for its support towards the publication of this monograph and also to the authors for their contributions which, no doubt, open new and constant challenges both scientifically and technically.

Pecvnia Monográfico 2012, pp. 1-20

MIDIENDO LA VARIABILIDAD EN CARACTERES CUALITATIVOS / *MEASURING VARIABILITY IN QUALITATIVE CHARACTERISTICS*

Jesús Basulto Santos[1]
basulto@us.es
José Antonio Camúñez Ruiz[1]
camunez@us.es
Francisco Javier Ortega Irizo[1]
fjortega@us.es
María Dolores Pérez Hidalgo[1]
mdperez@us.es

Universidad de Sevilla

Resumen

El estudio de la variabilidad en caracteres categóricos rara vez es abordado. A partir de un enfoque menos usado de la variabilidad en variables cuantitativas, el de la disparidad, distinto al de la dispersión que, por ejemplo, proporciona la varianza, se propone la construcción de dos coeficientes de medida de la variabilidad en variables cualitativas o categóricas a los que llamamos coeficientes de disparidad. La sencillez y proximidad de los mismos permiten que sean abordados en un curso introductorio de estadística descriptiva. Ejemplos sencillos son ofrecidos para introducir las medidas y para, también, que el profesor constate la idea que el alumno tiene sobre variabilidad, dispersión y disparidad.

Palabras clave: Variables cualitativas o categóricas; Variabilidad; Dispersión; Disparidad.

Abstract

The study of variability in categorical characteristics is rarely discussed. From a less used viewpoint of variability in quantitative variables, as it is the one of dissimilarity, which is different from the dispersion that, for example, the variance provides, we propose the construction of two coefficients that measure the variability in qualitative or categorical variables, which we call

[1] Departamento de Economía Aplicada I. Facultad de Ciencias Económicas y Empresariales, Universidad de Sevilla, Avda. Ramón y Cajal, 1, 41018-Sevilla.

coefficients of dissimilarity. Simple examples are provided to introduce the measures, so that the teacher can also evaluate the idea students have about variability, dispersion and dissimilarity.

Keywords: Qualitative or categorical variables; Variability; Dispersion; Dissimilarity.

1. INTRODUCCIÓN

Las variables cualitativas o categóricas siempre han ocupado un mínimo espacio en los cursos introductorios de estadística. Se suelen definir, clasificar en nominales u ordinales, introducir la moda como una medida representativa y, en el caso de las ordinales, alguna medida similar a la mediana. También, representarlas gráficamente, siendo en este aspecto donde, quizás, encontramos más variedad de propuestas: diagramas de barras, de sectores, pictogramas, y una pluralidad de gráficos cuyo nivel de sofisticación depende, casi, de la imaginación de la persona interesada. La media aritmética, la que presenta mayores posibilidades de manipulación algebraica, la más conocida y utilizada, la medida por antonomasia en variables cuantitativas, no dispone de su equivalente entre las categóricas.

Prácticamente, nuestro trabajo en el aula se reduce a lo que acabamos de citar en el caso del estudio de una variable categórica aislada. Después, al tratar con dos variables cualitativas relacionadas entre si, las tablas de contingencia, con sus medidas asociadas, amplían un poco la visión sobre este tipo de estadísticas.

Desde luego, la variabilidad, (cualidad de variable, según el diccionario de la Real Academia Española) tan profusamente estudiada en cuantitativas, no es tratada en general en las categóricas, dando la sensación, entonces, de que este tipo de

1. INTRODUCTION

Qualitative or categorical variables have always been residually dealt with in introductory statistics courses. These courses usually include their definition, classification into nominal or ordinal variables, the presentation of the mode as a representative measure and, in the case of the ordinal variables, other kind of measures similar to the median. They are also graphically represented, following a variety of options: bar chart, pie chart, pictograms, and a diversity of charts whose level of sophistication, it can be said, depends on the imagination of the person in question. The arithmetic mean, which presents the largest possibility of being algebraically manipulated, which is the most known and used and the measure *par excellence* in the case of quantitative variables, has no counterpart in the case of the categorical ones.

From a practical point of view, our work in the classroom is reduced to what we have just mentioned in the case of the study of a separate categorical variable. After that, when dealing with two related categorical variables, the use of contingency tables and their associated measures allow spreading a bit the idea of this kind of statistics.

Certainly, variability –defined as the quality of variable, according to the Academy of Spanish Language (RAE)–, which has been so profusely studied in the case of quantitative variables, is not usually dealt with for the categorical ones, which seems to mean that this type of

medidas no existe. Es claro que esa idea de variabilidad alrededor de la media, significado habitual que damos a varianza o desviación típica, no tiene sentido. Se suele usar el término "dispersión" para esta forma de variabilidad.

Pero hay otra manera de entender la variabilidad, la que se detiene en el análisis comparativo de respuestas donde la comparación se reduce a igualdad o desigualdad de las mismas, sin pararse en medir la magnitud de esa desigualdad. Podemos usar en este caso el término "disparidad" (desemejanza, desigualdad y diferencia de unas cosas respecto de otras, según el diccionario de la Real Academia Española). Estas medidas, que se emplean aunque con menos frecuencia en variables cuantitativas, pueden extenderse a las cualitativas, pues la disparidad existe siempre que se manifiesten opiniones distintas. O sea, la variabilidad existe en las categóricas (no tendría sentido cualquier estudio estadístico si no fuese así). Creemos que es algo que debemos inculcar a nuestros alumnos y que, si es posible, construir medidas o indicadores de dicha variabilidad.

En este trabajo presentamos un par de medidas sencillas para casos categóricos (aunque en concepto podríamos hablar de una sola, dado que la diferencia entre ambas es la misma que la existente entre varianza y cuasivarianza), a las que proponemos llamar "coeficientes de disparidad", y las aplicamos a ejemplos sencillos que nos permiten observar, en el aula, si la percepción de variabilidad que tienen nuestros estudiantes es coherente con la que mide estos coeficientes.

En algunos trabajos hemos comprobado la utilidad de estas medidas que, acompañada de lo intuitivas que resultan, creemos, deben ser medidas que engrosen el contenido de una asignatura dedicada a Estadística Descriptiva.

measures does not exist. It is clear that the idea of variability around the mean, which is the usual meaning given to the variance or standard deviation, makes no sense in the case of categorical variables. For this kind of variability, the term 'dispersion" is generally used.

However, there is another way of understanding variability, which is the one that focuses on the comparative analysis of responses, where the comparison is reduced to their similarity or disparity, but it does not deal with measuring the amount of disparity. In this case, the term 'dissimilarity' can be used (which is defined as disparity, inequality or difference of some things with regard to others by the Academy of Spanish Language). These measures, which are used for quantitative variables but less frequently, can be spread to the qualitative ones, since dissimilarity exists as long as there are different options. That is, variability exists in the case of categorical variables (otherwise, any statistical analysis would make no sense). We think this is something we must instil in our students and, if possible, construct measures or indicators of the above-mentioned variability.

In this paper we present a couple of simple measures for categorical cases (although strictly speaking we could talk about only one, since the difference between them is the same as the one between variance and quasivariance), which are proposed to be named as 'dissimilarity coefficients' and we apply them in simple examples which allow us to observe in the classroom if the perception of variability our students have is coherent with the one these coefficients measure.

We have checked in some papers the usefulness of these measures, which, together with the fact of being so intuitive, must widen the contents of a subject in Descriptive Statistics.

Dado que en variables categóricas la proporción de respuestas en un sentido u otro es uno de los primeros cálculos que realizamos y que, la idea de proporción enlaza con la de probabilidad para el caso de variables aleatorias, terminamos analizando la similitud entre una de las medidas propuestas y la varianza de una variable probabilística dicotómica tipo Bernoulli.

2. VARIABILIDAD EN CUANTITATIVAS: DISPERSIÓN Y DISPARIDAD

En variables cuantitativas nos encontramos como primeras medidas de dispersión la varianza y la cuasivarianza, cuyas definiciones recordamos: $S^2 = \frac{\sum_i (x_i - \bar{X})^2}{n}$ y $S_c^2 = \frac{\sum_i (x_i - \bar{X})^2}{n-1}$, respectivamente. Gini (1912), cuando estudia la variabilidad entre las cuantitativas distingue dos tipos de variables: las que se definen como un sólo valor real, μ, pero que al ser medido se producen diferentes mediciones debido a los errores asociados a las mismas, por lo que los valores observados u observaciones efectuadas son de la forma $x_i = \mu + \varepsilon_i$ (habla de variables relacionadas con la medición en astronomía), y las que presentan distintas modalidades cuantitativas que van surgiendo con las repetidas observaciones de las variables. Pues bien, para el primer tipo, Gini (1912) propone medidas del tipo de las citadas anteriormente, o sea, medidas de dispersión alrededor de la media (siendo ésta el valor real de la variable), mientras que para las del segundo formula medidas que recojan todas las posibles diferencias, por parejas, entre los valores observados. Serían, pues, medidas construidas a partir de los siguien-

Provided in categorical variables the proportion of responses in one and another sense is one of the first computations that are carried out, this idea of proportion is connected with the one of probability, so we conclude with analysing the similarity between one of the proposed measures and the variance of a Bernoulli dichotomous random variable.

2. VARIABILITY IN QUANTITATIVE: DISPERSION AND DISSIMILARITY

In quantitative variables, the first measures of dispersion are variance and quasi-variance, whose expressions are reminded: $S^2 = \frac{\sum_i (x_i - \bar{X})^2}{n}$ and $S_c^2 = \frac{\sum_i (x_i - \bar{X})^2}{n-1}$, respectively. When Gini (1912) studies variability in quantitative variables, he distinguishes between two types of variables: those which are defined as an only real value, μ, but when this is measured there are different measurements due to mistakes associated to the former, so the observed values or observations are in the form of $x_i = \mu + \varepsilon_i$ (actually, he talks about variables related to the measurement in astronomy); and those which present different qualitative categories that arise with the repeated observations of the variables. In this context, for the first type, Gini (1912) proposes measures which are similar to the ones we have mentioned above, that is, measures of dispersion around the mean (which is the real value of the variable), whereas for the second group he formulates measures that include all the possible pairwise differences among observed values. They would be, therefore, measures that would be constructed from the following expressions:

tes agregados: $\sum_i\sum_j (x_i - x_j)^2$, $\sum_i\sum_j |x_i - x_j|$, (las distancias entre observaciones son medidas mediante diferencias al cuadrado o diferencias en valor absoluto) donde este autor apuesta más por el segundo que por el primero, pues la que propuso es la conocida como media de las diferencias:

$$\Delta = \frac{\sum_i\sum_j |x_i - x_j|}{n(n-1)}.$$

Para el primer agregado es fácil demostrar la siguiente igualdad:

$$\sum_i\sum_j (x_i - x_j)^2 = 2n\sum_i (x_i - \overline{X})^2 .$$

De alguna forma, esta igualdad genera conciliación, tanto sobre la varianza como sobre la cuasivarianza, entre las dos formas de observar la dispersión desde los dos tipos de variables, según Gini (1912).

Podemos construir dos nuevas medidas usando el agregado del primer miembro de la igualdad anterior, a los que podemos llamar, por ejemplo, "promedios cuadráticos de diferencias por pares" y que definimos a continuación:

$$V^2 = \frac{\sum_i\sum_j (x_i - x_j)^2}{n^2} \text{ y } V_c^2 = \frac{\sum_i\sum_j (x_i - x_j)^2}{n(n-1)}$$

La igualdad de arriba nos permite escribir: $V^2 = 2 \cdot S^2$, $V_c^2 = 2 \cdot S_c^2$.

En todas las medidas citadas hasta ahora la variabilidad depende de dos factores, del número de valores diferentes que nos encontremos y de la distancia entre los mismos (influida por la magnitud de los correspondientes valores).

$\sum_i\sum_j (x_i - x_j)^2$, $\sum_i\sum_j |x_i - x_j|$, (the distances among observations are measured as squared differences or differences in absolute value), although this author banks on the first one rather than the second, since he proposed the measure known as differences mean:

$$\Delta = \frac{\sum_i\sum_j |x_i - x_j|}{n(n-1)}.$$

For the first expression it is easy to prove the following equality:

$$\sum_i\sum_j (x_i - x_j)^2 = 2n\sum_i (x_i - \overline{X})^2$$

Somehow, this equality makes agreement come, on both the variance and the quasivariance, about the two ways of observing the dispersion from both types of variables, according to Gini (1912).

We can construct two new measures using the first side in the previous equality, which can be called, for example, 'squared means of pairwise differences" and which are defined as follow:

$$V^2 = \frac{\sum_i\sum_j (x_i - x_j)^2}{n^2} \text{ and } V_c^2 = \frac{\sum_i\sum_j (x_i - x_j)^2}{n(n-1)}.$$

The above equality allows us to write the expressions as: $V^2 = 2 \cdot S^2$, $V_c^2 = 2 \cdot S_c^2$.

In all the aforementioned measures, variability depends on two factors: on the number of different values that can be found and on the distance among them

Dos valores, x_i y x_j, que estén muy separados entre sí, por ser dos cantidades muy distintas, aportan mucho peso a la hora de calcular la dispersión mediante cualquiera de esas medidas. Serían éstas las que al principio hemos llamado "medidas de dispersión".

Ahora, podemos plantearnos la variabilidad sólo desde el punto de vista de la disparidad, del número de posibles parejas de componentes distintos que se pueden formar, lo que depende del número de valores distintos que presente una variable, sin tener en cuenta la magnitud de dichos valores. Así, bajo este punto de vista se nos ocurren dos posibles medidas a las que podemos llamar "coeficientes de disparidad" (Perry y Kader, 2005):

(affected by the magnitude of the respective values). Two values, x_i and x_j, which are very separated from each other, as they are two very different quantities, present a lot of weight in order to calculate dispersion through any of those measures. These would be what at the beginning we have called 'measures of dispersion'.

In this point, we can consider variability from the viewpoint of dissimilarity, of the number of possible pairs of different components, which depends on the number of different values a variable presents, without taking into account their magnitude. Thus, from this point of view, two measures can be defined, which can be called 'coefficients of dissimilarity' (Perry and Kader, 2005):

$$D_1 = \frac{\sum_i \sum_j c\left(x_i, x_j\right)}{n^2} \text{ y/}and\ D_2 = \frac{\sum_i \sum_j c\left(x_i, x_j\right)}{n(n-1)}, \text{ con/}being\ c\left(x_i, x_j\right) = \begin{cases} 1, \text{ si } x_i \neq x_j \\ 0, \text{ si } x_i = x_j \end{cases}.$$

Por tanto, el numerador de estos coeficientes cuenta el número de disparidades que encontramos entre los valores de la variable y, como se ha dicho, no tiene en cuenta la magnitud de dichos valores ni, por tanto, la distancia entre los mismos. Cada disparidad la cuenta dos veces, pues contamos la de x_i con x_j y la de x_j con x_i.

Hemos de destacar que estas dos medidas tienen carácter de coeficiente o indicador, por dos razones: no depende de las unidades de la variable y su recorrido es menor estricto que 1, en la primera, y menor o igual que 1 en la segunda. Téngase presente que en una muestra tamaño *n*, si todos los valores observados son distintos, el número total

Therefore, the numerator in both coefficients counts the number of dissimilarities that are found among the values of the variable and, as it has already been said, it does not take into account the magnitude of the values nor the distance among them. Every dissimilarity is counted twice, since it is counted the dissimilarity between x_i and x_j, and the one between x_j with x_i.

It must be emphasised that these two measures present the nature of coefficient or indicator for two reasons: they do not depend on the variable units and their range is less than 1, in the first measure, and less than or equal to 1, in the second one. It must also be considered that in a sample of size *n*, if all observed values are different, the total

de posibles parejas que se pueden formar, de x_i con x_j y de x_j con x_i, es n^2.

A ese número restamos las parejas del tipo (x_i, x_i), que son n en total, nos queda como número máximo de parejas con componentes distintos $n^2 - n = n(n-1)$. Podemos escribir:

$0 \le D_1 \le \frac{n-1}{n} < 1$ y $0 \le D_2 \le 1$. Cuando no hay disparidad, cuando todas las observaciones coinciden, ambos coeficientes toman el valor cero. Cuando se produce la máxima disparidad, cuando todas las observaciones son distintas, el primero toma el valor $\frac{n-1}{n}$ y el segundo el valor 1. En este aspecto, podríamos decir que se trata de medidas relativas de variabilidad.

Mostramos ejemplos ilustrativos sencillos:

Ejemplo 1: La variable X toma 5 valores siendo todos distintos, $X : \{1, 2, 3, 4, 5\}$.

La media aritmética es 3.

Calculamos en primer lugar las "medidas de dispersión" comentadas arriba.

number of possible pairs x_i with x_j and x_j with x_i) to be formed n^2. From that number we subtract the pairs of the form (x_i, x_i), which are n in total, so the highest number of pairs with different components is $n^2 - n = n(n-1)$. We can write: $0 \le D_1 \le \frac{n-1}{n} < 1$ and $0 \le D_2 \le 1$. When there is no disparity, when all the observations coincide, both coefficients take the value zero. When there is no dissimilarity, that is, when all observations are the same, both coefficients equal zero. When there is the highest dissimilarity, that is, when all observations are different, the first coefficient equals $\frac{n-1}{n}$ and the second one equals 1. In this regard, it could be said that they are relative measures of variability.

Next some simple illustrative examples are shown.

Example 1: Variable X presents 5 different values, $X : \{1, 2, 3, 4, 5\}$.

The arithmetic mean is 3.

We first calculate the aforementioned 'measures of dispersion'.

Tabla 1. Cálculo de las desviaciones al cuadrado respecto de la media
Table 1. Calculation of squared differences around the mean

x_i	$x_i - \bar{X}$	$(x_i - \bar{X})^2$
1	-2	4
2	-1	1
3	0	0
4	1	1
5	2	4
	Total	10

Entonces, $S^2 = \frac{10}{5} = 2$ y $S_c^2 = \frac{10}{4} = 2'5$.

Para la "media de las diferencias" construimos la siguiente tabla:

Then, $S^2 = \frac{10}{5} = 2$ and $S_c^2 = \frac{10}{4} = 2'5$.

For the 'differences mean' we construct the following table:

Tabla 2. Cálculo de las diferencias por parejas en valor absoluto
Table 2. Calculation of squared pairwise differences in absolute value

$\lvert x_i - x_j \rvert$						
x_i \ x_j	1	2	3	4	5	Suma de cada fila *Total in row*
1	0	1	2	3	4	10
2	1	0	1	2	3	7
3	2	1	0	1	2	6
4	3	2	1	0	1	7
5	4	3	2	1	0	10
					Total	40

$\Delta = \frac{40}{5 \cdot 4} = 2$. Para los promedios cuadráticos de diferencias por pares: / *For the squared means of pairwise differences*:

Tabla 3. Cálculo de las diferencias cuadráticas por parejas
Table 3. Calculation of squared pairwise differences

$(x_i - x_j)^2$						
x_i \ x_j	1	2	3	4	5	Suma de cada fila *Total in row*
1	0	1	4	9	16	30
2	1	0	1	4	9	15
3	4	1	0	1	4	10
4	9	4	1	0	1	15
5	16	9	4	1	0	30
					Total	100

$$V^2 = \frac{100}{5^2} = 4, \; V_c^2 = \frac{100}{5 \cdot 4} = 5.$$

Calculamos por último los "coeficientes de disparidad" / We finally calculate the 'coefficients of dissimilarity':

Tabla 4. Cálculo de las disparidades por parejas
Table 4. Calculation of pairwise dissimilarities

$c(x_i, x_j)$						
x_i \ x_j	1	2	3	4	5	Suma de cada fila *Total in row*
1	0	1	1	1	1	4
2	1	0	1	1	1	4
3	1	1	0	1	1	4
4	1	1	1	0	1	4
5	1	1	1	1	0	4
					Total	20

Por tanto / *Therefore* $D_1 = \frac{20}{25} = 0'8$ y / *and* $D_2 = \frac{20}{20} = 1$.

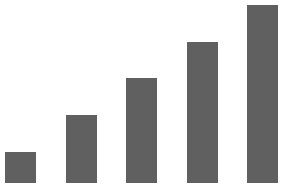
Estamos en un caso de máxima disparidad, todos los valores observados de la variable son distintos.

No hay cosa mejor para visualizar la variabilidad que observar los propios valores mediante alguna asociación geométrica, sobre todo cuando, como en este caso, tenemos pocos valores observados. Construimos, entonces, cinco barras cuyas longitudes son proporcionales a la magnitudes de los datos, y proponemos a los estudiantes su observación para que comparen con otras variables también representadas. Advertimos sobre la posible confusión de este gráfico con el diagrama de barras habitual en estadística. Aquí, y en los ejemplos que siguen, la longitud de cada barra no representa la frecuencia absoluta de un valor de la variable, sino que es el propio valor, y observamos en el conjunto cómo de diferentes son entre sí los valores observados.

This is the case of highest dissimilarity, since all observed values of the variable are different.

A better way of visualising variability is to observe the values through some geometric association, especially when there are few observed values, like in this case. Then, we draw five bars, whose height is proportional to the magnitudes of data and we suggest that our students observe them in order to compare with other variables which have also been presented. We warn them about mistaking this chart for the usual bar chart in statistics. In this example and for the following ones, every bar height does not represent the absolute frequency for a value of the variable, but the value itself, and we observe how different to each other the observed values are within the set.

Gráfico 1. Visualización de 5 valores observados

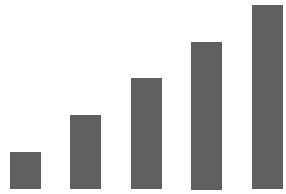


Ejemplo 2. La variable *X* toma también 5 valores distintos, $X:\{1,3,5,7,9\}$. La diferencia con la anterior está en la magnitud de los mismos. Procedemos con los mismos cálculos y representamos de manera similar al anterior. En la tabla resumen que ponemos más abajo (Tabla 5) aparecen los valores de los estadísticos de dispersión y de disparidad.

Ejemplo 3. La variable X toma estos cinco valores $X:\{1,1,3,5,5\}$. Aquí se da más paridad o, quizás mejor, menos disparidad. Resumiremos en la tabla. Igual haremos con el último de los cuatro ejemplos.

Ejemplo 4. La variable X toma estos cinco valores $X:\{1,1,5,9,9\}$.

Graph 1. Visualisation of 5 observed values

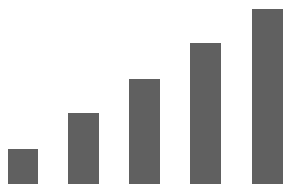


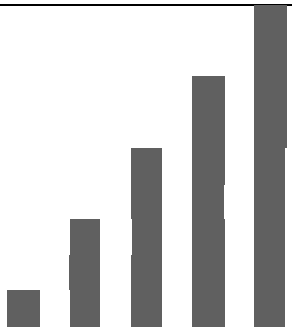
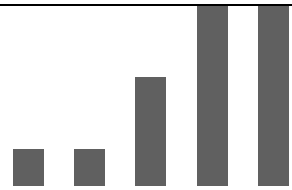
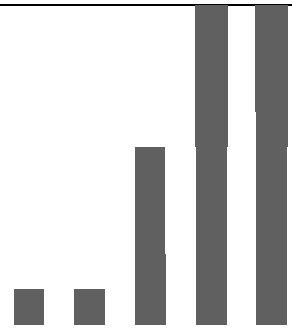
Example 2: Variable *X* also presents 5 different values, $X:\{1,3,5,7,9\}$. The difference with the previous one is their magnitude. We proceed with the same computations and present data in a similar way. In the summary table below (Table 5) all measures of dispersion and dissimilarity are presented.

Example 3. Variable X presents these 5 values, $X:\{1,1,3,5,5\}$. In this case there is more similarity, or more correct, less dissimilarity. This will also be summarised in the table, as it will be for the last example.

Example 4. Variable X presents these 5 values, $X:\{1,1,5,9,9\}$.

Tabla 5. Cuadro resumen de las medidas de variabilidad para los cuatro ejemplos / *Table 5. Summary Table of measures of variability for the four examples*

Ejemplo / Example	S^2	S_c^2	Δ	V^2	V_c^2	D_1	D_2
	2	2'5	2	4	5	0'8	1

	8	10	4	16	20	0'8	1
	3'2	4	2'4	6'4	8	0'64	0'8
	12'8	16	4'8	25'6	32	0'64	0'8

Comparamos entre sí los ejemplos:

- *Ejemplo* 1 y *Ejemplo* 3: mayor dispersión en el 3 y mayor disparidad en el 1. Podemos decir que en el ejemplo 3 hay mayor dispersión que en el 1 y, sin embargo, menor disparidad.
- *Ejemplo* 1 y *Ejemplo* 2: mayor dispersión en el 2 que en el 1 (las diferencias en cuanto a sus magnitudes son mayores en los valores observados en el 2 que en el 1) y la misma disparidad.
- *Ejemplo* 3 y *Ejemplo* 4: mayor dispersión en el 4 que en el 3 (las diferencias en cuanto a sus magnitudes son mayores en los valores observados en el 4 que en el 3) y la misma disparidad.

A comparison between examples is carried out:

- *Example* 1 and *Example* 3: higher dispersion in 3 and higher dissimilarity in 1. It can be said that in example 3 there is more dispersion than in 1 but less dissimilarity.
- *Example* 1 and *Example* 2: higher dispersion in 2 than in 1 (the differences in the magnitude are higher in the observed values in 2 than in 1) and same dissimilarity.
- *Example* 3 and *Example* 4: higher dispersion in 4 than in 3 (the differences in the magnitude are higher in the observed values in 4 than in 3) and same dissimilarity.

- De los cuatro ejemplos, el de mayor dispersión es el 4 y, sin embargo, es uno de los de menor disparidad.
- De los cuatro ejemplos, el de menor dispersión es el 1 y, sin embargo, es uno de los de mayor disparidad.

Por tanto, hemos de distinguir entre lo que es el "cuánto" de lo que es "con qué frecuencia", o sea, la distinción entre medidas basadas en la distancia (dispersión) de las más simples basadas en la disyuntiva entre igualdad o no igualdad (disparidad). Es interesante intentar captar la percepción que nuestros estudiantes tienen de la variabilidad mediante el ejercicio sencillo de mostrar representaciones similares a las anteriores para que se manifiesten sobre cuál presenta mayor o menor variabilidad.

3. MIDIENDO LA VARIABILIDAD EN CATEGÓRICAS: COEFICIENTES DE DISPARIDAD

De las dos formas de medir la variabilidad comentadas en el apartado anterior, la primera basada en las distancias no es aplicable en variables categóricas. Supongamos el caso más sencillo, una variable de carácter dicotómico donde las dos posibles respuestas son representadas por A y B. Esas respuestas no están definidas por magnitudes numéricas (salvo que codifiquemos arbitrariamente) por lo que no podemos medir la distancia entre A y B, o sea, no podemos construir una "medida de dispersión" para esta variable. Lo que sí podemos hacer es comparar las respuestas de los individuos y ver si las mismas coinciden o no. Por tanto, los dos coeficientes de disparidad introducidos para cuantitativas serían perfectamente válidos en las cualitativas y esas son las medidas de variabilidad que proponemos para las mismas.

- Out of the four examples, the highest dispersion is in 4 and, however, it is one of the examples where there is less dissimilarity.
- Out of the four examples, the smallest dispersion is in 1 and, however, it is one of the examples where there is more dissimilarity.

Therefore, we have to distinguish between what is "how much" of what is "with what frequency", or, the distinction between measures based on the distance (dispersion) of the simplest stocks in the dilemma between equality or not equality (disparity). It is interesting to try to catch the perception that our students have of the variability by means of the exercise simple to show representations similar to the previous ones in order that they demonstrate on which he presents major or minor variability.

3. MEASURING VARIABILITY IN CATEGORICAL CHARACTERISTICS: COEFFICIENTS OF DISSIMILARITY

Out of the two ways of measuring variability which were presented in the previous section, the first one based on distances is not applicable to categorical variables. Let us figure out the simplest case, a dichotomous variable where the two possible responses are represented by A and B. These responses are not defined as numerical magnitudes (unless they are arbitrarily codified), so we cannot measure the distance between A and B, that is, we cannot construct a 'measure of dispersion' for this variable. What we can do is to compare the individuals' responses and observe whether they are the same or not. Therefore, both coefficients of dissimilarity which were presented for quantitative variables could also be valid in the case of qualitative ones, and they are the measure of variability we propose for them.

Planteamos tres ejemplos de variables dicotómicas en los que, para los tres casos, requerimos las respuestas de 6 individuos. Visualizamos las respuestas y calculamos los dos coeficientes de disparidad en cada uno de los tres casos:

Ejemplo 1: $X:\{A,B,B,B,B,B\}$.

Ejemplo 2: $X:\{A,A,A,B,B,B\}$.

Ejemplo 3: $X:\{A,A,B,B,B,B\}$.

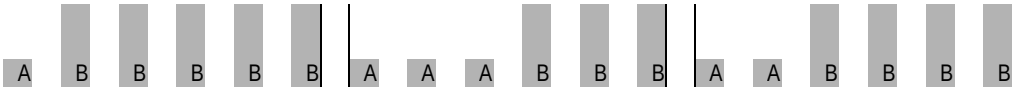
We set out three examples of dichotomous variables where the response of 6 individuals is needed. We show the responses and calculate both coefficients of dissimilarities for each of the three cases:

Example 1: $X:\{A,B,B,B,B,B\}$.

Example 2: $X:\{A,A,A,B,B,B\}$.

Example 3: $X:\{A,A,B,B,B,B\}$.

Gráfico 2. Visualización de tres variables categóricas con dos posibles respuesta cada una / *Graph 2. Visualisation of three categorical variables with two possible responses*



Ejemplo 1/*Example 1* Ejemplo 2/*Example 2* Ejemplo 3/*Example 3*

Calculamos los coeficientes para los tres ejemplos consecutivamente, usando tablas de disparidades similares a los ejemplos de cuantitativas, donde hemos sombreado las "cajas" donde aparecen 1 (disparidades).

We calculate the coefficients for the three examples consecutively, using tables of dissimilarities which are similar to the ones used in the examples of quantitative variables. The cells in the tables where the value is 1 (there is dissimilarity) have been shaded.

Tabla 6. Disparidades del *Ejemplo* 1 / *Table 6. Dissimilarities in Example 1*

$c(x_i, x_j)$							
	A	B	B	B	B	B	Suma de cada fila *Total in row*
A	0	1	1	1	1	1	5
B	1	0	0	0	0	0	1
B	1	0	0	0	0	0	1
B	1	0	0	0	0	0	1
B	1	0	0	0	0	0	1
B	1	0	0	0	0	0	1
						Total	10

$$D_1 = \frac{10}{6^2} = 0'277\,, \qquad D_2 = \frac{10}{6\cdot 5} = 0'333\,.$$

Tabla 7. Disparidades del *Ejemplo* 2 / *Table 7. Dissimilarities in Example 2*

$c(x_i, x_j)$							
	A	A	A	B	B	B	Suma de cada fila *Total in row*
A	0	0	0	1	1	1	3
A	0	0	0	1	1	1	3
A	0	0	0	1	1	1	3
B	1	1	1	0	0	0	3
B	1	1	1	0	0	0	3
B	1	1	1	0	0	0	3
						Total	18

$$D_1 = \frac{18}{6^2} = 0'5\,, \qquad D_2 = \frac{18}{6 \cdot 5} = 0'6\,.$$

Tabla 8. Disparidades del *Ejemplo* 3 / *Table 8. Dissimilarities in Example 3*

$c(x_i, x_j)$							
	A	A	B	B	B	B	Suma de cada fila *Total in row*
A	0	0	1	1	1	1	4
A	0	0	1	1	1	1	4
B	1	1	0	0	0	0	2
B	1	1	0	0	0	0	2
B	1	1	0	0	0	0	2
B	1	1	0	0	0	0	2
						Total	16

$$D_1 = \frac{16}{6^2} = 0'444\,, \qquad D_2 = \frac{16}{6 \cdot 5} = 0'533\,.$$

Según estos coeficientes, la variable categórica donde hay menor variabilidad (en el sentido de disparidad), es la del *Ejemplo* 1, y la de mayor, la del *Ejemplo* 2.

En el *Ejemplo* 1, para el primer coeficiente podemos escribir, observando las dimensiones de las cajas donde aparecen 1:

According to these coefficients, the categorical variable where there is the smallest variability (in the sense of dissimilarity) is the one in *Example* 1, whereas the highest one is in *Example* 2.

In *Example* 1, observing the cells where a value 1 is present, the first coefficient can be written as:

$$D_1 = \frac{1 \cdot 5 + 5 \cdot 1}{6^2} = \frac{2 \cdot 1 \cdot 5}{6^2} = 2 \cdot \frac{1}{6} \cdot \frac{5}{6}$$

Obsérvese que la primera fracción, $\frac{1}{6}$, es la proporción de respuestas A que encontramos en esa variable categórica, mientras que la segunda, $\frac{5}{6}$, es la de respuestas B. Por tanto, en el caso de una variable categórica con dos posibles respuestas, si p_1 es la proporción de respuestas correspondientes a la primera categoría, o sea, $p_1 = \frac{n_1}{n}$, con n_1 número de veces que aparece la primera respuesta, y si p_2 es la proporción para la segunda respuesta, $p_2 = \frac{n_2}{n}$, podemos escribir el primer coeficiente de disparidad como:

$$\boxed{D_1 = 2 \cdot p_1 \cdot p_2},$$

o sea, 2 veces la varianza de una variable aleatoria Bernoulli (la misma relación que la existente entre varianza y cuasivarianza, por una parte, y los dos promedios cuadráticos de diferencias por pares, por la otra, en el caso cuantitativo). Podíamos evitar ese 2 si contásemos las disparidades de una pareja una sola vez. Como ya se ha comentado, en los coeficientes propuestos contamos la disparidad de x_i con x_j y la de x_j con x_i. A nivel práctico bastaría con dividir por 2 esos coeficientes. Ahora bien, al hacerlo cambiaríamos el recorrido de ambos. Por ejemplo, D_2, en lugar de tomar valores entre 0 y 1, los tomaría entre 0 y 0'5, como ocurre con los posibles valores de la varianza de una distribución Bernoulli.

Alguna manipulación más es posible:

The first fraction, $\frac{1}{6}$, is the proportion of A responses that are found in that categorical variable, whereas the second one, $\frac{5}{6}$, is the proportion of B responses. Therefore, in the case of a categorical variable with two possible responses, if p_1 is the proportion of responses corresponding to the first category, that is, $p_1 = \frac{n_1}{n}$, with n_1 being the number of times the first responses is found, and if p_2 is the proportion for the second response, that is, $p_2 = \frac{n_2}{n}$, then, the first coefficient of dissimilarity can be written as:

$$\boxed{D_1 = 2 \cdot p_1 \cdot p_2},$$

That is, twice the variance of a Bernoulli random variable (the same relationship as the one existing between variance and quasivariance, on the one hand, and both squared pairwise differences means, on the other, in the case of quantitative variables). We could eliminate '2' in the above expression by counting every pairwise dissimilarity only once. As it was already said, in the proposed coefficients, the dissimilarity of x_i with x_j and the one of x_j with x_i is counted. At a practical level, it would be enough to divide both coefficients by 2. Nevertheless, by doing that their range would be modified. For instance, D_2, instead of being between 0 and 1, would vary from 0 to 0,5, as it happens with the possible values of the variance in a Bernoulli distribution.

Some more manipulation is possible:

La suma de "unos" que aparece en cada tabla se podría construir así (mirar zonas sombreadas):

The addition of 'ones' in each table could be calculated as (observe the shaded areas):

$$n_1 \cdot n_2 + n_2 \cdot n_1 = n_1 (n - n_1) + n_2 (n - n_2)$$

Por tanto, / *Therefore*,

$$D_1 = \frac{n_1 (n - n_1) + n_2 (n - n_2)}{n^2} = \frac{n_1}{n} \cdot \frac{n - n_1}{n} + \frac{n_2}{n} \cdot \frac{n - n_2}{n}.$$

O sea, ese coeficiente se puede escribir también como:

That is, the coefficient can also be written as:

$$\boxed{D_1 = p_1 (1 - p_1) + p_2 (1 - p_2)}.$$

$$\boxed{D_1 = p_1 (1 - p_1) + p_2 (1 - p_2)}.$$

Aún otra expresión más. El número de "unos" que hay en la caja también se puede calcular restando al total de celdas de la tabla el número de "ceros". Así, en el *Ejemplo* 3 sería $16 = 6^2 - 2^2 - 4^2$. Por tanto,

$D_1 = \frac{6^2 - 2^2 - 4^2}{6^2} = 1 - \left(\frac{2}{6}\right)^2 - \left(\frac{4}{6}\right)^2$. En general, otra expresión más para el cálculo de este coeficiente es:

There is also another possible expression. The number of 'ones' in the table can also be calculated by subtracting the number of 'zeros' from the total number of cells in the table. Thus, in *Example* 3, it would be $16 = 6^2 - 2^2 - 4^2$. Therefore, $D_1 = \frac{6^2 - 2^2 - 4^2}{6^2} = 1 - \left(\frac{2}{6}\right)^2 - \left(\frac{4}{6}\right)^2$. In general, another expression for the computation of this coefficient is:

$$\boxed{D_1 = 1 - p_1^2 - p_2^2}.$$

$$\boxed{D_1 = 1 - p_1^2 - p_2^2}.$$

A continuación planteamos otro ejemplo de variable categórica en el que hay tres posibles respuestas, A, B y C, de una cuestión planteada a 8 individuos, dando como resultado la siguiente estadística ya agrupada por respuestas:

$X : \{A, B, B, C, C, C, C, C\}$. Visualizamos estas respuestas, pero en este caso evitamos la utilización de la longitud como elemento distintivo de las respuestas con el objeto de que las mismas pueden ejercer impacto visual ajeno a lo buscado.

We next set out another example of categorical variable where there are three possible responses, A, B and C, to a question posed to 8 individuals, resulting in the following statistics, which have already been grouped according to the responses:

$X : \{A, B, B, C, C, C, C, C\}$. We show these responses but in this case we avoid the use of the height as a distinguishing element among the responses, since they can have a visual impact different to the proper one.

Gráfico 3. Visualización de una variable categórica con tres posibles respuestas / *Graph 3. Visualisation of a categorical variable with three possible responses*

Calculamos para esta variable los dos coeficientes de disparidad. En primer lugar, la tabla de disparidades: / We calculate both coefficients of dissimilarity for this variable, We first present the table of dissimilarities:

Tabla 9. Disparidades de variable categórica con tres posibles respuestas ***Table 9. Dissimilarities of a categorical variable with three possible responses***

$c(x_i, x_j)$

	A	B	B	C	C	C	C	C	Suma de cada fila *Total in row*
A	0	1	1	1	1	1	1	1	7
B	1	0	0	1	1	1	1	1	6
B	1	0	0	1	1	1	1	1	6
C	1	1	1	0	0	0	0	0	3
C	1	1	1	0	0	0	0	0	3
C	1	1	1	0	0	0	0	0	3
C	1	1	1	0	0	0	0	0	3
C	1	1	1	0	0	0	0	0	3
								Total	34

Entonces,

$D_1 = \frac{34}{8^2} = 0'531$ y $D_2 = \frac{34}{8 \cdot 7} = 0'607$. Si observamos la tabla, el número de "unos" que hay en la misma es la suma del número de celdas contenidos en la tres cajas enmarcadas y sombreadas (la de A con B, la de A con C, y la de B con C) que, a su vez, están duplicadas. O sea, $34 = 2\left(1 \cdot 2 + 1 \cdot 5 + 2 \cdot 5\right)$. Por tanto,

$$D_1 = \frac{2(1 \cdot 2 + 1 \cdot 5 + 2 \cdot 5)}{8^2} = 2\left(\frac{1}{8} \cdot \frac{2}{8} + \frac{1}{8} \cdot \frac{5}{8} + \frac{2}{8} \cdot \frac{5}{8}\right).$$

Entonces, si p_1, p_2 y p_3 son las proporciones de individuos que escogen cada una de las tres respuestas, tenemos:

$$\boxed{D_1 = 2\left(p_1 p_2 + p_1 p_3 + p_2 p_3\right)}$$

Then,

$D_1 = \frac{34}{8^2} = 0'531$ and $D_2 = \frac{34}{8 \cdot 7} = 0'607$. If we observe the table, the number of 'ones' that exists in the same one is the sum of the number of cells contained in three boxes framed and shaded (her of A with B, her of A with C, and her of B with C) that, in turn, are duplicated. Or, $34 = 2\left(1 \cdot 2 + 1 \cdot 5 + 2 \cdot 5\right)$. Therefore,

$$D_1 = \frac{2(1 \cdot 2 + 1 \cdot 5 + 2 \cdot 5)}{8^2} = 2\left(\frac{1}{8} \cdot \frac{2}{8} + \frac{1}{8} \cdot \frac{5}{8} + \frac{2}{8} \cdot \frac{5}{8}\right).$$

If p_1, p_2 and p_3 are the proportions of individuals that who choose each of three responses, respectively, then:

$$\boxed{D_1 = 2\left(p_1 p_2 + p_1 p_3 + p_2 p_3\right)}$$

También, el número de "unos" de la caja anterior puede ser determinado mediante $34 = 1\cdot(8-1)+2\cdot(8-2)+5\cdot(8-5)$, por lo que el coeficiente sería,

The number of 'ones' in the table above can also be calculated as $34 = 1\cdot(8-1)+2\cdot(8-2)+5\cdot(8-5)$, so the coefficient would be:

$$D_1 = \frac{1\cdot(8-1)+2\cdot(8-2)+5\cdot(8-5)}{8^2} = \frac{1}{8}\cdot\frac{8-1}{8}+\frac{2}{8}\cdot\frac{8-2}{8}+\frac{5}{8}\cdot\frac{8-5}{8} = \frac{1}{8}\left(1-\frac{1}{8}\right)$$

$$+\frac{2}{8}\left(1-\frac{2}{8}\right)+\frac{5}{8}\left(1-\frac{5}{8}\right).$$ En general, / *In general,*

$$\boxed{D_1 = p_1(1-p_1)+p_2(1-p_2)+p_3(1-p_3)}.$$

Por último, la suma de "unos" puede calcularse también restando al total de celdas de la caja, 8^2, el total de ceros que hay en ella. O sea, $34 = 8^2-1^2-2^2-5^2$. Por tanto,

Finally, the sum of 'ones' can also be calculated by subtracting from the total number of cells in the table, 8^2, the number of 'zeros' in it. That is, $34 = 8^2-1^2-2^2-5^2$. Therefore,

$$D_1 = \frac{8^2-1^2-2^2-5^2}{8^2} = 1-\left(\frac{1}{8}\right)^2-\left(\frac{2}{8}\right)^2-\left(\frac{5}{8}\right)^2.$$ En general,

$$D_1 = \frac{8^2-1^2-2^2-5^2}{8^2} = 1-\left(\frac{1}{8}\right)^2-\left(\frac{2}{8}\right)^2-\left(\frac{5}{8}\right)^2.$$ In general,

$$\boxed{D_1 = 1-p_1^2-p_2^2-p_3^2}.$$

$$\boxed{D_1 = 1-p_1^2-p_2^2-p_3^2}.$$

A partir de los ejemplos analizados para dos o tres posibles respuestas de una variable cualitativa nos resulta relativamente fácil establecer diferentes expresiones para el primer coeficiente de disparidad: Si una variable categórica tiene k posibles respuestas o categorías y si disponemos de un número finito de observaciones, n, y si $n_1, n_2, ..., n_i, ..., n_k$ representan la frecuencia con que aparece cada una de las categorías con, naturalmente, $n_1+n_2+...+n_i+...+n_k = n$, llamamos $p_i = \frac{n_1}{n}$, $i = 1,2,...,k$, o sea, la proporción de respuestas que corresponde a la categoría i entre las observaciones.

From the examples above for two or three possible responses of a qualitative variable it is relatively easy to define different expressions for the first coefficient of dissimilarity: If there is a categorical variable with k possible responses o categories, if there are a finite number of observations, n, and if $n_1, n_2, ..., n_i, ..., n_k$, represent the frequency of appearance of every category, with $n_1 + n_2 + ... + n_i + ... + n_k = n$, then $p_i = \frac{n_1}{n}$, $i = 1,2,...,k$, is the proportion of responses in the category i for the observations.

Entonces, podemos escribir para el primer coeficiente de disparidad las siguientes expresiones: / Then, the first coefficient of dissimilarity can be written as:

$$D_1 = 2\sum_{i<j} p_i p_j \ ,$$

$$D_1 = \sum_{i=1}^{k} p_i (1-p_i) \ ,$$

$$D_1 = 1-\sum_{i=1}^{k} p_i^2 \ .$$

4. CONCLUSIONES

El concepto de variabilidad es más amplio de lo que habitualmente se explica en los libros de texto y en clase. En variables cuantitativas, además de la idea de dispersión, en general ligada a la desviación respecto a la media, podemos introducir por ejemplo la de disparidad, que conduce a medidas sencillas e intuitivas. La distinción entre el "cuánto" y "con qué frecuencia" es la base de la separación entre dispersión y disparidad. Aunque el "cuánto se diferencian los datos" no se puede medir en variables categóricas, sí podemos contar "con qué frecuencia son distintas las respuestas". Por tanto, medidas relacionadas con la disparidad son posibles en variables cualitativas. Creemos que dichas medidas, a las que hemos llamado "coeficientes de disparidad", por su naturalidad y sencillez, deben ser abordadas en un curso de introducción a la estadística descriptiva llenando así uno de los vacíos tradicionales de la enseñanza de esta disciplina. La estadística existe al existir variabilidad dentro de un carácter medido en una población y dicho carácter puede ser cuantitativo o cualitativo. Es función del usuario de la estadística poder medir dicha variabilidad. La visualización de ejemplos simples por parte de los alumnos permitirá al profesor la captación de las ideas que sobre variabilidad tienen los mismos.

4. CONCLUSIONS

The concept of variability is wider than the one that is usually discussed in literature and classroom. In the case of quantitative variables, apart from the idea of dispersion, in general related to the deviation around the mean, it can be introduced, for example, the idea of dissimilarity, which results in simple and intuitive measures. The difference between 'when' and 'how frequently' is the base to distinguish between dispersion and dissimilarity. Even though 'how much data are different' cannot be measured in categorical variables, it is possible to count 'how frequently responses are different'. Therefore, measures related to dissimilarity are possible to be defined in qualitative variables. We think that these measures, which we have called 'coefficients of dissimilarity', due to their naturalness and simplicity, must be dealt with in a descriptive statistics introductory course, filling in this way one of the gags in the teaching of this subject. Statistics exists because variability inside a characteristic measured on a population exists, and that characteristic can be quantitative or qualitative. To measure that variability is a role corresponding to the user of statistics. The visualization of simple examples by students will allow the teacher to catch the idea they have on variability.

BIBLIOGRAFÍA/REFERENCES

Agresti, A. (1990). *Categorical data analysis*. John Wiley & Sons.

Blasius, J. y Greenacre, M. (1998). *Visualization of categorical data.* San Diego (CA): Academic Press.

Gini, C.W. (1912). Variability and mutability, contribution to the study of statistical distributions and relations. Estudi Economico-Giuricici della R. Universita de Cagliari.

Gordon, T. (1986). Is the standard deviation tied to the mean? *Teaching Statistics*, *8*(2), 67-70.

Kader, G.D. y Perry, M. (2007). Variability for categorical variables. *Journal of Statistics Education*, *15*(2), 1-17.

Loosen, F., Lioen, M. y Lacante, M. (1985). The standard deviation: Some drawbacks to an intuitive approach. *Teaching Statisctics*, *7*(1), 2-5.

Perry, M. y Kader, G. (2005). Variation as unalikeability. *Teaching Statistics*, *27*(2), 58-60.

Pecvnia Monográfico 2012, pp. 21-30

ORDEN DE LORENZ EN LA FAMILIA DE DISTRIBUCIONES GAMMA TRIPARAMÉTRICAS
LORENZ ORDERING OF THREE PARAMETER GAMMA DISTRIBUTIONS

Héctor M. Ramos[1]
hector.ramos@uca.es
Miguel A. Sordo[1]
mangel.sordo@uca.es

Universidad de Cádiz

Resumen

El orden de Lorenz es una herramienta adecuada para comparar la desigualdad de dos distribuciones de rentas. En este artículo obtenemos una condición suficiente para que dos distribuciones sean comparables en el orden de Lorenz y aplicamos el resultado para ordenar la familia de distribuciones Gamma triparamétricas.

Palabras clave: Orden de Lorenz; Distribución Gamma.

Abstract

Lorenz ordering is an useful tool for comparing inequality of income distributions. In this note, we give a sufficient condition for this order and then apply it to order the three parameter family of Gamma distributions.

Keywords: Lorenz ordering; Gamma distribution.

[1] Departamento de Estadística e I.O., Facultad de CC. Económicas y Empresariales, Universidad de Cádiz, Duque de Nájera 8, 11002 Cádiz (España).

1. INTRODUCCIÓN

El orden de Lorenz permite la comparación en desigualdad de dos distribuciones de renta. Sea $F_X(x)$ la función de distribución de una variable aleatoria X no negativa y con media finita μ_X. La curva de Lorenz, también llamada *curva de concentración,* correspondiente a X se define (Gastwirth, 1971) como:

$$L_X(p) = \frac{1}{\mu_X}\int_0^p F_X^{-1}(t)dt \quad 0 \le p \le 1 \quad (1)$$

Donde por F_X^{-1} denotamos la inversa de F_X definida por

$$F_X^{-1}(p) = \inf\{x : F_X(x) \ge p\}, \quad p \in [0,1].$$

Si X representa ingresos anuales, entonces $L_X(p)$ es la proporción del total de ingresos que corresponde a los individuos que se encuentran en el $100p\%$ de ingresos más bajos. Un detallado estudio de la curva de Lorenz podemos encontrarlo en Gail y Gastwirth (1978), así como una relación precisa de sus propiedades en Dagum (1985). La curva de Lorenz nos permite definir el siguiente orden parcial ($\le_L$) sobre la clase de variables aleatorias no negativas:

$$X \le_L Y \Leftrightarrow L_X(p) \ge L_Y(p) \text{ para todo } 0 \le p \le 1. \quad (2)$$

Si $X \le_L Y$, entonces diremos que X es menos desigual que Y en el sentido de Lorenz. Algunas referencias clásicas sobre el orden de Lorenz son Atkinson (1970), Dasgupta, Sen y Starrett (1973), Rothschild y Stiglitz (1973), Kakwani (1984) y Arnold (1987). De (1) y (2) se deduce fácilmente que el orden de Lorenz es invariante frente a transformaciones de escala; es decir, $X \le_L Y$ si y sólo si $aX \le_L bY$ para todo $a > 0,\ b > 0$. En

1. INTRODUCTION

Lorenz ordering is used to compare the amounts of inequality in two income distributions. Let $F_X(x)$ be the distribution function of a non-negative random variable X with finite mean μ_X, then the Lorenz curve, also called *curve of concentration,* corresponding to X can be defined (Gastwirth,1971) as:

$$L_X(p) = \frac{1}{\mu_X}\int_0^p F_X^{-1}(t)dt \quad 0 \le p \le 1 \quad (1)$$

where we denote by F_X^{-1} the inverse of F_X defined by

$$F_X^{-1}(p) = \inf\{x : F_X(x) \ge p\}, \quad p \in [0,1].$$

If X represents annual income, $L_X(p)$ is the proportion of total income that accrues to individuals having the $100p\%$ lowest incomes. There is extensive discussion of the Lorenz curve in Gail and Gastwirth (1978) and a concise account of its properties in Dagum (1985). The Lorenz curve can be used to define a partial ordering (denoted $\le_L$) on the class of non-negative random variables, as follows:

$$X \le_L Y \Leftrightarrow L_X(p) \ge L_Y(p) \text{ for every } 0 \le p \le 1. \quad (2)$$

If $X \le_L Y$, then X is said to exhibit less inequality in the Lorenz or relative Lorenz sense than Y. Some standard references for the Lorenz order are Atkinson (1970), Dasgupta, Sen and Starrett (1973), Rothschild and Stiglitz (1973), Kakwani (1984) and Arnold (1987). It is obvious from (1) and (2) that the Lorenz order is scale-invariant, that is, $X \le_L Y$ if and only if $aX \le_L bY$ for all $a > 0,\ b > 0$. Arnold [1] shows that

$$X + a \le_L X, \text{ for all } a > 0, \quad (3)$$

Arnold [1] puede verse que

$$X + a \leq_L X \text{, para todo } a > 0, \qquad (3)$$

para toda variable aleatoria X no negativa con media finita.

El orden de Lorenz nos permite establecer una comparación en desigualdad de forma absoluta, es decir, al margen de cualquier medida concreta de desigualdad que pudiera considerarse, las cuales, necesariamente, deben ser coherentes con dicho orden.

Desafortunadamente, no siempre dos variables aleatorias X e Y con distribuciones conocidas son susceptibles de ser comparadas en desigualdad en el sentido del orden de Lorenz. En este trabajo proponemos una condición suficiente para el orden de Lorenz cuyo cumplimiento es fácilmente contrastable. Esta condición suficiente será la que utilizaremos en la Sección 3 para la ordenación en desigualdad de la familia Gamma triparamétrica que puede ser usada como modelo probabilístico para la distribución de la renta.

Consideraremos los siguientes resultados previos (Shaked, 1982):

Definición 1 *Sea* $h(x)$ *una función real definida en* $I \subset \mathbf{R}$. *El número de cambios de signo de* h *en* I *se define como sigue:*

$$S(h) = \sup S[h(x_1), h(x_2), ..., h(x_m)] \quad (4)$$

donde $S[h(x_1), h(x_2), ..., h(x_m)]$ *es el número de cambios de signo una vez eliminados los términos iguales a cero, y el supremo en (4) se extiende a todos los conjuntos* $x_1 < x_2 < ... < x_m \quad (x_i \in I)$, $m < \infty$.

Necesitaremos el siguiente resultado.

for every non-negative random variable X with finite mean.

Lorenz ordering may be viewed as the maximal ranking generated by relative inequality measures, that is, if $X \leq_L Y$ then, whatever the measures of relative inequality one may choose, Y must not be judged as less unequal that X and conversely.

Unfortunately, for the two random variables X and Y with known distributions, it is sometimes not clear how to verify the relation $X \leq_L Y$. In this paper, we give a simple condition that ensure these orderings. This condition will be used in Section 3 for the ordering of the three parameter Gamma income distribution model. As in Shaked (1982), the following notation is used:

Definition 1 *Let* $h(x)$ *be a real function defined in* $I \subset \mathbf{R}$. *The number of sign changes of* h *in* I *is defined by*

$$S(h) = \sup S[h(x_1), h(x_2), ..., h(x_m)] \quad (4)$$

where $S[h(x_1), h(x_2), ..., h(x_m)]$ *is the number of sign changes of the indicated sequence, zero terms being discarded, and the supremum in (4) is extended over all sets* $x_1 < x_2 < ... < x_m \quad (x_i \in I)$, $m < \infty$.

We require the following well known result.

Teorema 2 *Sean* X *e* Y *variables aleatorias continuas con igual media* $\mu_X = \mu_Y$ *y sean* F *y* G *sus correspondientes funciones de densidad. Si* $S(F-G)=1$ *y la secuencia de signos es* $-,+,$ *entonces*

$$\int_0^u F^{-1}(t)dt \geq \int_0^u G^{-1}(t)dt, \quad \text{para todo } 0 \leq u \leq 1. \tag{5}$$

Demostración. De acuerdo con lo asumido en el enunciado, tendremos que

$$S\left(F^{-1}-G^{-1}\right)=1$$

siendo la secuencia de signos $+,-$.

Por lo tanto, la integral

$$\int_0^u \left[F^{-1}(t)-G^{-1}(t)\right]dt$$

alcanza su menor valor cuando $u=1.$

Por otra parte, de la igual de medias se sigue que

$$\int_0^u \left[F^{-1}(t)-G^{-1}(t)\right]dt \geq \int_0^1 \left[F^{-1}(t)-G^{-1}(t)\right]dt=0,$$

y, en consecuencia, se verifica (5).

2. CONDICIONES SUFICIENTES PARA EL ORDEN DE LORENZ

Utilizaremos el siguiente resultado de Arnold (1987).

Teorema 3 *Sean* X *and* Y *variables aleatorias no negativas con medias finitas* μ_X *y* μ_Y*, respectivamente, y sean* F *y* G *sus correspondientes funciones de distribución. Si* $S\left(F\left(x\mu_X\right)-G\left(x\mu_Y\right)\right)=1$ *y la secuencia de signos es* $-,+,$ *entonces* $X \leq_L Y.$

Theorem 2 *Let* X *and* Y *be continuous random variables with equal means* $\mu_X = \mu_Y$ *and let* F *and* G *the corresponding densities. If* $S(F-G)=1$ *and the sign sequence is* $-,+,$ *then*

$$\int_0^u F^{-1}(t)dt \geq \int_0^u G^{-1}(t)dt, \quad \textit{for all } 0 \leq u \leq 1. \tag{5}$$

Proof. By the assumptions on F and G we have that

$$S\left(F^{-1}-G^{-1}\right)=1$$

with the sequence $+,-$.

Therefore, the integral

$$\int_0^u \left[F^{-1}(t)-G^{-1}(t)\right]dt$$

assumes its smallest value for $u=1.$ From the equality of the means it follows that

$$\int_0^u \left[F^{-1}(t)-G^{-1}(t)\right]dt \geq \int_0^1 \left[F^{-1}(t)-G^{-1}(t)\right]dt=0,$$

and consequently (5) holds.

2. SUFFICIENT CONDITIONS FOR LORENZ ORDERING

The next theorem, due to Arnold (1987), will be used below.

Theorem 3 *Let* X *and* Y *be non-negative random variables with finite means* μ_X *and* μ_Y*, respectively, and let* F *and* G *be the corresponding distribution functions. If* $S\left(F\left(x\mu_X\right)-G\left(x\mu_Y\right)\right)=1$ *and the sign sequence is* $-,+,$ *then* $X \leq_L Y.$

Demostración. Sean $F_{\frac{X}{\mu_X}}$ y $G_{\frac{Y}{\mu_Y}}$ las funciones de distribución de $\frac{X}{\mu_X}$ y $\frac{Y}{\mu_Y}$, respectivamente. Puesto que

$$S\left(F_{\frac{X}{\mu_X}}(x)-G_{\frac{Y}{\mu_Y}}(x)\right)=S(F(x\mu_X)-G(x\mu_Y))=1$$

y la secuencia de signos es $-,+,$ del Teorema 2 se sigue que $\frac{X}{\mu_X}\leq_L\frac{Y}{\mu_Y}$. Y ya que el orden de Lorenz es invariante frente a transformaciones de escala, tendremos que $X\leq_L Y$.

Basándonos en el Teorema 3, obtenemos en el siguiente corolario una condición suficiente para la comparación en desigualdad en el sentido de Lorenz de dos variables aleatorias absolutamente continuas.

Corolario 4 *Sean X e Y variables aleatorias absolutamente continuas y no negativas, con medias finitas μ_X y μ_Y, y soportes $supp(X)$ y $supp(Y)$, respectivamente, y sean f y g sus correspondientes funciones de densidad. Asumiremos que $supp\ (\frac{X}{\mu_X})\subseteq supp\ (\frac{Y}{\mu_Y})$. En estas condiciones, si $f(\mu_X x)/g(\mu_Y x)$ es unimodal para valores de x restringidos al $supp\left(\frac{Y}{\mu_Y}\right)$, donde la moda es un supremo, entonces $X\leq_L Y$.*

Demostración. Puesto que

$$f_{\frac{X}{\mu_X}}(x)=\mu_X f(\mu_X x),\quad g_{\frac{Y}{\mu_Y}}(x)=\mu_Y g(\mu_Y x)$$

(6)

Proof. Let $F_{\frac{X}{\mu_X}}$ and $G_{\frac{Y}{\mu_Y}}$ be the distribution functions of $\frac{X}{\mu_X}$ and $\frac{Y}{\mu_Y}$, respectively. Since

$$S\left(F_{\frac{X}{\mu_X}}(x)-G_{\frac{Y}{\mu_Y}}(x)\right)=S(F(x\mu_X)-G(x\mu_Y))=1$$

and the sign sequence is $-,+,$ it follows from Theorem 2 that $\frac{X}{\mu_X}\leq_L\frac{Y}{\mu_Y}$. Since the Lorenz order is invariant under scale transformations we have that $X\leq_L Y$.

Based on Theorem 3 the next corollary gives a sufficient condition for the Lorenz comparison of two absolutely continuous random variables.

Corollary 4 *Let X and Y be non-negative and absolutely continuous random variables with finite means and supports $supp(X)$ and $supp(Y)$, respectively, and let f and g be the corresponding densities. Assume that $supp\ (\frac{X}{\mu_X})\subseteq supp\ (\frac{Y}{\mu_Y})$. If $f(\mu_X x)/g(\mu_Y x)$ is unimodal for x restricted to $supp\left(\frac{Y}{\mu_Y}\right)$, where the mode is a supremum, then $X\leq_L Y$.*

Proof. Since

$$f_{\frac{X}{\mu_X}}(x)=\mu_X f(\mu_X x),\quad g_{\frac{Y}{\mu_Y}}(x)=\mu_Y g(\mu_Y x)$$

(6)

y $f(\mu_X x)/g(\mu_Y x)$ es unimodal en $supp(\frac{Y}{\mu_Y})$, entonces también lo será $f_{\frac{X}{\mu_X}}(x)/g_{\frac{Y}{\mu_Y}}(x)$, siendo la moda un supremo. Por lo tanto,

$$S\left(f_{\frac{X}{\mu_X}}-g_{\frac{Y}{\mu_Y}}\right)=S\left(\frac{f_{\frac{X}{\mu_X}}}{g_{\frac{Y}{\mu_Y}}}-1\right)\le 2.$$

Como el orden estocástico no puede darse ya que $\frac{X}{\mu_X}$ y $\frac{Y}{\mu_Y}$ tienen la misma media, tendremos que $S\left(f_{\frac{X}{\mu_X}}-g_{\frac{Y}{\mu_Y}}\right)=2$ siendo la secuencia de signos $-,+,-$ y, en consecuencia, $S\left(F_{\frac{X}{\mu_X}}-G_{\frac{Y}{\mu_Y}}\right)=1$ siendo la secuencia de signos $-,+$. Finalmente, del Teorema 3 se sigue que $X \le_L Y$.

Observación 5 *Una condición suficiente para que f/g sea unimodal es que f/g sea log-concava* (Keilson y Gerber, 1971*).*

Corolario 6 *Sean* X *e* Y *variables aleatorias absolutamente continuas con medias finitas y soportes respectivos* $supp(X)=(a,\infty)$ *y* $supp(Y)=(b,\infty)$, $a>0$, $b\ge 0$, *y sean* f *y* g *sus correspondientes funciones de densidad. Si* $f(\mu_X x)/g(\mu_Y x)$ *decrece en* x *para* $x\ge\frac{a}{\mu_X}$, *entonces* $X \le_L Y$.

and $f(\mu_X x)/g(\mu_Y x)$ is unimodal on $supp(\frac{Y}{\mu_Y})$, so is $f_{\frac{X}{\mu_X}}(x)/g_{\frac{Y}{\mu_Y}}(x)$, with the mode yielding a supremum. Hence

$$S\left(f_{\frac{X}{\mu_X}}-g_{\frac{Y}{\mu_Y}}\right)=S\left(\frac{f_{\frac{X}{\mu_X}}}{g_{\frac{Y}{\mu_Y}}}-1\right)\le 2.$$

Since $\frac{X}{\mu_X}$ and $\frac{Y}{\mu_Y}$ have the same mean, ordinary stochastic order is not possible, so that $S\left(f_{\frac{X}{\mu_X}}-g_{\frac{Y}{\mu_Y}}\right)=2$ and the sign sequence is $-,+,-$. Then $S\left(F_{\frac{X}{\mu_X}}-G_{\frac{Y}{\mu_Y}}\right)=1$ and the sign sequence is $-,+$. From theorem 3 it follows that $X \le_L Y$.

Remark 5 *A sufficient condition for f/g to be unimodal is for f/g to be log-concave* (Keilson and Gerber, 1971).

Corollary 6 *Let* X *and* Y *be absolutely continuous random variables with finite means and supports* $supp(X)=(a,\infty)$ *and* $supp(Y)=(b,\infty)$, $a>0$, $b\ge 0$, *and let* f *and* g *be the corresponding densities. If* $f(\mu_X x)/g(\mu_Y x)$ *decreases in* x *for* $x\ge\frac{a}{\mu_X}$, *then* $X \le_L Y$.

Demostración. En primer lugar probaremos que $\frac{a}{\mu_X} > \frac{b}{\mu_Y}$. Consideremos, por reducción al absurdo, que $supp(\frac{Y}{\mu_Y}) \subseteq supp(\frac{X}{\mu_X})$, es decir que $\frac{b}{\mu_Y} \geq \frac{a}{\mu_X}$. Entonces, teniendo en cuenta (6), definiremos

$$\frac{g_{\frac{Y}{\mu_Y}}(x)}{f_{\frac{X}{\mu_X}}(x)} = \begin{cases} \frac{\mu_Y}{\mu_X}\frac{g(\mu_Y x)}{f(\mu_X x)} & \text{if} \quad x > \frac{b}{\mu_Y} \\ 0 & \text{if} \quad \frac{a}{\mu_X} < x \leq \frac{b}{\mu_Y} \end{cases}$$

Puesto que $f(\mu_X x)/g(\mu_Y x)$ es decreciente para $x \geq \frac{a}{\mu_X}$, tendremos que $g(\mu_Y x)/f(\mu_X x)$ es creciente para $x > \frac{b}{\mu_Y}$. En consecuencia,

$$S\left(g_{\frac{Y}{\mu_Y}} - f_{\frac{X}{\mu_X}}\right) = S\left(\frac{g_{\frac{Y}{\mu_Y}}}{f_{\frac{X}{\mu_X}}} - 1\right) \leq 1$$

lo que no es posible ya que $\frac{Y}{\mu_Y}$ y $\frac{X}{\mu_X}$ tienen la misma media. Por tanto, se cumple que $supp(\frac{X}{\mu_X}) \subset supp\ (\frac{Y}{\mu_Y})$, es decir, $\frac{a}{\mu_X} > \frac{b}{\mu_Y}$. Finalmente, puesto que $f(\mu_X x)/g(\mu_Y x)$ es decreciente para $x \geq a/\mu_X$, tenemos que $f(\mu_X x)/g(\mu_Y x)$ es unimodal para x restringido a $supp\left(\frac{Y}{\mu_Y}\right)$. Finalmente, aplicando el Corolario 4, el resultado queda demostrado.

Proof. First, we prove that $\frac{a}{\mu_X} > \frac{b}{\mu_Y}$. Assume, by the way of contradiction, that $supp(\frac{Y}{\mu_Y}) \subseteq supp(\frac{X}{\mu_X})$, i.e. that $\frac{b}{\mu_Y} \geq \frac{a}{\mu_X}$. Then, using (6) we define

$$\frac{g_{\frac{Y}{\mu_Y}}(x)}{f_{\frac{X}{\mu_X}}(x)} = \begin{cases} \frac{\mu_Y}{\mu_X}\frac{g(\mu_Y x)}{f(\mu_X x)} & \text{if} \quad x > \frac{b}{\mu_Y} \\ 0 & \text{if} \quad \frac{a}{\mu_X} < x \leq \frac{b}{\mu_Y} \end{cases}$$

Since $f(\mu_X x)/g(\mu_Y x)$ decreases in x for $x \geq \frac{a}{\mu_X}$, it follows that $g(\mu_Y x)/f(\mu_X x)$ increases in x for $x > \frac{b}{\mu_Y}$. Hence

$$S\left(g_{\frac{Y}{\mu_Y}} - f_{\frac{X}{\mu_X}}\right) = S\left(\frac{g_{\frac{Y}{\mu_Y}}}{f_{\frac{X}{\mu_X}}} - 1\right) \leq 1$$

which is not possible because $\frac{Y}{\mu_Y}$ and $\frac{X}{\mu_X}$ have the same mean. Thus, it follows that $supp(\frac{X}{\mu_X}) \subset supp\ (\frac{Y}{\mu_Y})$, i.e. $\frac{a}{\mu_X} > \frac{b}{\mu_Y}$. Again, since $f(\mu_X x)/g(\mu_Y x)$ decreases in x for $x \geq a/\mu_X$, we have that $f(\mu_X x)/g(\mu_Y x)$ is unimodal for x restricted to $supp\left(\frac{Y}{\mu_Y}\right)$. The result follows by applying Corollary 4.

3. APLICACIÓN

En esta sección aplicaremos nuestro anterior resultado (Corolario 6) para la ordenación de la familia de distribuciones Gamma triparamétricas.

3.1. Orden de Lorenz en la familia de distribuciones Gamma triparamétricas

Sea X la distribución Gamma triparamétrica cuya función de densidad es

$$f_X(x) = \frac{\beta^{-\alpha}(x-\theta)^{\alpha-1} e^{-\left(\frac{x-\theta}{\beta}\right)}}{\Gamma(\alpha)}, \theta \geq 0, \alpha > 0, \beta > 0, \ x > \theta. \tag{7}$$

Puesto que el orden de Lorenz es invariante frente a transformaciones de escala, podemos considerar, sin pérdida alguna de generalidad, que el parámetro de escala β es igual a 1 por lo que

$$f_X(x) = \frac{(x-\theta)^{\alpha-1} e^{-(x-\theta)}}{\Gamma(\alpha)}, \alpha > 0, \theta \geq 0, x > \theta. \tag{8}$$

Llamaremos $G(\theta,\alpha)$. a la correspondiente función de distribución. Arnold, Brockett, Robertson y Shu (1987) consideran $X_1 \sim G(\theta,\alpha_1)$ y $X_2 \sim G(\theta,\alpha_2)$ (θ fijo) y demuestran que $X_2 \leq_L X_1$ para $\alpha_1 \leq \alpha_2$. Por otra parte, si consideramos $X_1 \sim G(\theta_1,\alpha)$ y $X_2 \sim G(\theta_2,\alpha)$ α (fijo) puede verse que $X_2 \leq_L X_1$ para $\theta_1 \leq \theta_2$. En efecto, puesto que $X_2 = X_1 + (\theta_2 - \theta_1)$, con $\theta_2 - \theta_1 > 0$, se sigue de (3) que $X_2 \leq_L X_1$. Estos resultados pueden ser extendidos para $X_1 \sim G(\theta_1,\alpha_1)$ y $X_2 \sim G(\theta_2,\alpha_2)$. Aplicando la transitividad del orden de Lorenz y los anteriores resultados, es fácil probar que si $\alpha_1 \leq \alpha_2$ y $\theta_1 \leq \theta_2$, entonces $X_2 \leq_L X_1$. Sin embargo, no sabemos qué ocurre cuando $\alpha_1 < \alpha_2$ y $\theta_1 > \theta_2$. Esta cuestión nos conduce al siguiente resultado.

3. APPLICATION

3.1. Lorenz ordering of three parameter Gamma distributions

Let X be the three-parameter Gamma distribution with density function

Since the Lorenz order is invariant under scale changes, the scale parameter β can be set equal to 1 without loss of generality. This gives

The corresponding distribution will be denoted by $G(\theta,\alpha)$. Arnold, Brockett, Robertson and Shu (1987) consider $X_1 \sim G(\theta,\alpha_1)$ and $X_2 \sim G(\theta,\alpha_2)$ (θ fixed) and prove that $X_2 \leq_L X_1$ for $\alpha_1 \leq \alpha_2$. On the other hand, if we consider $X_1 \sim G(\theta_1,\alpha)$, and $X_2 \sim G(\theta_2,\alpha)$ (α fixed) it can be shown that $X_2 \leq_L X_1$ for $\theta_1 \leq \theta_2$. In fact, since $X_2 = X_1 + (\theta_2 - \theta_1)$, with $\theta_2 - \theta_1 > 0$, from (3) it follows that $X_2 \leq_L X_1$. These results can be completed considering $X_1 \sim G(\theta_1,\alpha_1)$ and $X_2 \sim G(\theta_2,\alpha_2)$. Using the transitivity of the Lorenz order and the above results, it is easy to prove that if $\alpha_1 \leq \alpha_2$ and $\theta_1 \leq \theta_2$ then $X_2 \leq_L X_1$. However, what happen if $\alpha_1 < \alpha_2$ and $\theta_1 > \theta_2$? This leads to the following result.

Teorema 7 *Si* $X_1 \sim G(\theta_1,\alpha_1)\,(\alpha_1 \le 1)$ *y* $X_2 \sim G(\theta_2,\alpha_2)$, *con* $\theta_1-\theta_2 > \alpha_2-\alpha_1 > 0$, *entonces* $X_1 \le_L X_2$.

Demostración. En primer lugar observemos que la relación

$$\theta_1-\theta_2 > \alpha_2-\alpha_1 > 0 \qquad (9)$$

es posible únicamente si $\alpha_1 < \alpha_2$ y $\theta_1 > \theta_2$. Teniendo en cuenta que la función de densidad de $X \sim G(\theta,\alpha)$ es (8) y que $\mu = E[X] = \alpha+\theta$, tendremos que

Theorem 7 *For* $X_1 \sim G(\theta_1,\alpha_1)\,(\alpha_1 \le 1)$ *and* $X_2 \sim G(\theta_2,\alpha_2)$, *with* $\theta_1-\theta_2 > \alpha_2-\alpha_1 > 0$, *we have* $X_1 \le_L X_2$.

Proof. First note that the relationship

$$\theta_1-\theta_2 > \alpha_2-\alpha_1 > 0 \qquad (9)$$

is only possible if $\alpha_1 < \alpha_2$ and $\theta_1 > \theta_2$. Now, taking into account the fact that the density of $X \sim G(\theta,\alpha)$ is (8) and $\mu = E[X] = \alpha+\theta$, we have that

$$\frac{f_1(\mu_1 x)}{f_2(\mu_2 x)} = k\cdot\left[\frac{(\alpha_1+\theta_1)x-\theta_1}{(\alpha_2+\theta_2)x-\theta_2}\right]^{\alpha_1-1}\cdot\left[\frac{1}{(\alpha_2+\theta_2)x-\theta_2}\right]^{\alpha_2-\alpha_1}\cdot$$

$$\cdot\exp\left[(\alpha_2-\alpha_1+\theta_2-\theta_1)x\right]$$

Donde/*were*, $k = \dfrac{\Gamma(\alpha_2)}{\Gamma(\alpha_1)}\cdot\exp(\theta_1-\theta_2) > 0.$

De (9) se sigue claramente que $\exp\left[(\alpha_2-\alpha_1+\theta_2-\theta_1)x\right]$ es decreciente en x. Supongamos ahora que $x \ge \theta_1/\mu_1 = \theta_1/[\alpha_1+\theta_1]$. Puesto que $\alpha_1 < \alpha_2$, se deduce que

$$\left[\frac{1}{(\alpha_2+\theta_2)x-\theta_2}\right]^{\alpha_2-\alpha_1} \qquad (10)$$

es también decreciente en x para $x \ge \theta_2/[\alpha_2+\theta_2]$. De (9) se sigue que $\dfrac{\theta_1}{\alpha_1+\theta_1} > \dfrac{\theta_2}{\alpha_2+\theta_2}$.

Por lo tanto, (10) es decreciente en x para $x \ge \theta_1/\mu_1$. Finalmente, llamando

$$h(x) = \left[\frac{(\alpha_1+\theta_1)x-\theta_1}{(\alpha_2+\theta_2)x-\theta_2}\right]^{\alpha_1-1},\ \alpha_1 < 1,$$

From (9) we clearly have that $\exp\left[(\alpha_2-\alpha_1+\theta_2-\theta_1)x\right]$ decreases in x. Now, suppose $x \ge \theta_1/\mu_1 = \theta_1/[\alpha_1+\theta_1]$. Since $\alpha_1 < \alpha_2$, it follows that

$$\left[\frac{1}{(\alpha_2+\theta_2)x-\theta_2}\right]^{\alpha_2-\alpha_1} \qquad (10)$$

is also decreasing in x for $x \ge \theta_2/[\alpha_2+\theta_2]$. From (9) it follows that $\dfrac{\theta_1}{\alpha_1+\theta_1} > \dfrac{\theta_2}{\alpha_2+\theta_2}$.

Thus, (10) decreases in x for $x \ge \theta_1/\mu_1$. Finally, denoting

$$h(x) = \left[\frac{(\alpha_1+\theta_1)x-\theta_1}{(\alpha_2+\theta_2)x-\theta_2}\right]^{\alpha_1-1},\ \alpha_1 < 1,$$

se demuestra a partir de (9) que $h'(x) \leq 0$ si y sólo si $\alpha_2\theta_1 > \alpha_1\theta_2$. Por consiguiente, la razón $f_1(\mu_1 x)/f_2(\mu_2 x)$ es decreciente para $x \geq \theta_1/\mu_1$. Por tanto, aplicando el Corolario 6, el teorema queda demostrado.

it can be proven that $h'(x) \leq 0$ if and only if $\alpha_2\theta_1 > \alpha_1\theta_2$, which follows from (9). Therefore, the assumption implies that the ratio $f_1(\mu_1 x)/f_2(\mu_2 x)$ is decreasing for $x \geq \theta_1/\mu_1$. Thus, applying Corollary 6, the proof is complete.

BIBLIOGRAFÍA/REFERENCES

Arnold, B.C. (1987). *Majorization and the Lorenz Order: A brief introduction*. Berlin: Springer-Verlag.

Arnold, B.C., Brockett, P.L., Robertson, C.A. y Shu, B. (1987). Generating ordered families of Lorenz curves by strongly unimodal distributions. *Journal of Business and Economic Statistics*, *5*, 305-308.

Atkinson, B. (1970). On the measurement of inequality. *Journal of Economic Theory, 2*, 244-263.

Dagum, C. (1985). *Lorenz Curve, encyclopedia of statistical sciencies* (5, pp. 156-161). S. Kotz, N.L. Johnson & C.B. Read (Eds.). New York: Wiley.

Dasgupta, P., Sen, A.K. y Starrett, D. (1973). Notes on the measurement of inequality. *Journal of Economic Theory, 6*, 180-187.

Gail, M.H. y Gastwirth,J.L. (1978). A scale-free goodness-of-fit test for the exponential distribution based on the Lorenz curve. *Journal of the American Statistical Association, 73*, 786-793.

Gastwirth, J.L. (1971). A general definition of the Lorenz curve. *Econometrica*, *39*, 1037-1039.

Kakwani, N. (1984). Welfare ranking of income distributions. *Advances in Econometrics*, *3*, 191-213.

Keilson, J. y Gerber, H. (1971). Some result for discrete unimodality. *Journal of the American Statistical Association, 66*, 386-389.

Rothschild, M. y Stiglitz, J.E. (1973). Some further results on the measurement of inequality. *Journal of Economic Theory, 6*, 188-204.

Shaked, M. (1982). Dispersive ordering of distributions. *Journal of Applied Probability*, *19*, 310-320.

Pecvnia Monográfico 2012, pp. 31-51

DATOS TEXTUALES COMO ELEMENTOS ACTIVOS EN SENSOMETRÍA / *TEXTUAL DATA AS ACTIVE ELEMENTS IN SENSOMETRY*

Ramón Álvarez-Esteban[1]
ralve@unileon.es
Pedro Aguado Rodríguez[2]
pedro.aguado@unileon.es

Universidad de León

Resumen

La utilización de datos textuales en estudios estadísticos sobre sensometría generalmente se ha realizado tratando de explicar e interpretar los resultados alcanzados a partir de datos cuantitativos. Este trabajo muestra una metodología que permite utilizar datos textuales como elementos activos. Dos catas de vinos ilustran el procedimiento.

Palabras clave: Datos textuales; Sensometría; Análisis de Correspondencias; Análisis Factorial Múltiple.

Abstract

The use of textual data in statistical studies into sensometric field has been conducted generally seeking to explain and interpret results obtained from quantitative data. This work shows a methodology that allows use textual data as active elements. Two wine tastings illustrate the procedure.

Keywords: Textual data; Sensometry; Correspondence Analysis; Factorial Multiple Analysis.

[1] Facultad de Ciencias Económicas y Empresariales, Departamento de Economía y Estadística, Área de Estadística e Investigación Operativa. Universidad de León, Campus de Vegazana, 24071-León (España).

[2] Escuela Superior y Técnica de Ingeniería Agraria. Departamento de Ingeniería y Ciencias Agrarias. Área de Ingeniería Agroforestal. Universidad de León.

1. INTRODUCCIÓN

El análisis estadístico de datos sensoriales es un instrumento necesario para conocer las preferencias de los consumidores y comprender cómo perciben los productos.

El estudio de los vinos en sensometría utiliza información que contiene un elevado número de variables.

Generalmente se utilizan formularios estandarizados en los que un panel de expertos analiza y puntúa aspectos visuales, aromáticos y gustativos. La puntuación media global se obtiene como suma de puntuaciones parciales.

Los métodos estadísticos utilizados en análisis sensorial plantean dificultades cuando se aplican a personas no expertas, especialmente en el caso del vino. Expertos y consumidores pueden utilizar diferentes variables en su decisión de compra o bien valorarlas con distinta importancia.

Es frecuente encontrar personas que no son capaces de describir con palabras un vino. Este hecho ha llevado a suponer que todo conocimiento requiere conocer su lenguaje asociado (Brochet y Dubourdieu, 2001). De esta forma, enólogos, sumilleres y aficionados han modelado un lenguaje del vino para describir las propiedades sensoriales. En la mayor parte de los estudios, esta descripción se realiza analizando la presencia o ausencia de las propiedades sensoriales, más que cuantificando éstas (por ejemplo, utilizando adverbios de cantidad "poco", "mucho", etc.).

Tradicionalmente el análisis estadístico de productos u objetos se ha realizado utilizando métodos multidimensionales a partir de matrices de proximidad entre objetos (Takane, 1980, 1982), pero sin tener

1. INTRODUCTION

Statistical analysis of sensory data is a necessary tool to discover consumer preferences and to understand how products are perceived.

Sensometric study on wine field demands interpreting information that involves a large number of variables.

The classical approach is based on the score from a panel of expert tasters analyzing visual, smell and taste aspects, using standardized forms. The overall score is obtained as a sum of weighted partial scores.

Statistical methods used in sensory analysis have difficulties when they are applied to non-experts, especially in the case of wine. Experts and consumers can use different variables in their purchase decision or evaluate them with different emphasis.

It is common to find people who are not able to describe a wine using words. This problem has raised speculation that all knowledge requires the knowledge of the language associated (Brochet and Dubourdieu, 2001). Thus, winemakers, sommeliers and amateurs have built a language to describe wine sensory properties. In most studies, this description is done by analyzing the presence or absence of sensory properties rather than by quantifying these properties (e.g. using adverbs of quantity "little", "a lot", etc.).

Often the analysis of consumer judgment of products or objects has been conducted using statistical analysis based on multidimensional matrices of proximity between objects (Takane, 1980, 1982) but regardless of the language, the descriptions. The results of these tests

en cuenta el lenguaje, las descripciones. Los resultados suelen mostrar solamente una dimensión predominante, una dimensión hedónica que agrupa las preferencias de gusto de los catadores (Berglund et al., 1973).

No obstante, estos resultados parecen estar más relacionados con la forma de recoger la información que con la existencia de una sola dimensión. Entre estos estudios destacan las ordenaciones de productos (*sorting task*) en función de preferencias, obteniendo similares valoraciones entre expertos y no expertos (Lelièvre et al., 2008).

En otros casos se solicita que el catador agrupe productos en función de sus percepciones, obteniendo para cada catador una tabla productos *x* productos en la que "1" indica que los productos *i* y *j* se perciben dentro del mismo grupo y "0" en caso contrario. En la tabla global la frecuencia *ij* indica el número de veces (catadores) que han señalado el producto *i* dentro del mismo grupo que el producto *j* (Abdi et al., 2007). Las agrupaciones suelen ser similares para expertos y no expertos, pero la descripción de éstas suele ser diferente (por ejemplo, los enólogos buscan defectos en los vinos, mientras que los sumilleres buscan virtudes).

El perfil convencional (*conventional profile*, basado en la norma ISO 11035) es uno de los métodos más clásicos utilizados en los estudios sensométricos. Requiere utilizar un lenguaje común para todos los catadores, construyendo una lista consensuada de atributos (Delarue y Sieffermann, 2004) identificando qué descriptores están presentes o no en cada producto, para lo que se necesita la participación de catadores con gran experiencia. A continuación, se cuantifica

often show only one dominant dimension, a dimension that brings hedonic variables that explains the preferences of the tasters (Berglund et al., 1973).

However, these results seem to be more related to how collect the information than the existence of a single dimension. These studies highlight the arrangement of products (sorting task) according to the preferences, obtaining very similar ratings between experts and nonexperts (Lelièvre et al., 2008).

In other cases each taster is asked to perform product groups according to their perceptions. A table products *x* products is obtained for each taster where "1" indicates that the *i* and *j* products are perceived in the same group and "0" otherwise. In the global table of all the tasters, frequency *ij* indicates the number of times (tasters) who have noted the product *i* within the same group as the product *j* (Abdi et al., 2007). Clusters are often similar for experts and non-experts, but the description of these groups is often different (e.g. winemakers said that they look for defects in wines, while sommeliers look for virtues).

The conventional profile (based on ISO 11035) is one of the most classical methods used in sensometric studies. It requires the use of a common language for all the tasters and building a unique list of attributes through consensus (Delarue and Sieffermann, 2004). First we have to identify which descriptors are present or are not in every product, so we need the participation of highly experienced tasters. Then the intensity of these descriptors is measured. Results of conventional profile seem to be more accurated when tested products are simple

la intensidad de los descriptores. El perfil convencional parece ofrecer mejores resultados cuando los productos evaluados son simples, mientras que en productos complejos, los resultados son peores (Lawless, 1995). Otra desventaja es el tiempo elevado que se necesita para el entrenamiento de los expertos.

El principal objetivo de este trabajo es evaluar si las técnicas textuales aplicadas como elementos activos (y no solamente como información suplementaria o ilustrativa) permiten obtener configuraciones de productos que sean estables, a pesar de las variaciones entre expertos señaladas.

Para ello, se han utilizado nuevas técnicas de recogida de información como el napping y las descripciones textuales (ver Campo et al., 2010, sobre una comparación de la frecuencia de citación y el análisis descriptivo en el caso del vino y Sauvageot et al., 2006, sobre grandes variaciones encontradas en descripciones textuales en expertos).

La sección segunda contiene el proceso de recogida de información con dos catas de vinos relacionadas. A partir de las descripciones textuales de ocho vinos y dieciocho catadores se construyen las tablas de frecuencias. Se aplicó un Análisis de Correspondencias (AC) con cada tabla. En la sección cuarta se realiza un Análisis Factorial Múltiple (AFM) para construir la configuración compromiso a partir de las configuraciones del AC con los distintos umbrales. En la sección quinta se analizan las diferencias de los factores y subespacios del AC en relación a esta nueva configuración compromiso. Finalmente se comparan los resultados de esta métodología con los obtenidos en la primera cata.

and worse using complex products (Lawless, 1995). Another disadvantage is the long time required for the necessary training of experts.

The main objective of this study is to assess if textual techniques applied as active data (and not only as supplementary or illustrative information) provides product stable configurations, in spite of the variations among experts pointed out.

New techniques for gathering information such as napping and textual descriptions have been applied (see Campo et al., 2010, for a comparison of descriptive analysis and frequency of citation in the wine case and Sauvageot et al., 2006, about the large variation in textual descriptions among experts).

Second section describes the process of gathering information through two related wine tasting. From textual descriptions of eight wines and eighteen tasters several contingency tables are built using different thresholds of lexical forms. Correspondence Analysis (CA) is applied for each table in section three. Multiple Factorial Analysis (MFA) in section four allows to obtain an average configuration from the coordinates of all CA analyzed. Section five measures the overall differences between the MFA average configuration and individual CA configurations. Finally the results of this methodology are compared with those obtained from the first tasting

2. MATERIAL Y MÉTODOS

Se han seleccionado ocho vinos tintos correspondientes a dos denominaciones de origen (Bierzo y Tierra de León). Los vinos del Bierzo (etiquetados con los números 1, 3, 6 y 7) son del tipo de uva Mencía y los de Tierra de León del tipo Prieto Picudo. Cuatro de estos son vinos jóvenes, dos con seis meses de madera y los otros dos con doce meses (Tabla 1).

2. MATHERIAL AND METHODS

Eight red Spanish wines from two AOC designations (Bierzo and Tierra de León) were selected. The wines of Bierzo are the type of grape Mencía (wines labelled 1, 3, 6 and 7) and the wines of Tierra de León are the type of grape Prieto Picudo. Four wines are young, two with six months in oak barrels and two with twelve months (Table 1).

Tabla 1. Etiquetas de los vinos / *Table 1. Labels of wines*

DO/*AOC*	Meses en barrica *Months in oak barrels*		
	0	6	12
Bierzo (uva Mencía) / (*grape* Mencía)	1; 7	3	6
Tierra de León (uva Prieto Picudo) / (*grape* Prieto Picudo)	4; 8	2	5

Se conocen varios datos químicos: grado de alcohol, ph, acidez total, acidez volátil real, anhídrido sulfuroso libre y el anhídrido sulfuroso total. En esta experiencia han participado dieciocho personas, nueve enólogos, cuatro sumilleres y cinco aficionados con experiencia.

Cada catador tenía en su mesa ocho catavinos homologados, numerados del uno al ocho. El orden de los vinos fue diferente para catador, utilizando el diseño de cuadrados latinos de Williams (Williams, 1946) con objeto de que el orden de cata de los vinos no produzca efectos sobre los resultados finales.

Se realizaron dos catas sucesivas, con una separación de quince minutos.

2.1. Primera cata

Cada catador dispuso de una hoja de papel (mantel) de 60 centímetros de ancho por 40 centímetros de alto (Pagès, 2003, 2005). Las ocho copas con el vino se colocaron en la parte superior, fuera del

Chemical data are obtained: alcohol, pH, total acidity, volatile acidity, free sulfur dioxide and total sulfur dioxide.

In this experiment, eighteen people participated of which nine are oenologists, four sommeliers and five experienced amateurs.

Eight homologated wine glasses numbered from one to eight are placed in each table. Williams Latin-squares design (Williams, 1946) was used for assign the presentation order of wines for each taster looking for the order did not produce effects in final results.

There was two successive tastings, with a break of fifteen minutes between them.

2.1. First wine tasting

Each taster has a sheet of paper (tablecloth) of 60 cm wide by 40 cm high (Pagès, 2003, 2005). The glasses were placed in the top, outside the tablecloth (nappe). A taster will place next two glasses

mantel (*nappe*). Un catador situará dos vinos tanto más próximos cuanto más se parezcan, utilizando su opinión personal (los criterios que el catador considera, de forma global). De la misma forma, dos vinos se encontrarán tanto más alejados cuanto más diferentes sean percibidos.

Dos catadores tendrán configuraciones distintas si no han considerado los mismos descriptores o no los han ponderado de la misma manera. No hay respuestas buenas ni malas, no hay una configuración de mantel a la que haya que aproximarse. Tampoco deben justificar el motivo por el que han posicionado los vinos sobre el mantel.

Esta disposición será registrada en coordenadas numéricas, obteniendo una configuración individual para cada catador.

Este método *napping* de recogida de información permite obtener las diferencias entre vinos y/o catadores, así como la configuración conjunta, pero es necesario utilizar información adicional para conocer los motivos por los que cada catador decide posicionar las copas, para explicar el perfil sensorial de cada catador. Con el fin de obtener esta información, se solicitó que cada catador realizara una breve descripción de cada vino (*ultra-flash profile*) y la escribiera sobre el mismo mantel.

Por último, los catadores dibujaron en el mantel tantos círculos o elipses como desearon, agrupando los vinos que consideraron parecidos en función de sus percepciones.

Para cada catador se construye una tabla vinos *x* vinos que contiene "1" si los dos vinos han sido agrupados juntos y "0" en caso contrario. Tendremos tantas tablas como catadores (18 tablas). Si sumamos todas las tablas podemos establecer un indicador de la distancia entre cada pareja de vinos.

in the tablecloth if he perceives the wines as resemblance, using his personal opinion (the criteria considered as a whole). In the same way, two wines will be so much far away the more different they are perceived.

Two tasters will build two different configurations if they do not consider the same descriptors or they weight them in a different way. There are not right or wrong answers, there is not a reference tablecloth setting to approach. The tasters should not justify the reasons for locating each glass in the tablecloth.

This layout is recorded in numerical coordinates, obtaining individual settings for each taster.

Napping method of gathering information allows obtaining the differences between the wines and/or tasters as well as a joint configuration, but it is necessary to use additional information to know why each taster has placed the glasses in his tablecloth, to explain the sensorial profile of each taster. To obtain this information, each taster was asked to conduct a brief description of each wine (ultra-flash profile) and write it on the same tablecloth.

Finally, the tasters drew so many circles or ellipses in the tablecloth as they would wish, grouping the similar wines depending on their perceptions.

For each taster we build a table wines *x* wines. Each cell contains "1" if the two wines have been grouped together and "0" otherwise. We will have as many tables as tasters (18 tables). If we add all the tables we can obtain an indicator of the distance between each pair of wines.

2.2. Segunda cata

Se construyó el vocabulario de las palabras elegidas en la primera cata. No se incluyeron artículos, preposiciones, etc. Las formas en masculino y femenino son agrupadas, lo mismo que singulares y plurales (Labbé, 1990). Algunos términos que los catadores consideran sinónimos en la primera cata se agrupan como un solo descriptor en la segunda cata. Por último, los términos que tienen una frecuencia baja no son incluidos en la lista ordenada alfabéticamente con la frecuencia de repetición de cada palabra. Esta lista se muestra a los catadores.

El objetivo es el de consensuar las palabras conservadas, permitiendo añadir nuevos términos para atributos relevantes, obteniendo una nueva lista de descriptores consensuados que puedan ser utilizados en la segunda cata.

La segunda cata fue realizada con los mismos vinos, presentados en un orden distinto, con el fin de que los resultados de la primera cata no pudieran condicionar los de la segunda. Cada catador caracterizó los ocho vinos, asociando los descriptores de la lista consensuada, teniendo en cuenta características visuales, aromas directos y retronasales, sensaciones en boca, etc. El objetivo es la obtención de información que permita explicar las configuraciones obtenidas, así como la variabilidad de los datos recogidos.

Una vez realizada la segunda cata se aplicó la lematización. Se añadieron a los términos los adjetivos y palabras cuantificadoras (e.g. "roble viejo", "roble francés", "roble nuevo", "fruta negra", "bien equilibrado", etc.) (Perrin y Pagès, 2009). En algunos casos descriptores con baja frecuencia se agruparon en otro más amplio para no perder información (e.g. "almizcle" es agrupado bajo el descriptor "animal").

2.2. Second wine tasting

A vocabulary was built from the words used in the first tasting. Articles, prepositions, etc. are not included. Masculine and feminine terms are grouped. The same is done for singular and plural forms (Labbé, 1990). Some words that the tasters consider synonymous are grouped as a single descriptor in the second tasting. Words with low frequency are not included in the agreed list because of their little relevance. The total vocabulary arranged alphabetically with the frequency of repetition of the words was given to each taster.

It was allowed to add new terms for relevant attributes. The goal is to agree the words that we must retain for obtaining a new list of agreed descriptors that can be used in the second tasting.

The second tasting was conducted with the same wines, presented to each taster with a different order, so that the results of the first tasting could not prejudge the outcome of the second. Each taster characterized eight wines, associating the descriptors of the agreed list, taking into account the visual characteristics, direct and retronasal aromas, sensations in the mouth, etc. The objective is to obtain information explaining the configurations drawn in the first tasting, as well as the data variability.

Once the second taste has been performed, lemmatisation is applied. Quantifiers and adjectives were added to terms (e.g. "old oak", "French oak", "new oak", "black fruit", "well balanced", etc.) (Perrin and Pagès, 2009). In some cases specific descriptors with low frequency are grouped into a broader descriptor in order to not lose information (e.g. "musk" was grouped under the descriptor "animal").

El nuevo corpus tiene 948 ítems, de los que 151 son diferentes y 44 son *hapax*. La distribución de los ítems para cada vino se muestra en la Tabla 2.

A continuación se muestran los treinta ítems más utilizados, indicando la frecuencia entre paréntesis: astringente (33), madera (32), lácteo (26), capa alta (24), violáceo (24), acidez (22), fruta (22), especias (21), rojo picota (19), capa media (18), alcohólico (17), carbónico (17), cereza (16), amargo (15), fruta roja (15), tánico (15), tostado (15), vainilla (15), capa media-alta (14), cereza picota (14), cuero (14), grosella (14), corto (13), equilibrado (13), seco (13), glicérico (12), ligero (12), mora (12), redondo (12) y teja (12).

Entre los ítems utilizados al menos tres veces, 52 están asociados a características positivas del vino, 17 a negativas y un tercer grupo de 20 palabras que son positivas o negativas dependiendo del contexto.

The new corpus has 948 total items, of which 151 items are different and 44 are hapax. Frequency citation analysis shows the distribution of items for each wine (Table 2).

The thirty most frequently used items, indicating the frequency in parentheses, are: astringent (33), wood (32), lactic (26), high layer (24), violet (24), acid (22), fruit (22), spice (21), picota-red colour (19), medium layer (18), alcoholic (17), carbonic (17), cherry (16), bitter (15), red fruit (15), tanic (15), roasted (15), vanilla (15), medium-high layer (14), picota-cherry (14), leather (14), redcurrant (14), short (13), balanced (13), dry (13), glyceric (12), light (12), blackberry (12), rounded (12) and russet (12).

Among the items used at least three times, 52 different items are associated with positive characteristics of wine, 17 to negative characteristics and a third group of 20 words that they depending on which context they are used.

Tabla 2. Distribución de los ítems / *Table 2. Distribution of items*

Vino / ***wine***	1	2	3	4	5	6	7	8	Total
Número de ítems */number of items*	111	118	122	110	125	127	112	123	948
% total	11.71	12.45	12.87	11.60	13.19	13.40	11.81	12.97	100
Media de ítems por vino y catador *Average items for wine and taster*	6.2	6.6	6.8	6.1	6.9	7.1	6.2	6.8	6,6
Número diferentes ítems del vino *Number of different items of wine*	57	53	55	54	54	59	48	55	-
% diferentes ítems del vino *% different items of wine*	51.35	44.92	45.08	49.09	43.20	46.46	42.86	44.72	-

3. AC SOBRE LA TABLA VINOS X PALABRAS

Entre las diversas formas de tratar la información recogida en la segunda cata, presentamos la realización del AC sobre la tabla de contingencia formada por los

3. CA FROM TABLE WINES X WORDS

Several ways of managing the information can be used. We present the CA implementation on the occurrence table wines *x* words obtained from the second tasting. Frequencies indicate the number

ocho vinos en fila y un número elevado de columnas (palabras), en la que las frecuencias indican el número de veces que cada palabra aparece en la descripción de un vino, como suma de las opiniones de los catadores. Este resultado puede enriquecerse utilizando información de tipo suplementario (e.g. diferenciando entre sumilleres, enólogos y amateurs).

El umbral de palabras escogido para realizar el AC determinará el número de columnas (Tabla 3).

of times that each word appears in the description of a wine, as the sum of the opinions of the tasters. Supplementary information can be added (e.g. differentiating between oenologists, sommeliers and amateurs).

The occurrence table has eight rows (wines) and a high numbers of columns (words). The threshold of words chosen to carry out the CA determines the number of columns (Table 3).

Tabla 3. Número de palabras diferentes y frecuencia total para los umbrales de dos a diez palabras / *Table 3. Number of different words and total frequency for thresholds from two to ten words*

Umbral/***threshold***	1	2	3	4	5	6	7	8	9	10
Palabras distitnas / *Different words*	151	107	89	79	69	59	51	43	39	36
Total palabras / *Total words*	948	904	868	838	798	748	700	644	612	585

La gran variabilidad de los catadores a la hora de describir los vinos (Labbè et al., 2004), la baja frecuencia de las palabras utilizadas y la posibilidad de trabajar con diferentes umbrales hacen que nos planteemos a priori la posibilidad de obtener resultados estables utilizando AC.

Lebart señala que hay que establecer un umbral para que los resultados del AC tengan sentido estadístico (Lebart et al., 1998). Además, estos resultados debieran ser consistentes con los obtenidos a partir de otras metodologías.

Hay que considerar dos aspectos para elegir el umbral con el que efectuar el AC. Primero, elegir un umbral suficientemente elevado para obtener estabilidad en los resultados (en este caso de los vinos). Segundo, elegir un umbral suficientemente pequeño para que el número de formas léxicas utilizadas sea suficiente para interpretar los resultados.

The great variability of the tasters when describing wines (Labbè et al., 2004), the high number of items, the low frequency of items and the possibility of working with different thresholds lead us to consider a priori the possibility of obtaining stable results using CA.

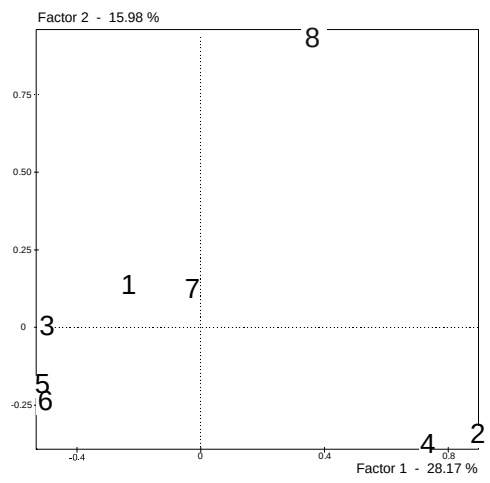
Lebart pointed out that a threshold must be established to achieve than CA results can be the most significant in a statistical sense (Lebart et al., 1998). Moreover, these results should be consistent with those obtained from other methodologies.

The choice of a threshold to carry out the CA can be split into two aspects. First, to choose a high enough threshold that it allows to obtain stability in the results of the CA (in this case of the wines). Secondly, to choose a threshold low enough that the number of used lexical forms is sufficient to interpret the results.

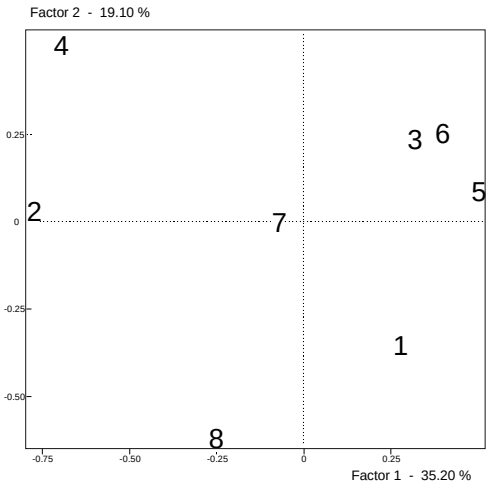
Evidentemente, no es posible comparar las coordenadas de las columnas (palabras) obtenidas de los diferentes AC ya que las palabras son diferentes, pero sí es posible comparar las configuraciones de los ocho vinos. La Figura 1 muestra el primer plano factorial para el AC con umbrales de dos y diez palabras.

Obviously it is not possible to compare the coordinates of the columns (words) obtained from different CA because the words are different, but it is possible to compare the configurations of the eight wines. Figure 1 shows the first factorial plane for CA with thresholds of two and ten words.

Figura 1. Plano factorial del AC para los umbrales de dos y diez palabras
Figure 1. First factorial plane of wines for CA with thresholds of two and ten words



Umbral de dos palabras / *Two words threshold*



Umbral de diez palabras / *Ten words threshold*

La comparación directa de resultados obtenidos a partir de perfiles léxicos diferentes plantea algunos problemas.

En primer lugar, los puntos (vinos) están centrados cuando se ponderan por su frecuencia marginal (número total de palabras que se han retenido para un vino teniendo en cuenta un umbral determinado). En segundo lugar, es posible encontrar reflejos (en la Figura 1 hay reflejos tanto en el eje1 como en el eje2). En tercer lugar, en ocasiones los ejes se intercambian (especialmente cuando los valores propios son muy parecidos). En cuarto lugar, es posible encontrar sub-

The direct comparison of these results obtained from different lexical profiles arise some problems.

First of all, the points (wines) are centered when they are weighted by their marginal frequency (number of words that have been retained for a wine considering a threshold). Secondly, it is possible to find reflections (in Figure 1 we can see reflections in axis 1 and axis 2). In the third place, sometimes the axes are interchangeable (especially when the eigenvalues are very close). Fourth, it is possible to find similar subspaces when using orthogonal rotations (e.g. the first factorial plane is

espacios factoriales similares cuando se utilizan rotaciones ortogonales. Por último, la dilatación de una de las figuras puede hacer disminuir las distancias entre los puntos de las dos configuraciones.

4. AFM DESDE EL AC

De cada AC realizado en el apartado anterior se obtiene una matriz de coordenadas de ocho vinos por siete factores. Esta matriz ha sido centrada por columnas, con el fin de dar a cada uno de los vinos la misma ponderación. Este proceso se repite para todos los umbrales analizados.

A continuación se ha aplicado un AFM (Escofier y Pagès, 1990, 1994), utilizando el paquete de R FactoMineR (Husson et al., 2007). En el AFM se realiza un Análisis de Componentes Principales ACP para cada matriz de coordenadas X_i (ocho vinos *x* siete dimensiones), obteniendo los valores propios λ_l^j, indicando "i" el número del valor propio y "j" el grupo (en nuestro caso el índice del umbral elegido).

Las nueve tablas X_i correspondientes a las coordenadas para los distintos umbrales (de dos a diez) se yuxtaponen en una tabla X, ponderando cada una por la inversa de la raíz cuadrada del primer valor propio λ_1^j procedente del ACP de las j tablas individuales

the same but the factors are not). Finally, the dilation of one configuration can decrease the distances between the points of the two configurations.

4. MFA FROM CA

From the coordinates of the CA in previous section, we will get a matrix of eight wines by seven factors. This matrix has been centered by columns, in order to give the same weight to each wine. This process is repeated for all the thresholds analyzed.

Next a MFA (Escofier and Pagès, 1990, 1994) has been performed using R FactoMineR package (Husson et al., 2007). In MFA, a Principal Component Analysis PCA is carried out for each coordinates matrix X_i (eight wines *x* seven dimensions), obtaining the eigenvalues λ_l^j, notating "i" the number of the eigenvalue and "j" the group (in our case the index of threshold chosen).

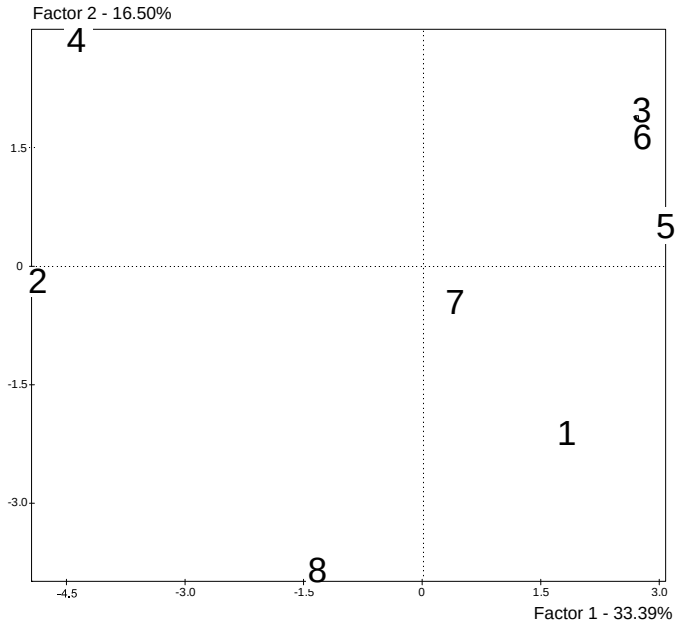
All the nine coordinates tables X_i for the different thresholds (from two to ten) are juxtaposed in a table X, weighting each one by the inverse of the square root of the first eigenvalue λ_1^j from the j PCA individual tables.

$$X = \left(\frac{1}{\sqrt{\lambda_1^1}} X_1, \frac{1}{\sqrt{\lambda_1^2}} X_2, \ldots, \frac{1}{\sqrt{\lambda_1^J}} X_J \right)$$

A partir de la diagonalización de la tabla *X'X* se obtiene la configuración compromiso del AFM (Figura 2).

MFA average configuration is obtained diagonalizing *X'X*. Figure 2 shows the first factorial plane.

Figura 2. Primer plano factorial para la configuración compromiso obtenida del AFM / *Figure 2. First factorial plane for MFA average configuration*



5. DISTANCIA INDIVIDUAL Y GLOBAL

La Figura 2 indica que tan solo hay pequeñas variaciones con las gráficas de la Figura 1 obtenidas para los umbrales 2 y 10.

Entre las medidas para comparar la similitud entre dos configuraciones (distancia global entre la configuración media del AFM y la del AC) se encuentra el coeficiente RV (Krzanowski, 1990), basado en la similaridad normalizada Procrustres.

Si X e Y son dos matrices con las coordenadas de los vinos (ocho) por las dimensiones (siete) y se da la misma ponderación a cada uno de los vinos, siendo *tr()* la traza, el coeficiente *RV* se define como:

5. GLOBAL AND INDIVIDUAL DISTANCE

There are only small variations between the MFA average configuration and the obtained for thresholds two and ten of CA (Figure 2).

Among the measures to compare the similarity between two configurations (global distance between MFA average configuration and CA configuration) we can use the *RV* coefficient (Krzanowski, 1990), based on the normalized Procrustres similarity.

If X and Y are matrices with the coordinates of wines (eight) by the dimensions (seven), giving the same weight to each wine, and notating *tr()* as the trace, *RV* coefficient can be computed as:

$$RV = \frac{tr(XX'YY')}{\sqrt{tr(XX'XX')tr(YY'YY')}}$$

RV varía entre 0 y 1 (Josse et al., 2008). Un valor de RV igual a 1 indica que las dos configuraciones son idénticas. Un valor de RV igual a 0 indica que cada punto de la primera tabla está incorrelado con los de la segunda.

RV varies between 0 and 1 (Josse et al., 2008). A RV value of 1 indicates that both configurations are identical. A RV value of 0 indicates that each point of the first table is uncorrelated to each point of the second table.

La Tabla 4 muestra los coeficientes RV entre las configuraciones para los diferentes umbrales (U) y la compromiso (AFM en itálica).

Table 4 shows the RV coefficient between configurations for the different thresholds (U) of CA and the MFA average configuration (italic letters).

Si solamente se deseara comparar un subespacio es necesario rotar previamente la tabla Y hacia la tabla X.

If only a subspace must be compared it is necessary to rotate the Y table toward X table.

Tabla 4. *RV* entre vinos de las configuraciones AFM y AC para umbrales dos al diez / *Table 4. RV between wines MFA and CA configuration for thresholds from two to ten*

	U2	U3	U4	U5	U6	U7	U8	U9	U10	*MFA*
U2	1.000	0.996	0.991	0.985	0.981	0.971	0.962	0.958	0.955	*0.986*
U3	0.996	1.000	0.998	0.993	0.991	0.981	0.973	0.968	0.967	*0.993*
U4	0.991	0.998	1.000	0.996	0.992	0.984	0.978	0.974	0.973	*0.995*
U5	0.985	0.993	0.996	1.000	0.995	0.991	0.981	0.978	0.977	*0.996*
U6	0.981	0.991	0.992	0.995	1.000	0.996	0.989	0.985	0.984	*0.998*
U7	0.971	0.981	0.984	0.991	0.996	1.000	0.990	0.985	0.984	*0.994*
U8	0.962	0.973	0.978	0.981	0.989	0.990	1.000	0.998	0.997	*0.992*
U9	0.958	0.968	0.974	0.978	0.985	0.985	0.998	1.000	1.000	*0.990*
U10	0.955	0.967	0.973	0.977	0.984	0.984	0.997	1.000	1.000	*0.989*
MFA	*0.986*	*0.993*	*0.995*	*0.996*	*0.998*	*0.994*	*0.992*	*0.990*	*0.989*	*1.000*

RV es muy alto para todos los casos, es decir, las configuraciones de los vinos obtenidas del AC utilizando diferentes umbrales son muy parecidas a la configuración compromiso del AFM. Además la compara-

RV is very high for all cases considered, that is to say, the wines configurations from CA using different thresholds are very similar to the average configuration of the MFA. Furthermore, the comparison

ción de vinos con diferentes umbrales (parte central de la Tabla 4) indica que las configuraciones son muy parecidas, independientemente del umbral elegido.

Altos valores *RV* indican que las configuraciones globalmente son muy parecidas, pero no que todos los factores sean iguales, no que todos los factores puedan ser seleccionados para su interpretación.

Por ello, se han calculado los coeficientes de correlación de Pearson entre cada factor del AFM con el correspondiente factor del AC (Tabla 5).

of wines configurations between different thresholds (middle part of Table 4) shows that the configurations are very close, regardless of the chosen threshold.

High *RV* values indicate that overall configurations are very similar, but not that the factors are equal, not that all factors can be selected for their interpretation.

Therefore, Pearson's correlation coefficients between each MFA factor and the corresponding CA factor have been computed (Table 5).

Tabla 5. Coeficientes de correlación de Pearson entre las coordenadas de los vinos del AFM y las coordenadas del AC para distintos umbrales. Los menores coeficientes están marcados en negro / *Table 5. Pearson's correlation between wines MFA coordinates and CA coordinates for different threshold. Lowest coefficients are pointed out in black*

Umbral \ *Threshold*	r_{11}	r_{22}	r_{33}	r_{44}	r_{55}	r_{66}	r_{77}
2	-0.996	-0.830	-0.699	0.422	0.616	0.937	0.940
3	0.995	0.472	0.524	0.855	0.906	-0.935	0.950
4	0.997	-0.770	-0.793	0.865	0.894	0.968	0.993
5	-0.999	0.786	-0.820	-0.956	-0.862	0.936	0.926
6	0.996	0.940	0.876	0.880	0.934	0.932	0.925
7	0.997	0.967	-0.642	0.591	0.886	0.974	-0.975
8	0.996	-0.979	0.951	0.946	-0.915	0.907	-0.845
9	0.995	0.981	0.929	0.942	0.897	-0.399	0.326
10	0.995	0.993	0.911	-0.925	0.905	-0.289	-0.230

La primera columna r_{11} muestra altas correlaciones entre las coordenadas del primer factor del AFM con las coordenadas del primer factor para el AC con diferentes umbrales. Esto indica que es un factor estable desde el punto de vista de los umbrales escogidos y es posible interpretar las configuraciones de los vinos utilizando las formas lé formas léxicas como descriptores.

First column r_{11} shows high correlations between the coordinates of the first MFA factor and the coordinates for first CA factor for the different thresholds. This means that it is a stable factor from the point of view of the thresholds chosen and it is possible to interpret the wine configuration using the lexical forms as descriptors.

La Tabla 6 muestra un resumen de las formas léxicas que caracterizan el primer factor del AC utilizando un umbral de 5 palabras. Las coordenadas negativas corresponden a características negativas de los vinos.

Para los factores segundo y tercero (Tabla 5) las correlaciones con los tres umbrales más grandes indican factores similares. Las correlaciones son más bajas para los umbrales pequeños (marcadas en negro), por lo que podríamos pensar que no hay estabilidad en el segundo y tercer factor, por lo que no debieran ser interpretados.

Las mayores diferencias entre los factores 2 y 3 del AFM (Figura 3a) con los del AC corresponden al umbral de tres palabras (Figura 3b) con correlaciones de 0.472 y 0.524.

The following table contains an excerpt from the lexical forms that characterize the first CA factor using a threshold of five words (Table 6). Negative coordinates are related to negative characteristics of wines.

For the second and third factors (Table 5), correlations with the three highest thresholds (8, 9 and 10) are very high, indicating similar factors. Smaller thresholds show low correlations (highlighted in black), so we might think that there is no stability in the second and third factors, so they should not be interpreted.

The biggest differences of MFA factors 2 and 3 (Figure 3a) with the CA correspond to three words threshold (Figure 3b) with correlations 0.472 and 0.524.

Tabla 6. Resumen de formas léxicas características del primer factor del AC con umbral de cinco palabras. / *Table 6. Excerpt of lexical forms for the first CA factor using a threshold of five words*

Coordenadas negativas / *negative coordinates* Características negativas / *negative characteristics* a) siempre / *always* b) depende del contexto / *depending on the context*	Coordenadas positivas / *positive coordinates* Características positivas / *positive characteristics*
a) capa media-baja / *medium-low layer*, evolucionado / *evolved*, oxidado / *oxidised*, licor / *liquor*, desequilibrado / *unbalanced*, sucio / *dirty*, herbáceo / *herbaceous*, corto / *short*, carbónico / *carbonic*, acidez / *acid*, ligero / *light*, madera vieja / *old oak*, capa media / *medium layer*, verde / *green*, desagradable / *unpleasant*, amargo / *bitter*, intensidad baja / *low intensity*, intensidad media / *medium intensity.* b) teja / *russet*, caramelo / *caramel*, cereza / *cherry*, color rojo cereza / *cherry-red colour*, fresco / *fresh*, fresa / *strawberry*, frambuesa / *raspberry.*	persistente / *persistent*, capa alta / *high layer*, madera nueva / *new oak*, tostado / *roasted*, redondo / *rounded*, *suave* / *smooth*, intenso / *intense*, fruta negra / *black fruit*, roble francés / *French oak*, rojo cereza / *red cherry*, tonos violáceos / *hints of violet*, color cereza picota / *picota-red colour*, madera / oak, estructurado / *structured*, largo / *long*, cálido / *hot*, torrefacto / *high-roast*, agradable / *pleasant*, vainilla / *vanilla*, glicérico / *glyceric*, toffee / *toffee*, equilibrado / *balanced*, animal /*animal*, violáceo / *violet*, tánico / *tannic*, mora / *blackberry*, capa medio-alta / *high-medium layer*, mineral / *mineral*, coco / *coconut.*

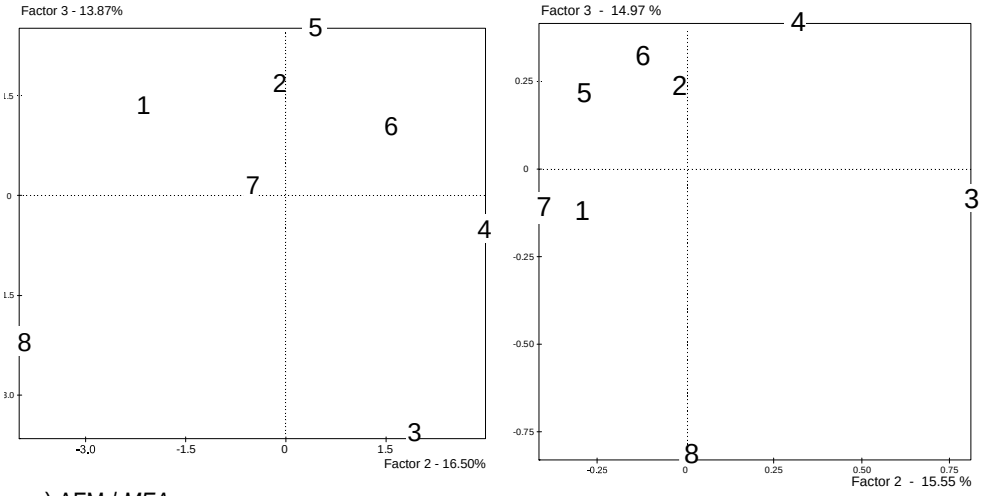
Como indicamos anteriormente, hay que considerar la posibilidad de que haya reflejos, intercambios, rotaciones y dilataciones cuando se comparan configuraciones, especialmente cuando los valores propios son cercanos (15.55% de la varianza del factor dos y 14.97% del factor tres de la Figura 3b).

Si X es la configuración de destino o referencia del MFA e Y es la configuración de origen CA, la matriz de rotación H puede ser obtenida minimizando $\|X\text{-}YH\|$.

As indicated above, there must be considered the possibility of reflections, interchanges, rotations and dilations when comparing configurations, especially when eigenvalues are close (15.55% of variance for factor two and 14.97% for factor three in Figure 3b).

If X is the target or reference MFA configuration and Y is the source CA configuration, H rotation matrix can be obtained minimizing $\|X\text{-}YH\|$.

Figura 3. Plano factorial para los factores dos y tres del AFM y el AC con umbral de tres palabras / *Figure 3. Factorial plane factors two and three from MFA and CA (threshold of three words)*



a) AFM / *MFA*

b) AC para tres palabas / *CA for a threshold of three words*

H se determina descomponiendo $X^{\mathrm{T}}Y$ en las matrices de vectores propios U y V, y la matriz de los valores S, estableciendo la relación $X^{\mathrm{T}}Y=USV^{\mathrm{T}}$. La matriz de rotación se define como $H=VU^{\mathrm{T}}$.

En ocasiones es necesario realizar un cambio de escala obteniendo la matriz dilatada como $Z = \mathrm{c}YH$, siendo *tr()* la traza, c es la

H is determined decomposing $X^{T}Y$ in U and V eigenvectors matrices and a S matrix of eigenvalues, establishing the equation $X^{\mathrm{T}}Y=USV^{\mathrm{T}}$. Rotation matrix is defined as $H=VU^{\mathrm{T}}$.

Sometimes is necessary to carry out a scale exchange. Dilated configuration Z is defined as $Z = \mathrm{c}YH$. Constant dilation c is

constante de dilatación:
$c=tr(YHX^T)/tr(YY^T)$.

Los factores en la Figura 3 obviamente no son los mismos, pero si realizamos una rotación ortogonal (Gower y Dijksterhuis, 2004) de AC hacia la configuración del AFM, las diferencias decrecen mucho (Figura 4).

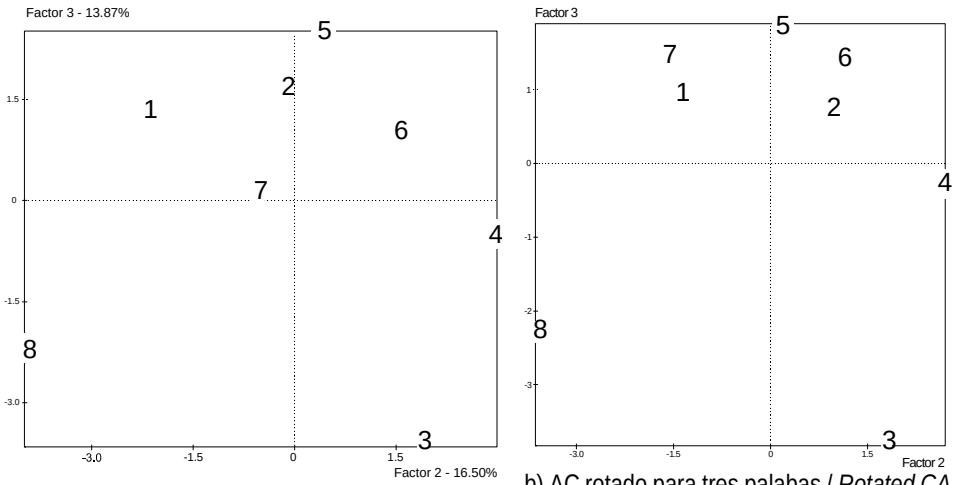
Estos resultados indican que hay diferencias entre los factores considerados uno a uno, pero hay estabilidad en el plano factorial formado por los factores dos y tres con distintos umbrales.

computed as $c=tr(YHX^T)/tr(YY^T)$ when *tr()* is the trace.

Factors in Figure 3 obviously are not the same, but if we perform an orthogonal rotation (Gower and Dijksterhuis, 2004) of CA to MFA configuration differences decrease a lot (Figure 4).

These results show that there are differences between the factors taken one by one but there is stability in the factorial plane made up by factors two and three using different thresholds.

Figura 4. Plano factorial para los factores dos y tres del AFM y el AC rotado con umbral de tres palabras / *Figure 4. Factorial plane factors two and three from MFA and rotated CA (threshold of three words)*



a) AFM / *MFA*

b) AC rotado para tres palabas / *Rotated CA for three words*

La configuración media del AFM ha sido comparada con los resultados del napping de la primera cata, obteniendo resultados consistentes (no incluidos en este trabajo).

The MFA average configuration has been compared with the results of napping from the first taste, getting consistent results (not included in this paper).

6. ANALYSIS CLUSTER

Con el fin de contrastar la consistencia de estos resultados se seleccionaron las coor-

6. CLUSTER ANALYSIS

In order to compare the consistency of these results MFA factorial coordinates of

denadas factoriales del AFM de la segunda cata a partir de los datos textuales considerados como elementos activos. Con estas coordenadas se realizó un análisis cluster con método el promedio entre grupos y la distancia euclídea al cuadrado (Figura 5.a). También se realizó un análisis cluster con la tabla agregada obtenida a partir de los clusters de los vinos creados por los catadores de la primera cata (Figura 5.b).

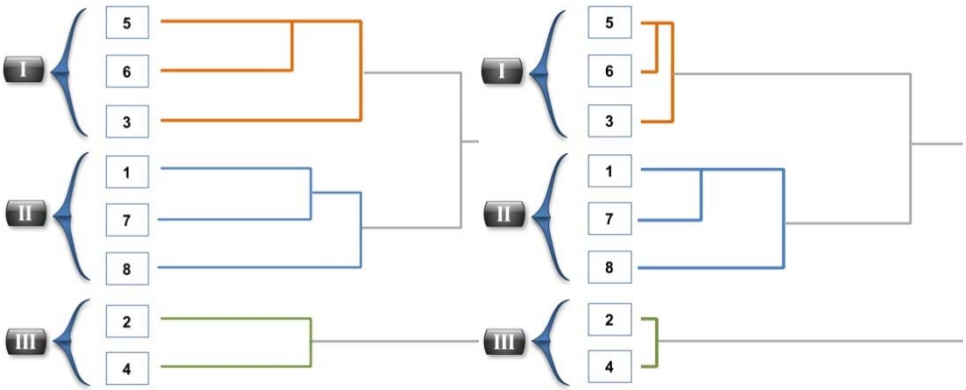
Como puede observarse en la siguiente figura, los tres grupos obtenidos son iguales utilizando las dos metodologías.

the second tasting from textual data considered as active data were selected. With these coordinates, a cluster analysis was performed using the average linkage between groups method and the squared Euclidean distance (Figure 5.a).

A cluster analysis with the aggregated table from the wine clusters created by the tasters in the first tasting was also carried out (Figure 5b).

As shown in the following figure, the three resulting groups are equal using the two methodologies

Figura 5. Análisis clusters utilizando a) las palabras como elementos activos y b) a partir de las agrupaciones de la primera cata / *Figure 5. Clusters analysis using a) words as active elements and b) from groups in the first tasting*



a) Clusters con los factores del AFM de la segunda cata de vinos
Clusters from MFA factors in the second wine tasting

b) Clusters de vinos para la primera cata *Clusters for first wine tasting*

7. CONCLUSIONES

La descripción de propiedades sensoriales de los productos no es una tarea fácil, incluso para los expertos, ya que pueden utilizar diferentes variables y ponderaciones. En la mayoría de los estudios, esta descripción se realiza mediante el análisis de la

7. CONCLUSIONS

Describing sensory properties of products is not an easy task, even for experts because they can use different variables or weighting. In most studies, this description is done by analyzing the presence or absence of sensory properties.

presencia o ausencia de las propiedades sensoriales.

En este trabajo, los datos textuales se aplican como datos activos, con el objetivo de construir configuraciones estables de productos, a pesar de las variaciones entre tres tipos de expertos.

Se realizaron dos sucesivas catas de vino. En la primera se utilizó el "napping" como método de recopilación de información y un "ultra-flash profile", creando un vocabulario y la lista de descriptores. Los expertos consensuaron una nueva lista a partir de este vocabulario.

Antes de aplicar el AC con las palabras de este segundo vocabulario es necesario fijar un umbral para la selección de palabras. Diferentes umbrales proporcionan diferentes configuraciones de coordenadas. Estas coordenadas del AC son seleccionadas para construir una configuración media utilizando el AFM. El coeficiente *RV* indicó que la configuración media del AFM era muy similar a las obtenidas para los diferentes umbrales.

Los resultados del AFM y del AC pueden ser comparados, pero es preciso utilizar rotaciones ortogonales. La metodología utilizada se ha mostrado útil para comparar configuraciones, factores individuales y subespacios.

La utilización de datos textuales como elementos activos ha proporcionado resultados consistentes, incluso al utilizar umbrales de palabras distintos. Esta consistencia se ha manifestado comparando los resultados de datos textuales con los obtenidos a partir de un cluster con los datos no textuales de la primera cata.

In this paper, textual data are applied as active data with the aim of building product stable configurations, in spite of the variations among three types of experts.

There was two successive wine tastings. Napping method of gathering information and ultra-flash profile were used in the first one, building a vocabulary and obtaining a list of descriptors. Experts agreed a new list from the results of the first tasting vocabulary.

Before applying the CA with the words of the second vocabulary it is necessary to fix a threshold for the selection of words. Different thresholds lead to different coordinates configurations. These coordinates from CA are selected to build an average configuration using MFA. *RV* coefficient indicated that average MFA configuration was similar to those obtained for various thresholds.

MFA and CA results can be compared but it can be necessary using orthogonal rotations. Methodology applied is useful to compare global configurations, individual factors and subspaces.

The use of textual data as active elements has led to similar wines configurations, even using different frequency thresholds of words.

This consistency is shown by comparing the textual data results with those obtained from a cluster with non-textual data from the first tasting.

BIBLIOGRAFÍA/REFERENCES

Abdi, H., Valentin, D., Chollet, D. y Chrea, C. (2007). Analyzing assessors and products in sorting tasks: DISTATIS, theory and applications. *Food Quality and Preference*, *18*, 627-640.

Berglund, B., Berglund, U., Engen, T. y Ekman, G. (1973). Multidimensional analysis of twenty-one odors. *Scandinavian Journal of Psychology*, *14*, 131-137.

Brochet, F. y Dubourdieu, D. (2001). Wine Descriptive Language Supports Cognitive Specificity of Chemical Senses. *Brain and Language*, *77*(2), 187-196.

Campo, E., Ballester, J., Langlois, J., Dacremont, C. y Valentin, D. (2010). Comparison of conventional descriptive analysis and a citation frequency-based descriptive method for odor profiling: An application to Burgundy Pinot noir wines. *Food Quality and Preference*, *21*, 44-55.

Delarue, J. y Sieffermann, J-M. (2004). Sensory mapping using Flash profile. Comparison with a conventional descriptive method for the evaluation of the flavour of fruit dairy products. *Food Quality and Preference*, *15*, 383-392.

Escofier, B. y Pagès, J. (1990). *Analyses factorielles simples et multiples: Objectifs, méthodes, interprétations*. Paris: Dunod.

Escofier, B. y Pagès, J. (1994). Multiple factor analysis (AFMULT package). *Computational Statistics and Data Analysis*, *18*, 121-140.

Gower, J.C. y Dijksterhuis, G.B. (2004). *Procrustes problems*. Oxford University Press.

Husson, F., Lê, S. y Mazet, J. (2007). FactoMineR: Factor Analysis and Data Mining with R.R package version 2.4.0, URL http://factominer.free.fr/ (accessed September 2010).

Josse, J., Pagès, J. y Husson, F. (2008). Testing the significance of the RV coefficient. *Computational Statistics and Data Analysis*, *53*, 82-91.

Krzanowski, W.J. (1990). Principles of multivariate analysis, a user's perspective. Oxford Statistical Science Series.

Lawless, H.T., Sheng, N. y Knoops, S. (1995). Multidimensional-scaling of sorting data applied to cheese perception. *Food Quality and Preference*, *6*, 91-98.

Labbé, D. (1990). Normes de dépouillement et procédures d'analyse des textes politiques, CERAT.

Labbé, D., Rytz, A. y Hugi, A. (2004). Training is a critical step to obtain reliable product profiles in a real food industry context. *Food Quality and Preference*, *15*, 341-348.

Lebart, L., Salem, A. y Berry, L. (1998). Exploring textual data. Dordrecht, Boston: Kluwer Academic Publisher.

Lelièvre, M., Chollet, S., Abdi, H. y Valentin, D. (2008). What is the validity of the sorting task for describing beers? A study using trained and untrained assessors. *Food Quality and Preference, 19*, 697-703.

Pagès, J. (2003). Recueil direct de distances sensorielles: Application à l'évaluation de dix vins blancs du Val-de-Loire. *Sciences des Aliments*, *23*, 679-688.

Pagès, J. (2005). Collection and analysis of perceived product interdistances using multiple factor analysis: application to the study of 10 white wines from the Loire Valley. *Food Quality and Preference*, *16*(7), 642-649.

Perrin, L. y Pagès, J. (2009). Construction of a product space from the ultra-flash profiling method: application to 10 red wines from the Loire Valley. *Journal of Sensory Studies*, *24*(3), 372-395.

Sauvageot, F., Urdapilleta, I. y Peyron, D. (2006). Within and between variations of texts elicited from nine wine experts. *Food Quality and Preference*, *17*(6), 429-444.

Takane, Y. (1980). Analysis of categorizing behavior by a quantification method. *Behaviormetrika*, *8*, 75-86.

Takane, Y. (1982). IDSORT: An individual differences multidimensional scaling for sorting data. *Behavior Research Methods and Instrumentation*, *14*, 546.

Williams, E.J. (1949). Experimental designs balanced for the estimation of residual effects of treatments. *Australian Journal of Scientific Research*, Ser. A 2, 149-168.

Pecvnia Monográfico 2012, pp. 53-83

ANÁLISIS DEL FRACASO EMPRESARIAL POR SECTORES: FACTORES DIFERENCIADORES
CROSS-INDUSTRY ANALYSIS OF BUSINESS FAILURE: DIFFERENTIAL FACTORS

Mª Jesús Mures Quintana[1]
mj.mures@unileon.es
Ana García Gallego[1]
ana.ggallego@unileon.es
Mª Eva Vallejo Pascual[1]
eva.vallejo@unileon.es

Universidad de León

Resumen

El objetivo de este trabajo se centra en el análisis del fracaso empresarial por sectores, a fin de identificar los factores explicativos y predictivos de este fenómeno que son diferentes en tres de los principales sectores que se distinguen en toda economía: industria, construcción y servicios.
Para cada uno de estos sectores, seguimos el mismo procedimiento. En primer lugar, aplicamos un análisis de componentes principales con el que identificamos los factores explicativos del fracaso empresarial en los tres sectores. A continuación, consideramos dichos factores como variables independientes en un análisis discriminante, que aplicamos para predecir el fracaso de una muestra de empresas, utilizando no sólo información financiera en forma de ratios, sino también otras variables no financieras relativas a las empresas, así como información externa a las mismas que refleja las condiciones macroeconómicas bajo las que desarrollan su actividad.

Palabras clave: Fracaso empresarial; Ratios financieros; Información no financiera; Análisis discriminante; Sectores.

Abstract

This paper focuses on a cross-industry analysis of business failure, in order to identify the

[1] Facultad de Ciencias Económicas y Empresariales, Departamento de Economía y Estadística, Área de Estadística e Investigación Operativa. Universidad de León, Campus de Vegazana, 24071-León (España).

explanatory and predictor factors of this event that are different in three of the main industries in every economy: manufacturing, building and service.
For each one of these industries, the same procedure is followed. First, a principal components analysis is applied in order to identify the explanatory factors of business failure in the three industries. Next, these factors are considered as independent variables in a discriminant analysis, so as to predict the firms' failure, using not only financial information expressed by ratios, but also other non-financial variables related to the firms, as well as external information that reflects macroeconomic conditions under which they develop their activity.

Keywords: Business failure; Financial ratios; Non-financial information; Discriminant analysis; Industries.

1. INTRODUCCIÓN

La predicción del fracaso empresarial es un importante campo de investigación dentro de la literatura financiera, que en los últimos tiempos ha adquirido gran trascendencia, debido a la severa crisis económica y financiera que está afectando a muchos países en Europa y el resto del mundo.

En este contexto de incertidumbre, las empresas, cuya actividad es la base para el desarrollo de las economías, sufren tales consecuencias que les llevan a su fracaso o crisis o, incluso, a su desaparición. Esto tiene importantes efectos en los agentes con los que la empresa se relaciona, tales como accionistas, empleados, clientes y proveedores, pero también en otros negocios que actúan en el sistema económico (Lang y Stulz, 1992), lo que puede provocar un impacto general de gran magnitud en las economías. Por tanto, se hace necesario disponer de herramientas que permitan anticipar el posible fracaso de una empresa y, por consiguiente, evitar los efectos que el fracaso empresarial tiene sobre los agentes económicos que participan en el sistema.

1. INTRODUCTION

Business failure prediction is an important research field in corporate finance literature, which has become topical in recent times, due to the serious economic and financial recession which is affecting many countries in Europe and all over the world.

In this uncertainty context, businesses, whose activity is the basis for the development of the economies, suffer from such serious consequences that lead them to their failure or crisis or even to their disappearance. That has important effects on the agents with whom the firm has relationships, such as stakeholders, employees, clients and suppliers, but also on other businesses acting in the economic system (Lang and Stulz, 1992), which can cause a general impact of great magnitude on the economies. Therefore, some tools are necessary in order to anticipate the possible failure of a firm and, consequently, to avoid the effects that business failure has on the economic agents taking part in the system.

Este tipo de escenarios ha motivado la elaboración de modelos de predicción del fracaso empresarial, cuyo objetivo es predecir las dificultades a que una empresa pueda estar sometida, puesto que tratan de predecir el fracaso, utilizando para ello la información que las empresas publican en sus estados financieros, además de otro tipo de variables no financieras relativas a la empresa e información externa a la misma. Pero la evolución no se observa sólo en la información considerada para predecir el fracaso, sino también en las metodologías aplicadas para obtener los respectivos modelos predictivos. Desde el estudio univariante de Beaver (1966) y el primer modelo multivariante discriminante de Altman (1968), que son considerados pioneros en este campo, se han aplicado los modelos de probabilidad condicional y herramientas procedentes del campo de la inteligencia artificial, con el objetivo de mejorar los modelos elaborados con anterioridad.

Una característica común de esos modelos es la utilización de muestras heterogéneas de empresas pertenecientes a diferentes sectores o de muestras centradas en empresas que operan en el sector industrial. Este tipo de muestras parece ser la razón de la discrepancia entre los buenos resultados de clasificación *ex-post* y los comparativamente decepcionantes resultados de validación (*ex-ante*) (Smith y Liou, 2007), ya que los ratios utilizados como variables independientes pueden verse afectados por efectos sectoriales diferenciales. Es decir, "los sectores pueden diferir con respecto a los factores de producción, ciclos de vida de los productos, estructura competitiva y modos de distribución, lo que provoca diferencias en varias medidas de condición financiera" (Platt y Platt, 1990: 32).

This kind of scenarios has motivated the development of business failure prediction models, whose aim is to foresee the difficulties which a firm can face as they try to predict business failure. These models use the information in the financial statements published by the firms, as well as another kind of non-financial firm-related variables and other firm information from external sources. Not only is this evolution in the information considered to predict failure observed, but also in the statistical methodologies applied in order to obtain the respective prediction models. From Beaver´s (1966) univariate study and Altman´s (1968) first multivariate discriminant model, which are considered to be pioneering in business failure prediction research, conditional probability models and tools from artificial intelligence field have been applied, in order to improve the previously developed models.

A common figure of most of those studies is the use of heterogeneous samples of firms belonging to different industries of the economy or samples focused on firms operating in the manufacturing industry. This kind of samples seems to be the reason of discrepancy between the good *ex-post* classification results and the comparatively disappointing validation outcomes (*ex-ante* results) (Smith and Liou, 2007), as the ratios used as independent variables can be affected by differential industry effects. That is, "industries can differ with respect to factors of production, product life cycles, competitive structure, and distribution modes which cause industry differences in various measures of financial condition" (Platt and Platt, 1990: 32).

La influencia de los efectos sectoriales en los ratios financieros y su aplicación en la predicción del fracaso empresarial ha sido estudiada por Lincoln (1984) y McDonald y Morris (1984). El primero, a fin de medir niveles de riesgo de insolvencia, analizó cuatro sectores (industria, comercio, inmobiliario y financiero), para los que desarrolló modelos específicos de cada sector, además de un modelo combinado basado en los cuatro y modelos a partir de muestras de dos sectores, por su similitud respecto a su estructura financiera (industria-comercio e inmobiliario-financiero). Utilizando diferentes combinaciones de ratios mediante la aplicación de análisis discriminante, obtuvo que "en la mayoría de los casos la capacidad clasificatoria de las funciones obtenidas de los datos específicos de cada sector era superior a la de las derivadas de los datos de dos y de los cuatro sectores", lo que puede explicarse por "la falta de homogeneidad en los datos […] cuando se combinan sectores" (Lincoln, 1984: 330).

McDonald y Morris (1984), por su parte, analizaron la validez estadística del método de ratios en el análisis financiero tanto en una muestra con varios sectores como en un sector homogéneo (el de servicios públicos), llegando a la conclusión de que el método es más válido para éste último, debido al hecho de que "los ratios no tienen características distributivas similares en diferentes sectores" (McDonald y Morris, 1984: 94).

Con el objetivo de tener en cuenta las diferencias entre sectores, algunas autores han adoptado un enfoque denominado "relativo al sector" (*industry-relative*), ajustando los ratios de las empresas por el valor mediano (Izan, 1984) o medio (Platt y Platt, 1990) para el respectivo ratio

Influence of industry effects on financial ratios and their application on business failure prediction has been studied by Lincoln (1984) and McDonald and Morris (1984). The former, in order to measure levels of insolvency risk, analysed four industries (manufacturing, retail, property, and finance), for which he developed not only single-industry models, but also a combined model based on all four, as well as other models based on two-industry samples, due to their similarity in their financial structure (manufacturing-retail and property-finance). Using different combinations of ratios in a discriminant analysis, he found that "in most cases the classification accuracy of functions derived from single-industry data was superior to those derived from two-industry data and four-industry data", which can be explained by "the lack of homogeneity in the data […] when industries are combined" (Lincoln, 1984: 330).

McDonald and Morris (1984), on the other hand, analysed the statistical validity of the ratio method in financial analysis both in a cross-industry sample and in one homogeneous industry (the utility one), drawing the conclusion that ratio method is more valid for the latter, due to the fact that "ratios do not have similar distributional characteristics across various industries" (McDonald and Morris, 1984: 94).

In order to take into account the differences across industries, some researchers have adopted an industry-relative approach, adjusting the firms' raw ratios for the median (Izan, 1984) or the mean (Platt and Platt, 1990) value for the respective ratio in each firm's industry at a point in time. In both studies, the classification results in the models with unadjusted ratios were lower

en el sector de cada empresa en un momento de tiempo. En ambos trabajos, los resultados de clasificación en los modelos con ratios no ajustados eran inferiores a los obtenidos con ratios ajustados al sector, confirmando que este enfoque es útil y "especialmente atractivo para modelos donde las empresas representan una extensa selección de sectores industriales" (Izan, 1984: 319).

Chava y Jarrow (2004) también evaluaron la importancia de incluir los efectos sectoriales en la predicción del fracaso, analizando diez sectores industriales principales clasificados en cuatro grupos. Incluyeron esta agrupación en los modelos mediante variables *dummy*, que resultaron significativas.

Otra posibilidad para solucionar el problema de las diferencias sectoriales en los ratios financieros es la elaboración de modelos específicos de predicción del fracaso empresarial referidos a un solo sector, en especial aquéllos con peculiaridades respecto a su actividad o la información que publican, como el financiero (Laffarga *et al*., 1985; Pina, 1989) y el asegurador (Mora, 1994). Sin embargo, hay un conjunto de trabajos sobre el sector industrial (Beaver, 1966; Altman, 1968; Deakin, 1972; Ohlson, 1980; Taffler, 1982). Estos modelos específicos tienen la ventaja de centrarse sólo en un sector, por lo que no es necesario ajustar por las diferencias sectoriales, lo que exige conocer el valor medio o mediano para cada sector incluido en una muestra que incluya varios sectores.

Con el objetivo de comprobar si había diferencias sectoriales entre las empresas españolas respecto al fracaso empresarial, Román *et al*. (2001) y la Junta de Andalucía (2004) han desarrollado distintos

than the ones achieved with industry-relative ratios, confirming that the industry-relative approach is useful, and "particularly appealing for models where firms represent a broad cross-section of industrial sectors" (Izan, 1984: 319).

Chava and Jarrow (2004) also evaluated the importance of including industry effects in bankruptcy prediction, analysing ten major industrial sectors classified in four groups. They included this grouping into the models by using dummy variables, which resulted to be significant.

Another possibility to solve the problem of industry differences in financial ratios is the development of specific business failure prediction models referred to just one sector, especially those with peculiarities regarding their activity or the information they publish, such as the financial (Laffarga *et al*., 1985; Pina, 1989) or the insurance industries (Mora, 1994). However, there is a whole set of studies on the manufacturing industry (Beaver, 1966; Altman, 1968; Deakin, 1972; Ohlson, 1980; Taffler, 1982). These specific models have the advantage of focusing on one industry alone, so it is not necessary to adjust for industry differences, which requires knowing the mean or median value for each industry included in a cross-industry sample.

With the aim of proving whether there were industry differences among the Spanish firms regarding business failure, Román *et al*. (2001) and Junta de Andalucía (2004) have developed different prediction models in both a heterogeneous sample of firms belonging to several industries and some single-industry samples.

modelos predictivos, tanto en una muestra heterogénea de empresas pertenecientes a varios sectores como en muestras específicas de un sector. En general, "el grado de acierto alcanzado en el modelo global es claramente inferior a los que se llega en cada uno de los modelos sectoriales" (Román *et al*., 2001: 9). Asimismo, en ambos estudios puede observarse que hay factores del fracaso que son comunes a todos los sectores, pero también algunos son específicos de un sector.

Considerando el contexto descrito, el objetivo de nuestro trabajo es analizar el fracaso empresarial en España desde una perspectiva sectorial, a fin de concluir si las variables que mejor explican y describen el fracaso difieren por sectores, además de comparar los resultados de clasificación en cada sector. Para ello, consideramos un muestra de empresas de pequeño y mediano tamaño con domicilio social en la Comunidad Autónoma de Castilla y León, que dividimos en cuatro sectores principales según la actividad de las empresas, de las que recogemos tanto información financiera como variables no financieras y externas. Una vez recogida la información, aplicamos en primer lugar un Análisis de Componentes Principales (ACP), con el fin de reducir el número de variables potencialmente explicativas del fracaso empresarial. A continuación, los factores identificados los consideramos como variables independientes a entrar en una función discriminante, que obtenemos para predecir el fracaso de las empresas de la muestra en los diferentes sectores.

A fin de alcanzar nuestro objetivo, el artículo se organiza del siguiente modo: en el siguiente epígrafe hacemos referencia a la recogida de información, que implica el desarrollo de tres etapas:

In general, "the hit rate achieved in the global model is clearly lower than the ones which are obtained in each one of the single-industry models" (Román *et al*., 2001: 9). Furthermore, in both studies it can be observed that there are failure factors which are common to all industries, but several of them are also specific to one industry.

Taking into consideration the presented context, the aim of this paper is to analyse business failure in Spain from a cross-industry perspective, so as to conclude if the variables that best explain and predict failure are different across industries, as well as to compare the classification results in each industry. In order to do that, we consider a sample of small and medium-sized firms with head office in the region of Castile and León (Spain), which is divided into four main industries according to the firms' activity, from which we collect financial information as well as non-financial and external variables. Once information is collected, we first apply a Principal Components Analysis (PCA), so as to reduce the number of variables which can potentially explain business failure. Next, the identified factors are considered as independent variables to enter in a discriminant function, which is obtained in order to predict the failure of the firms in the sample across industries.

In order to reach our target, the paper is organised as follows: In the next section we will deal with the data collection, which involves the development of three stages: the definition of what it is understood by business failure, the sample

la definición de lo que se entiende por fracaso empresarial, el proceso de selección de la muestra y la elección de las variables financieras y no financieras que contribuyen a explicar y predecir el fenómeno del fracaso. En el epígrafe 3 presentamos los resultados empíricos obtenidos en las diferentes muestras de empresas que operan en cada sector, tanto respecto a la selección de variables mediante el ACP como los resultados predictivos obtenidos con la aplicación del análisis discriminante. El trabajo finaliza con las principales conclusiones extraídas del mismo.

2. RECOGIDA DE INFORMACIÓN

La primera etapa en el desarrollo de un modelo de predicción del fracaso empresarial es la recogida de información necesaria para obtener el modelo. Puesto que el objetivo de cualquiera de estos modelos es identificar las variables que mejor discriminan entre empresas fracasadas y no fracasadas, la muestra de estudio debe incluir ambos tipos de empresas. Por tanto, antes de proceder a seleccionar la muestra, es preciso decidir lo que se entiende por fracaso empresarial, ya que hay una variedad de situaciones con influencia negativa en la actividad empresarial que podrían considerarse como definición de este fenómeno.

Una vez definido el fracaso empresarial, seleccionamos la muestra objeto de estudio. A diferencia de la mayoría de modelos previos, que han utilizado una muestra emparejada con el mismo número de empresas fracasadas y no fracasadas, seleccionamos una muestra aleatoria, sobre la base del tamaño y composición de la población. Elegida la muestra, la dividimos en tres de los principales sectores que se distinguen en toda economía: industria, construcción y servicios, a fin de obtener un modelo predictivo diferente para cada sector.

selection process, and the choice of the financial and non-financial variables that help explain and predict the failure event. Section 3 introduces and discusses the empirical results in the different samples of firms operating in each industry, referring to both the selection of variables by PCA and the prediction results achieved by discriminant analysis. The paper concludes with the summary remarks.

2. DATA COLLECTION

The first step in the development of a business failure prediction model is the collection of the data that are necessary to obtain the model. Provided the aim of any of these models is to identify the variables that best discriminate between failed and non-failed firms, the study sample must include both kinds of firms. Therefore, before selecting the sample, it is important to decide what it is understood by business failure, since there are a variety of situations with a negative influence on firms' activity, all of them could be considered as a definition of that event.

Once business failure is defined, the study sample is selected. Unlike most of the previous models, which have used a paired sample with the same number of failed and non-failed firms, a random sample is chosen, based on the population size and composition. Once it is selected, we divide it into three of the main industries in every economy: manufacturing, building and service, in order to develop a different prediction model for each industry.

Con el fin de desarrollar los modelos, es necesario considerar un conjunto de variables que contribuyen a explicar y predecir el fracaso. Además de ratios financieros, que reflejan la actividad de las empresas, ya que se calculan a partir de sus estados financieros, es importante tener en cuenta otro tipo de información no financiera que también influye en el futuro fracaso de una empresa.

2.1. Definición de fracaso empresarial

Como hemos mencionado, la primera decisión a tomar en el desarrollo de un modelo de predicción del fracaso empresarial corresponde a la definición de lo que se entiende por dicho fenómeno.

Es evidente que el fracaso empresarial hace referencia a una situación negativa que afecta a la actividad empresarial. Una revisión de la literatura previa[2] en este campo pone de manifiesto la existencia de diferentes definiciones de fracaso, según el objetivo del respectivo trabajo, que a su vez se define atendiendo a las necesidades de los potenciales usuarios de cada modelo. Si una empresa fracasa, las consecuencias que conlleva sobre los distintos agentes implicados en la empresa (como inversores, prestamistas y proveedores, clientes, trabajadores, gerentes o auditores) son diferentes. Por consiguiente, el fenómeno que se utilice como definición de tal situación también debería ser diferente, puesto que los mencionados agentes son los potenciales usuarios de cualquier modelo de predicción del fracaso y cada uno busca una aplicación diferente cuando utiliza el modelo a efectos de predecir el fracaso empresarial.

To develop the models, it is necessary to consider a set of variables that contribute to explain and predict failure. As well as financial ratios, which reflect the firms' activity out of their financial statements, it is important to take into account other kind of non-financial information that also has an influence on a firm´s future failure.

2.1. Definition of business failure

As it has been said, the first decision in the development of a business failure prediction model is stating the definition of what it is understood by that event.

It is clear that business failure refers to a negative situation affecting a firm's activity. A review of the previous literature[3] in this field shows different definitions of failure, depending on the aim of the respective model, defined according to the needs of the potential users. If a firm fails, the consequences on the different parties involved in the firm (such as investors, lenders and suppliers, clients, employees, managers or auditors) are different. Therefore, the event used as a definition of this situation should also be different, provided those parties are the potential users of any prediction model and they seek a different applicability when using the model in order to predict business failure.

[2] Además de los trabajos citados en este artículo, un resumen de los modelos más relevantes desarrollados tanto en Estados Unidos como en otros países europeos y de todo el mundo puede encontrarse en las revisiones bibliográficas realizadas por Altman (1984), Dimitras *et al*. (1996), Cybinski (2001) ó Ravi Kumar & Ravi (2007).

[3] As well as the cited works in the paper, a summary of the most relevant models developed both in the United States and other European and all over the world countries can be found in the literature reviews published by Altman (1984), Dimitras *et al*. (1996), Cybinski (2001) or Ravi Kumar & Ravi (2007).

Entre las definiciones más utilizadas en los principales estudios realizados podemos considerar las siguientes:

– la declaración formal de quiebra o cualquier otro procedimiento legal (Altman, 1968; Taffler, 1982; Laffarga *et al*., 1985; Ohlson, 1980; Zmijewski, 1984; Peel *et al*., 1986; Pina, 1989; Theodossiou, 1991; Odom y Sharda, 1992; Dimitras *et al*., 1999; Charitou *et al*., 2004);
– fracaso en el sentido de insolvencia, como la incapacidad de la empresa para pagar sus deudas a medida que vencen (Edmister, 1972; Laitinen, 1991), o
– un conjunto de situaciones diferentes, además de las dos anteriores (Altman *et al*., 1994; Laitinen y Laitinen, 1998), como el descubierto bancario y la falta de pago a accionistas preferentes (Beaver, 1966; Deakin, 1972) o un acuerdo explícito con los acreedores para reducir deudas (Blum, 1974; Elam, 1975).

Como puede observarse, hay un gran listado de situaciones negativas que pueden considerarse como definición del fracaso empresarial. Sin embargo, la mayoría de modelos han utilizado una definición jurídica del fracaso, ya sea quiebra o liquidación o cualesquiera otros conceptos, según la legislación vigente en cada país, debido a la ventaja de ser un acontecimiento sumamente evidente que puede fecharse de manera objetiva (Keasey y Watson, 1991). Además, la mayoría de modelos contienen ratios financieros como variables independientes para predecir el fracaso, por lo que esta definición evitaría los problemas que podría ocasionar el hecho de que tanto las variables predictoras como el suceso que tratan de predecir estuvieran basados en los mismos estados financieros, si se utilizara un criterio más económico, como el nivel de ingresos o la posición de liquidez (Jones, 1987).

Among the most widespread definitions in the mainstream studies we can consider the following ones:

– A firm's formal declaration of bankruptcy or another legal proceeding (Altman, 1968; Taffler, 1982; Laffarga *et al*., 1985; Ohlson, 1980; Zmijewski, 1984; Peel et al., 1986; Pina, 1989; Theodossiou, 1991; Odom and Sharda, 1992; Dimitras *et al*., 1999; Charitou *et al*., 2004);
– Failure in the sense of insolvency, as the inability of a firm to pay debts as they fall due (Edmister, 1972; Laitinen, 1991), or
– A group of different situations, as well as the two previous ones (Altman *et al*., 1994; Laitinen and Laitinen, 1998), such as an overdrawn account and the nonpayment of a preferred stock dividend (Beaver, 1966; Deakin, 1972) or an explicit agreement with creditors to reduce debts (Blum, 1974; Elam, 1975).

As it can be observed, there is a large list of negative situations that can be considered as a definition of business failure. Nevertheless, the majority of the developed models have used a juridical definition of failure, either bankruptcy or liquidation or whatever other used concepts, attending the current legislation in each country, because it has the advantage of being a highly visible event that can be objectively dated (Keasey and Watson, 1991). Furthermore, most models contain financial ratios as independent variables to predict failure, so this definition would avoid the problems involved by the fact that both the predictor variables and the event they try to predict are based on the same financial statements if a more economic criterion, such as income level or liquidity position, was used (Jones, 1987).

Teniendo en cuenta las ventajas de la definición legal de fracaso sobre otras situaciones negativas que afectan a la actividad de las empresas, también consideramos esta situación como subrogado del fracaso empresarial, definiendo éste como la declaración formal, por parte de la empresa, de uno de los tres posibles procedimientos concursales recogidos en la legislación española, y que están incluidos bajo el término general de *bankruptcy*.

2.2. Muestra de empresas

El siguiente paso en el desarrollo de los modelos de predicción del fracaso es la selección de la muestra. Para ello, nos centramos en la Comunidad Autónoma de Castilla y León y utilizamos la base de datos SABI para recoger la información relativa a la muestra seleccionada.

El objetivo de nuestro estudio es comparar el fracaso empresarial en tres de los principales sectores que pueden identificarse en toda economía: industria, construcción y servicios. Sin embargo, con el fin de seleccionar una muestra representativa de la población, en primer lugar consideramos toda la población de empresas en la base de datos y, una vez seleccionada una muestra aleatoria de empresas, la dividimos en tres submuestras según la actividad de cada empresa.

En los modelos desarrollados con anterioridad, el método de muestreo más utilizado ha sido el de obtener una muestra denominada "basada en el estado" (*state-based sample*, Zmijewski 1984), ya que consiste en seleccionar la muestra de empresas fracasadas, según la definición de fracaso empresarial considerada, y emparejarlas con empresas sanas del mismo sector y tamaño, lo que resulta en una muestra constituida por el

Taking into consideration the advantages of the legal definition of failure over other negative situations affecting firms' activity, we also consider this state as a surrogate of business failure, defining this as the firm's formal declaration of one of the three possible proceedings in the Spanish law, which are included under the general terminology of bankruptcy.

2.2. Sample of firms

The next step forward in the development of prediction models is the sample selection. In order to do that, we focused on the region of Castile and León (Spain) and used the database SABI to collect the information related to the selected sample.

The target of our study is to compare business failure in three of the main industries that can be identified in every economy: manufacturing, building and service. Nevertheless, in order to select a representative sample of the population, we first consider the whole firms' population in the database. Once a random sample of firms is selected, we divide it into the three subsamples according to each firm´s activity.

In the previously developed models, the most common sampling method has been to derive a so-called state-based sample (Zmijewski, 1984), which consists in selecting the sample of failed firms, according to the used definition of business failure. Next, these firms are matched to the non-failed ones devoted to the same industry and being of the same size, which results in a sample composed by the same number of companies in both groups. This kind of

mismo número de empresas en los dos grupos. Si bien esta clase de muestra tiene la ventaja de asegurar un número suficiente de empresas fracasadas, pues hay una baja tasa de empresas que fracasan en la economía, se obtiene aplicando un método de muestreo no aleatorio que no respeta las proporciones poblacionales en la muestra, cuando los métodos estadísticos clásicos que se utilizan en los modelos de predicción del fracaso están basados en un diseño muestral aleatorio (Balcaen y Ooghe, 2006). Por consiguiente, los estimadores de los parámetros son inconsistentes y sesgados, lo que lleva a una sobrestimación de la capacidad predictiva del modelo (Palepu, 1986), ya que la tasa de error para las empresas fracasadas está subestimada (Balcaen y Ooghe, 2006).

Con el fin de solventar todos los inconvenientes planteados, aplicamos un procedimiento de muestreo "mixto", que combina las ventajas tanto del muestreo aleatorio como no aleatorio. En primer lugar, identificamos la población de empresas en la base de datos SABI, utilizada para la recogida de información, con el requisito de disponibilidad en la misma de los estados financieros para un periodo de tres ejercicios económicos consecutivos.

Teniendo en cuenta el criterio elegido para el fracaso empresarial, identificamos 59 empresas fracasadas. Debido a la baja tasa de fracaso en la población (un total de 41.584 compañías), elegimos estas 59 empresas para formar la muestra de fracasadas, como una especie de muestreo no aleatorio, a fin de asegurar un número suficientemente grande de este grupo de empresas en la misma. Respecto a las empresas sanas o no fracasadas, seleccionamos una muestra aleatoria, sobre la base del tamaño y la composición de la población. Utilizando

sample has the advantage of assuring a big enough number of failed firms, as there is a low frequency rate of failing firms in the economy. However, it is obtained by applying a non-random sampling method that does not respect the population proportions in the sample. On the contrary, classical statistical methods used in failure prediction models are based on the assumption of a random sampling design (Balcaen and Ooghe, 2006). Therefore, parameter estimates are inconsistent and biased, which leads to an overstatement of the model's ability to predict (Palepu, 1986), as the misclassification error rate for the failed firms is understated (Balcaen and Ooghe, 2006).

In order to solve all these drawbacks, we applied a "mixed" sample derivation, which combines the advantages of both random and non-random sampling. First of all, we identified the firms' population in the database SABI, used to collect the information, with the requirement of availability of financial statements for three consecutive economic years.

Taking into account our criterion for business failure, there were 59 failed firms. Due to the low failure rate in the population (41,584 companies altogether), we chose the 59 firms to derive the failed sample, as a kind of non-random sampling, in order to ensure a big enough number of firms in this group. Regarding the non-failed firms, a random sample was selected, on the basis of their population size and composition. Using the formulae appropriate to calculate the size for the non-failed firms group, it resulted in a sample size of 396 companies. In order to respect characteristics and peculiarities

la fórmula adecuada para determinar el tamaño de este grupo de empresas, resultó un tamaño muestral de 396 compañías. Por otro lado, al objeto de respetar las características y peculiaridades de los diferentes sectores, las empresas no fracasadas fueron seleccionadas del mismo sector en que operaban las empresas fracasadas, atendiendo al tamaño poblacional en cada uno.

Una vez seleccionada la muestra total, cada empresa fue clasificada según su actividad, tal como están codificadas en la Clasificación Nacional de Actividades Económicas (CNAE-93) a nivel de dos dígitos. Atendiendo a esta clasificación, la muestra fue dividida en cuatro submuestras, correspondientes a los cuatro sectores principales identificados en toda economía: agricultura, industria, construcción y servicios, tal como puede observarse en la Tabla 1, donde se recoge un resumen de la muestra objeto de estudio.

of different industries, non-failed firms were selected from the same industry in which failed companies developed its activity, according to each industry population size.

Once the total sample was selected, each firm was classified according to their activity as it is coded in the Spanish Industrial Classification Code (CNAE-93) considering two digits. Taking into account this classification, the sample was divided into four subsamples, corresponding to the four main industries identified in every economy: agriculture, manufacturing, building and service, as it can be observed in Table 1, where there is a summary of the study sample.

Tabla 1. Muestra de empresas / *Table 1. Firms' sample*

Sector económico / *Industry*		Empresas fracasadas *Failed firms*		Empresas no fracasadas *Non-failed firms*	
Actividad principal *Activity*	Código CNAE–93 / *CNAE–93 Code*	Número *Number*	Porcentaje *Percentage*	Número *Number*	Porcentaje *Percentage*
Agricultura *Agriculture*	01	5	8,5	14	3,5
Industria *Manufacturing*	14-36	22	37,3	59	14,9
Construcción *Building*	45	12	20,3	85	21,5
Servicios *Service*	50-85	20	33,9	238	60,1
Total		59	100	396	100

Debido al bajo número de empresas en el sector de la agricultura, decidimos eliminarlo del estudio y realizar el análisis respecto a los otros tres sectores, con el objetivo de comparar los resultados sobre fracaso empresarial en cada sector.

2.3 Variables del estudio

Dado que el objetivo de nuestro estudio es caracterizar el fracaso empresarial en cada uno de los tres sectores identificados en la muestra total, es necesario seleccionar un conjunto de variables que se supone discriminan entre empresas fracasadas y no fracasadas, contribuyendo de este modo a la explicación y predicción del fenómeno de interés.

En primer lugar, es obvio que el fracaso de una empresa depende fundamentalmente de la actividad que desarrolla, que se refleja en la información que publica en sus estados financieros. Por tanto, el primer tipo de información a considerar para predecir el fracaso es una variedad de ratios financieros que se calculan relacionando diferentes partidas contables, como una forma más fácil de tratar toda la información financiera contenida en los estados contables. Debido a la falta de una teoría económica sobre fracaso empresarial que sirviera de guía para elegir ratios, la selección ha sido básicamente empírica, basada en su popularidad en la literatura y en su frecuencia y nivel de significación en la investigación previa, en la línea iniciada por Beaver (1966), si bien la consideración de estos criterios ha dado lugar a un amplio listado de ratios potencialmente explicativos del fracaso empresarial.

En un intento por reducir ese gran número, para seleccionar los ratios financieros de nuestro estudio nos hemos limitado a los ratios utilizados (y que

Due to the small number of firms in the agriculture industry, we decided to eliminate it from the study and to run the analysis concerning the other three industries, with the aim of comparing the results regarding business failure in each industry.

2.3 Variables of study

Provided the target of our study is to characterize business failure in each of the three industries identified in the total sample, it is necessary to choose a set of variables that are supposed to discriminate between failed and non-failed firms, contributing in that way to the explanation and prediction of the event of interest.

First of all, it is obvious that a firm's failure mainly depends on the activity it develops, which is reflected in the information published in its financial statements. Therefore, the first type of information to consider when predicting failure is a variety of financial ratios that are computed relating different accounting entries, as an easier way of treating all the financial information in the statements. Due to the lack of an economic theory of business failure that could be a guide to select ratios, the selection has been basically empirical, based on their popularity in literature and their use and predictive success in previous research, as Beaver (1966) did, although the use of these criteria has resulted in a huge list of ratios potentially explanatory of business failure.

In an attempt to reduce that large list, in order to select the financial ratios for our study we have limited to the ratios used (and significant) in several of the previously developed models and especially those of

resultaron significativos) en varios de los modelos desarrollados previamente y en especial los de Beaver (1966) y Altman (1968), dado que sus trabajos son considerados pioneros en este campo y sus ratios han sido utilizados en gran cantidad de modelos desarrollados con posterioridad.

En todo caso, la selección de variables también ha estado influenciada por la disponibilidad de información para las empresas de la muestra, puesto que la información fue recogida para un periodo de tres ejercicios consecutivos: en el caso de las empresas fracasadas, el periodo de tres años anteriores al fracaso y los tres últimos años de actividad, para las no fracasadas.

Beaver (1966) and Altman (1968), provided they are considered to be pioneering in this field and because most of their ratios have also been used in a big number of subsequently developed models.

In any case, the variable selection has also been influenced by information availability for the firms in our sample, since the information was collected for a period of three consecutive years: in the case of failed firms the three-year period before failure, and the last three years of activity for the non-failed ones.

Tabla 2. Ratios financieros utilizados como variables independientes
Table 2. Financial ratios used as independent variables

Categoría *Category*	Nombre *Name*	Definición / *Definition*
Liquidez *Liquidity*	RCI	Ratio de circulante o liquidez general: Activo circulante ÷ Pasivo circulante *Current ratio: Current assets ÷ Current liabilities*
	PAC	Prueba ácida: (Activo circulante – Existencias) ÷ Pasivo circulante *Acid test: (Current assets – Inventories) ÷ Current liabilities*
	LIQ	Liquidez inmediata: Disponible (Tesorería) ÷ Pasivo circulante *Quick ratio: Cash ÷ Current liabilities*
	CCA	Capital circulante: Capital circulante ÷ Activo total *Working capital ÷ Total assets*
	CCFO	Capital circulante: Capital circulante ÷ Fondos propios / *Working capital ÷ Equity*
Rentabilidad *Profitability*	ROA	Rentabilidad económica: Resultado del ejercicio ÷ Activo total *Return on assets: Net income ÷ Total assets*
	ROE	Rentabilidad financiera: Resultado del ejercicio ÷ Fondos propios *Return on equity: Net income ÷ Equity*
	REAC	Rentabilidad sobre fondos de accionistas: Resultado antes de impuestos ÷ Fondos propios / *Earnings before taxes ÷ Equity*
	ROAII	Rentabilidad económica: Resultado antes de impuestos ÷ Activo total *Earnings before taxes ÷ Total assets*

Endeudamiento y solvencia *Leverage and solvency*	REP	Nivel de endeudamiento: Pasivo exigible ÷ Activo total *Total liabilities ÷ Total assets*
	RECP	Endeudamiento a corto plazo: Pasivo circulante ÷ Activo total *Current liabilities ÷ Total assets*
	RELP	Endeudamiento a largo plazo: Pasivo fijo ÷ Activo total *Fixed liabilities ÷ Total assets*
	NPA	Autonomía financiera (solvencia): Fondos propios ÷ Activo total *Equity ÷ Total assets*
	FPPC	Fondos propios ÷ Pasivo circulante / *Equity ÷ Current liabilities*
	EQUI	Cobertura de inmovilizado o equilibrio: (Fondos propios + Pasivo fijo) ÷ Activo fijo / *(Equity + Fixed liabilities) ÷ Fixed assets*
	CCF	Cobertura de cargas financieras: Resultado de explotación ÷ Gastos financieros / *Operating result ÷ Financial expenses*
	GFV	Cobertura de cargas financieras: Gastos financieros ÷ Importe neto cifra de ventas / *Financial expenses ÷ Sales*
Rotación y actividad *Turnover and activity*	RAC	Rotación de activo: Importe neto de la cifra de ventas (INCV) ÷ Activo total *Sales ÷ Total assets*
	Var(INCV)	Crecimiento de la cifra de ventas: $INCV_t \div INCV_{t-1}$ / $Sales_t \div Sales_{t-1}$
	CCV	Capital circulante ÷ Importe neto de la cifra de ventas / *Working capital ÷ Sales*
	PPAG	Rotación de activo circulante: Activo circulante ÷ Ingresos de explotación *Current assets ÷ Operating income*
Recursos generados *Cash-flow*	CFAT	Recursos generados sobre estructura económica: *Cash-flow* ÷ Activo total *Cash flow ÷ Total assets*
	CFDT	Capacidad de devolución de la deuda: *Cash-flow* ÷ Pasivo exigible *Cash flow ÷ Total liabilities*
	CFPC	Capacidad de devolución de la deuda a corto plazo: *Cash-flow* ÷ Pasivo circulante / *Cash flow ÷ Current liabilities*
Estructura *Economic structure*	AC	Activo circulante ÷ Activo total / *Current assets ÷ Total assets*
	AF	Activo fijo ÷ Activo total / *Fixed assets ÷ Total assets*
	TES	Tesorería ÷ Activo total / *Cash ÷ Total assets*

Todos los criterios señalados resultan en un listado final de 27 ratios financieros, que son clasificados en los tradicionales grupos de liquidez, rentabilidad, endeudamiento y solvencia, rotación y actividad, *cash-flow* y estructura. Aparecen recogidos en la Tabla 2, junto con su respectiva definición.

Además de ratios financieros, la información no financiera también tiene influencia en el fracaso empresarial, por lo que su inclusión en los modelos de predicción podría mejorar los resultados obtenidos cuando sólo se consideran variables financieras. En este sentido, Jones (1987) señala que no hay razón para limitar el estudio a

All the mentioned criteria result in a final list composed of 27 financial ratios, which are classified in the traditional groups of liquidity, profitability, leverage and solvency, turnover and activity, cash-flow, and economic structure. They are shown in Table 2, together with their respective definition.

As well as financial ratios, non-financial information also has an influence on failure business, so its inclusion in the prediction models could improve the results achieved when considering only financial variables. In this sense, Jones (1987) pointed out that there is no reason to limit the study to

los ratios financieros y que los modelos multivariantes podrían incrementar su poder predictivo incorporando otro tipo de información, como variables cualitativas y macroeconómicas. Esta es la razón por la que consideramos como potenciales variables independientes para predecir el fracaso variables no financieras relativas a la empresa, cuya definición se recoge en la Tabla 3.

Por otro lado, las empresas desarrollan su actividad bajo ciertas condiciones macroeconómicas, por lo que es importante considerar este tipo de información en cualquier modelo de predicción del fracaso. Estas variables externas tratan de medir el estado general de la economía y están definidas como la variación porcentual respecto al año anterior, como puede observarse en la Tabla 3.

financial ratios and that multivariate models could increase prediction power by incorporating another kind of information, such as qualitative and macroeconomic variables. That is the reason why we consider as potential independent variables to predict failure other firm-related non-financial variables whose definition is shown in Table 3.

On the other hand, firms develop their activity under some specific macroeconomic conditions, so it is important to take into consideration this kind of information in any business failure model. These external variables try to measure the general state of the economy and are defined as the percentage change on previous year, as it can also be observed in Table 3.

Tabla 3. Variables independientes no financieras
Table 3. Non-financial independent variables

Categoría *Category*	Nombre *Name*	Definición / *Definition*
Relativas a la empresa *Firm-related*	VIDA	Tiempo (en meses) desde la constitución de la empresa *Time (in months) from firm incorporation date*
	FORMA	1, si la empresa es sociedad de responsabilidad limitada; 0, si es sociedad anónima *1, if the firm is a private limited company (Ltd); 0, if it is a public limited company (plc)*
	AÑO	Último año disponible de las cuentas anuales / *Year of last financial statements*
	SECTOR	Sector de actividad, a nivel de dos dígitos de la CNAE-93 / *Firm economic activity industry, defined by two-digit Spanish Industrial Classification code (CNAE-93)*
Externas *External*	PIB_Nac	Variación porcentual del Producto Interior Bruto (a nivel nacional) *Percentage change in Gross National Product*
	IPI_Nac	Variación porcentual del Índice de Precios Industriales a nivel nacional *Percentage change in National Producer Price Index*
	IPI_CyL	Variación porcentual del Índice de Precios Industriales en Castilla y León *Percentage change in Castilla y León Producer Price Index*
	IPC_Nac	Variación porcentual del Índice de Precios de Consumo a nivel nacional *Percentage change in National Consumer Price Index*
	IPC_CyL	Variación porcentual del Índice de Precios de Consumo en Castilla y León *Percentage change in Castilla y León Consumer Price Index*
	Tac_Nac	Variación porcentual de la Tasa de actividad a nivel nacional *Percentage change in National Activity Rate*
	Tac_CyL	Variación porcentual de la Tasa de actividad en Castilla y León *Percentage change in Castilla y León Activity Rate*
	Tip_inte	Variación porcentual del Tipo de interés legal del dinero *Percentage change in Legal Interest Rate*

3. ESTUDIO EMPÍRICO: ANÁLISIS DE RESULTADOS

Una vez seleccionada la muestra y recogida la información correspondiente a las variables independientes, elaboramos un modelo de predicción del fracaso empresarial para cada sector mediante la aplicación del análisis discriminante, en el que consideramos como variables predictoras los ratios financieros y la información no financiera. Como paso previo, aplicamos un ACP, a fin de identificar aquellos factores con un alto poder explicativo del fracaso empresarial y reducir el número de variables seleccionadas como potencialmente predictoras de este fenómeno.

3.1. Selección de variables: Análisis de Componentes Principales (ACP)

El principal objetivo del ACP es obtener un conjunto de factores que resumen la información proporcionada por las variables independientes, lo que nos permite reducir el número de variables a entrar en los modelos predictivos.

Como hemos señalado, consideramos tres tipos de variables: ratios financieros, información no financiera relativa a la empresa y variables externas. Por lo que se refiere a la información relativa a la empresa, sólo FORMA y VIDA fueron consideradas como variables potencialmente explicativas, puesto que la variable AÑO fue tenida en consideración al seleccionar la muestra y estamos desarrollando un modelo para cada sector (variable SECTOR). Por tanto, debido a su pequeño número, las consideramos directamente como variables independientes a entrar en la función discriminante y aplicamos el ACP sólo sobre los ratios financieros y las variables externas.

3. EMPIRICAL ANALYSIS AND RESULTS

Once the study sample was selected and the information concerning the independent variables collected, a business failure prediction model for each industry was developed by applying a discriminant analysis, where financial ratios and non-financial information were considered as predictor variables. As a previous step, a PCA was applied in order to identify those factors with a high explanatory power over business failure and reduce the number of variables chosen as potentially predictor of this event.

3.1. Variable selection: Principal Components Analysis (PCA)

The main target of the PCA is to obtain a set of factors summarizing the information provided by the independent variables, which allows us to reduce the number of variables to enter the prediction models.

As it has been mentioned, three kinds of variables are considered: financial ratios, non-financial firm-related information and external variables. Regarding the firm-related information, only FORMA and VIDA were considered as potentially predictor variables, since AÑO was taken into consideration in the sample derivation and we are developing a prediction model for each industry (variable SECTOR). Therefore, due to their small number, they were directly considered as independent variables to enter the discriminant function and PCA was only applied on the financial ratios and external variables.

En primer lugar, aplicamos el análisis sobre la información financiera y externa, pero observamos un comportamiento claramente diferenciado entre los ratios financieros, por un lado, y las variables externas, por otro. Por consiguiente, decidimos desarrollar el análisis por separado con las variables pertenecientes a cada grupo.

En el caso de la información financiera, el ACP se aplicó sobre el listado inicial de 27 ratios, referidos al último año del periodo en estudio. Aquellos ratios no correlacionados con ninguno de los factores extraídos fueron eliminados en pasos sucesivos. Además, para incrementar el porcentaje de varianza explicada por los factores, los ratios que contenían información redundante también fueron eliminados del análisis. Todo el proceso fue realizado con el programa estadístico SPAD 6.0.

En cada sector, de acuerdo con el procedimiento descrito, se extrajeron seis factores. En el sector de la industria, 19 ratios se correlacionaban con los factores, que explicaban el 82,73% de la información original expresada por los 27 ratios financieros. En el sector de construcción, los seis factores extraídos explicaban el 84,86% de la varianza, correlacionándose con ellos 18 ratios. Por último, en el sector servicios, el porcentaje de varianza explicada fue del 70,96%, siendo 19 los ratios correlacionados con los seis factores obtenidos.

Los ratios financieros correlacionados con los factores extraídos en cada sector se recogen en la Tabla 4. A su vez, se muestra la definición dada a los diferentes factores, según las correlaciones entre ratios y factores.

Como puede observarse, hay varios factores (y ratios relacionados) que son comunes a las tres submuestras. Estos

First of all, the analysis was applied to both the financial and external information, but it was observed a clearly different behaviour among financial ratios, on the one hand, and external variables, on the other. Consequently, we decided to run the analysis separately with the variables belonging to each group.

In the case of financial information, PCA was applied on the initial list of 27 financial ratios, referred to the last year of the study period. Those ratios which did not correlate with any of the obtained factors were deleted in successive steps. Moreover, to increase the variance percentage explained by the factors, ratios containing redundant information were also removed from the analysis. The whole process was made with the statistical software SPAD 6.0.

In each industry, according to the described procedure, six factors were finally obtained. In the manufacturing industry, 19 ratios were correlated with the factors, which explained 82.73% of the original information expressed by the 27 financial ratios. In the building industry, the extracted six factors explained 84.86% of the variance and 18 ratios were correlated with any of them. Finally, in the service industries, the percentage of explained variance was 70.96%, with 19 ratios being correlated to the six obtained factors.

The ratios correlated with the extracted factors in each industry are shown in Table 4. Furthermore, the description given to the different factors is also shown, according to the correlations between ratios and factors.

As it can be observed, there are some factors (and related ratios) that are common to the three subsamples. These

son descritos como "estructura de recursos", puesto que está correlacionado con la proporción de fondos propios (NPA), deuda a corto plazo (RECP) y *cash-flow* (CFAT) sobre activo total; la capacidad de las empresas para pagar deudas con sus propios recursos, bien generados de forma interna (CFDT, CFPC) o externa (FPPC); y "liquidez", ya que este factor se correlaciona con varios ratios que miden este aspecto de la actividad de la empresa. A este respecto, el ratio de liquidez inmediata (TES) representa un factor propio en los sectores de construcción y servicios.

factors are described as: 'liability structure', since this factor is correlated to the proportion of equity (NPA), current debt (RECP) and cash-flow (CFAT) on total assets; the firms' ability to pay debts with their own resources, internally (CFDT, CFPC) or externally (FPPC) generated; and 'liquidity', as this factor is correlated to several ratios measuring this issue of the firm's activity. In this regard, the quick ratio (TES) represents a characteristic factor in the industries of building and service.

Tabla 4. Factores explicativos y ratios que los caracterizan
Table 4. Factors from PCA and variables related

Significado del factor *Factor description*	Variables			
	Industria *Manufacturing*	Construcción *Building*	Servicios *Service*	
Estructura de recursos *Liability structure*	NPA CFAT RECP	NPA CFAT RECP	NPA CFAT RECP	
Capacidad de devolución de deuda / *Ability to return debts*	CFDT CFPC FPPC	CFDT CFPC FPPC	CFDT CFPC FPPC	
Liquidez / *Liquidity*	RCI LIQ PAC	RCI LIQ PAC	RCI LIQ PAC	
Tesorería / *Cash*	TES	TES	TES	
Circulante / *Current position*	AC CCA	AC CCA	AC CCA	
Rotación / *Turnover*	PPAG GFV CCV RAC	PPAG GFV CCV –	PPAG – – RAC	
Rentabilidad económica *Economic profitabililty*	ROA –	ROA ROAII	ROA –	Margen de beneficios *Profit margin*
Fondos propios / *Equity*	ROE CCFO –	– – –	ROE CCFO EQUI	
Rentabilidad de accionistas *Stakeholders profitability*	–	REAC	REAC	

Otro factor común a las tres submuestras es el denominado "circulante", pues está correlacionado con los ratios de activo circulante (AC) y capital circulante (CCA) sobre activo total en los tres sectores, aunque en el de servicios, también se correlaciona con este factor el ratio PPAG. Dado que este ratio mide la rotación del activo circulante, en los otros dos sectores contribuye, junto con otros ratios que reflejan la rotación de diferentes partidas contables (GFV, CCV), a crear un factor definido como "rotación".

Un factor denominado "fondos propios" también es común a los sectores industrial y de servicios, mientras que éste último comparte con el de construcción un factor definido por la rentabilidad de accionistas (REAC).

Por último, podemos distinguir algunos factores específicos de un sector. En la construcción, se extrae un factor denominado "rentabilidad económica", ya que está correlacionado con los dos ratios que miden este aspecto (ROA, ROAII). En el sector servicios, ROA y la rotación del activo total (RAC) definen el factor "margen de beneficios".

Por lo que respecta a la información externa, los resultados del ACP son similares en los tres sectores, lo que es lógico, puesto que estas variables reflejen el entorno económico en el que las empresas desarrollan su actividad y es el mismo para todas las empresas, cualquiera que sea su actividad.

En cada sector, todas las variables se correlacionaron con los dos primeros factores extraídos, que explicaban más del 90% de la información original expresada por las ocho variables externas. En concreto, esos porcentajes fueron del 96,11, 95,45 y 91,74% de la varianza en los sectores de industria, construcción y servicios, respectivamente.

Another common factor to the three subsamples is the one termed as 'current position', provided it is correlated to the ratios of current assets (AC) and working capital (CCA) on total assets in the three industries. In the service industries the PPAG ratio is also correlated with this factor. Since this ratio measures the current assets turnover, in the other two industries it contributes, together with some other ratios reflecting the turnover of different accounting entries (GFV, CCV), to create a 'turnover' factor.

An 'equity' factor is also common to the manufacturing and the service industries, whereas the latter shares with the building industry a factor defined by the stakeholders' profitability (REAC).

Finally, some specific factors to an industry can be distinguished. In the building industry, a factor named as 'economic profitability' is extracted, provided it is correlated to the two ratios measuring this issue (ROA, ROAII). In the service industries, ROA and total assets turnover (RAC) define the 'profit margin' factor.

Regarding the external information, PCA results were similar in the three industries, which is logical since these variables reflect the economic environment in which firms develop their activity and this is the same for every firm, whatever its activity is.

In each industry, all the variables were correlated with the first two extracted factors, which explained more than 90% of the original information expressed by the eight external variables. Specifically, those percentages were 96.11, 95.45 and 91.74% of the variance in the manufacturing, building and service industries, respectively.

El primer factor fue definido como "actividad económica general" (FAC1_EXT), ya que se correlacionaba con el PIB, el tipo de interés, el Índice de Precios Industriales a nivel nacional y las tasas de actividad nacional y en Castilla y León, mientras que el segundo factor se correlacionaba con los Índices de Precios de Consumo e Industriales en Castilla y León, además del Índice de Precios de Consumo a nivel nacional, por lo que se denominó "nivel de precios" (FAC2_EXT).

3.2. Resultados de la predicción

Con el fin de predecir el fracaso empresarial de las empresas en cada submuestra, aplicamos un análisis discriminante. A pesar de los supuestos estadísticos exigidos, este análisis es uno de los métodos estadísticos más utilizados en el campo del fracaso empresarial, donde se han obtenido buenos resultados de predicción.

Los ratios financieros correlacionados con los factores extraídos por el ACP en cada sector y los dos factores que resumen la información externa, junto con las mencionadas variables relativas a la empresa, se consideraron como variables predictoras para estimar los diferentes modelos. Respecto a la información financiera, decidimos incluir los ratios medidos en los tres años de nuestro periodo de estudio, evitando de este modo el inconveniente de obtener un modelo para cada año del periodo de tres. El programa estadístico SPSS 19 se utilizó para desarrollar los modelos predictivos, que se obtuvieron aplicando un procedimiento sucesivo por pasos hacia delante considerando cualquiera de los criterios de selección de variables disponibles en el programa.

The first factor was defined as 'general economic activity' (FAC1_EXT), as it was correlated to GNP, interest rate, national producer price index, and both national and Castile and León activity rates, while the second one was correlated with both consumer and producer price indexes in Castile and León, as well as the national consumer price index, so it was named as 'price level' (FAC2_EXT).

3.2. Prediction results

In order to predict the failure of the firms in each subsample, discriminant analysis was applied. Despite its required statistical assumptions, this analysis is one of the most used statistical methods in business failure field, generally obtaining good prediction results.

The financial ratios correlated with the six extracted factors by PCA in each industry and the two factors summarizing the external information, together with the mentioned firm-related variables, were used as predictor variables in order to estimate the different models. Regarding financial information, we decided to include the ratios measured in the three years of our study period, avoiding in this way the drawback of obtaining a model for each year of the three-year period. Statistical software SPSS 19 was used to develop the prediction models, which were obtained applying a forward stepwise procedure according to any of the variable selection criteria available in the software.

Tabla 5. Resultados del análisis discriminante (Industria)
Table 5. Discriminant analysis results (Manufacturing industry)

Variable	Coeficientes de la función discriminante *Discriminant function coefficients*		*F* parcial *Partial F-value*	Valor *p* *P-value*	Coeficientes de estructura *Structure coefficients*
	No estandarizados *Non-standardized*	Estandarizados *Standardized*			
FAC1_EXT	–0,3233	–0,7433	20,4272	0,0000	–0,500
CFAT	–1,7077	–0,4156	3,3078	0,0744	–0,375
FORMA	–1,3851	–0,6153	12,4693	0,0008	–0,379
NPA	–0,0142	–0,5638	6,2835	0,0152	–0,424
Constante *Constant*	0,9832	–	–	–	–

Los resultados del análisis discriminante para el sector industrial se recogen en la Tabla 5. Cuatro variables entraron en la función discriminante, por su significación en la predicción del fracaso empresarial en este sector, aunque el ratio de cash-flow sobre activo total referido al último año anterior al fracaso (CFAT) sólo es significativo a un nivel del 10%, como indica el valor *p* asociado al estadístico *F* parcial. Las otras variables significativas son el porcentaje de fondos propios sobre activo total el mismo año (NPA), la variable que indica si la empresas es una Sociedad de Responsabilidad Limitada (FORMA) y uno de los factores externos, en concreto el denominado "actividad económica general" (FAC1_EXT). De acuerdo con los coeficientes estandarizados y de estructura, la variable más importante en la función discriminante es el factor externo, seguido de la FORMA y los dos ratios financieros. Todas las variables significativas tienen una influencia negativa en la variable dependiente, esto es, en la puntuación discriminante, como también indica el signo de los coeficientes.

Discriminant analysis results for the manufacturing industry are shown in Table 5. Four variables entered the discriminant function, because of their significance in predicting business failure in this industry, although the ratio of cash-flow on total assets referred to the last year before failure (CFTA) is only significant at a 10% level, as the *p*-value associated to the partial *F* shows. The other significant variables are the percentage of equity on total assets the same year (NPA), the variable indicating if the firm is Ltd (FORMA), and one of the external factors, specifically the one named as 'general economic activity' (FAC1_EXT). According to the structure and standardized coefficients, the most important variable in the discriminant function is this external factor, followed by FORMA and both financial ratios. All the significant variables have a negative influence on the dependent variable, that is, the discriminant score, as the sign of the coefficients also means.

La Tabla 6 muestra los resultados del análisis discriminante llevado a cabo en la submuestra de empresas de la construcción. En este caso, cinco variables resultaron significativas. Se trata del mismo factor externo que en la industria y de cuatro ratios financieros: la proporción de capital circulante sobre activo total (CCA) el último año previo al fracaso y tres ratios medidos el año anterior: los porcentajes de *cash-flow* (CFAT_1) y deuda a corto plazo (RECP_1) sobre activo total y la rentabilidad de accionistas (REAC_1).

Table 6 shows the results from the discriminant analysis carried out on the building subsample. In this case, five variables became significant. They are the same external factor than in the manufacturing industry and four financial ratios: the proportion of working capital on total assets (CCA) the last year previous to failure, and three more ratios measured the year before: the percentages of cash-flow (CFAT_1) and current debt (RECP_1) on total assets, and the stakeholders' profitability (REAC_1).

Tabla 6. Resultados del análisis discriminante (Construcción)
Table 6. Discriminant analysis results (Building industry)

Variable	Coeficientes de la función discriminante *Discriminant function coefficients*		*F* parcial *Partial F-value*	Valor *p* *P-value*	Coeficientes de estructura *Structure coefficients*
	No estandarizados *Non-standardized*	Estandarizados *Standardized*			
FAC1_EXT	–0,6756	–0,6811	18,1094	0,0001	–0,580
RECP_1	0,0174	0,4315	5,8854	0,0188	0,351
REAC_1	0,0092	0,5080	8,0288	0,0066	0,319
CCA	–1,9514	–0,5569	8,4460	0,0054	–0,099
CFAT_1	–6,9172	–0,5245	8,3399	0,0057	–0,450
Constante *Constant*	0,4450	–	–	–	–

Como puede observarse en la Tabla, la variable más importante en la función es, de nuevo, el factor externo que resume la actividad económica general y a continuación la proporción de capital circulante y de *cash-flow*, todos con influencia negativa en la variable dependiente. Los otros dos ratios tienen coeficientes positivos, lo que implica una relación directa con la puntuación discriminante.

As it can be observed in the Table, the most important variable in the function is again the external factor summarizing the general economic activity, and next the proportion of working capital and cash-flow, all with a negative influence on the dependent variable. The other two ratios have positive coefficients, which involve a direct relationship with the discriminant score.

Tabla 7. Resultados del análisis discriminante (Servicios)
Table 7. Discriminant analysis results (Service industries)

Variable	Coeficientes de la función discriminante *Discriminant function coefficients*		*F* parcial *Partial F-value*	Valor *p* *P-value*	Coeficientes de estructura *Structure coefficients*
	No estandarizados *Non-standardized*	Estandarizados *Standardized*			
CFAT	–2,6404	–0,6152	27,9982	0,0000	–0,427
FAC2_EXT	0,3938	0,6659	33,3676	0,0000	0,409
CFAT_1	–3,6436	–0,4610	14,4806	0,0002	–0,344
EQUI_2	0,0117	0,5367	19,6864	0,0000	0,259
VIDA	0,0050	0,4854	15,7598	0,0001	0,250
RECP	0,0082	0,2710	4,5569	0,0346	0,171
Constante *Constant*	–1,0793	–	–	–	–

Por último, los resultados respecto al sector servicios se recogen en la Tabla 7. La función discriminante está definida por seis variables que, según el valor *p* asociado a cada una, son significativas. La información externa vuelve a ser importante en predecir el fracaso empresarial aunque en este sector es el "nivel de precios" el factor significativo. El tiempo desde la constitución de la empresa (VIDA) también entró en el modelo, junto con cuatro ratios financieros, medidos en los tres años del periodo en estudio: el porcentaje de deudas a corto plazo sobre activo total (RECP) referido al último año previo al fracaso, el ratio de *cash-flow* sobre activo total tanto este año (CFAT) como el anterior (CFAT_1), y la proporción de fondos propios y fondos ajenos a largo plazo que financian el activo fijo el tercer año previo al fracaso (EQUI_2). Según los coeficientes en la tabla, sólo el ratio de *cash-flow* tiene una influencia negativa en la variable dependiente, mientras que los coeficientes correspondientes al resto de variables tienen signo positivo. Dicho ratio y el factor externo que describe al nivel de precios son las variables más importantes en la función discriminante, de acuerdo con los coeficientes estandarizados y de estructura.

Finally, results regarding the service industries are shown in Table 7. The discriminant function is defined by the six variables that, according to the *p*-value associated to each one, are significant. External information is again important in predicting business failure, although in this industry the 'level price' (FAC2_EXT) is the significant factor. The time from the firms incorporation date (VIDA) also entered the model, as well as four financial ratios, measured in the three years of our study period: the percentage of current liabilities on total assets (RECP) referred to the last year before failure, the ratio of cash-flow on total assets both that year (CFAT) and the previous one (CFAT_1), and the proportion of equity and fixed liabilities that finance fixed assets the third year before failure (EQUI_2). According to the coefficients in the table, only the ratio of cash-flow has a negative influence on the dependent variable, whereas the coefficients corresponding to the rest of variables have a positive sign. That ratio and the external factor which describes the level price are the most important variables in the discriminant function, according to both the standardized and structure coefficients.

A la vista de los resultados recogidos en las Tablas 5 a 7, pueden extraerse algunas conclusiones respecto al fracaso empresarial en los tres sectores:

– La información externa es importante en la explicación y predicción del fracaso empresarial, dado que uno de los dos factores que resumen esta clase de información ha resultado significativo en cada sector. Mientras que la actividad económica general tiene influencia en los sectores de industria y construcción, el factor que mide el nivel de precios es el significativo para las empresas que operan en el sector servicios, lo que es lógico, dada la clase de actividad que desarrollan.
– La información no financiera también es significativa en el fracaso empresarial, al menos en dos de los tres sectores considerados. En la industria, es la variable que distingue entre Sociedades Anónimas y de Responsabilidad Limitada la que entra en el modelo. Por otro lado, la vida de las empresas en el sector servicios tiene un efecto significativo sobre la variable dependiente en esta submuestra.
– Respecto a la información financiera, varios ratios han resultados significativos en los modelos elaborados para cada sector. La proporción de *cash-flow* sobre activo total en uno de los dos años anteriores al momento del fracaso forma parte de la función discriminante en las tres submuestras, lo que indica la importancia de generar recursos internamente a efectos de evitar el fracaso empresarial.
– Además de estos aspectos comunes, también hay algunas diferencias entre los distintos sectores. La función discriminante en la industria se completa con un ratio que refleja la importancia de los fondos propios. En el sector de construcción, hay otros tres ratios significativos

By observing the results in Tables 5 to 7, some conclusions about business failure in the three industries can be drawn:

– External information is important in explaining and predicting business failure, since one of the two factors summarizing this kind information have become significant in each industry. Whereas the general economic activity has influence on both the manufacturing and building industries, the factor measuring the price level is the significant one for the firms operating in the service industries, which is logical, provided the kind of activities they develop.
– Non-financial information is also significant in business failure, at least in two out of the three considered industries. In the manufacturing one, it is the variable distinguishing between public and private limited companies which enters the prediction model. On the other hand, the age of the firms in the service industries has a significant effect on the dependent variable in this subsample.
– Regarding financial information, several ratios have become significant in the developed models in each industry. The proportion of cash-flow on total assets for one of the two last years before failure is part of the discriminant function in the three subsamples, which shows the importance of generating internally resources in order to avoid business failure.
– As well as these common issues, there are some differences among the various industries. The discriminant function in the manufacturing industry is completed with a ratio reflecting the importance of equity. In the building one, there are three more significant ratios which measure the stakeholders' profitability and current issues of the firm, whereas

que miden la rentabilidad de accionistas y aspectos de circulante de la empresa, mientras que en los servicios, además del pasivo circulante, también es significativo el nivel de solvencia.

Una vez estimados los modelos predictivos, las empresas en cada submuestra sectorial fueron clasificadas como fracasadas o no fracasadas a partir de la información proporcionada por las variables significativas en cada función discriminante. Los resultados de clasificación se recogen en la Tabla 8. Como puede observarse, son bastante similares en los tres sectores, si bien los mejores resultados se obtienen en la construcción, tanto en cada grupo de empresas como en la submuestra total.

En los tres sectores, se alcanzan buenos resultados de clasificación, con un alto porcentaje de empresas sanas correctamente clasificadas, en especial en el sector de la construcción, donde la tasa de aciertos es del 100%. Por lo que se refiere a las empresas fracasadas, los porcentajes de clasificación correcta son bastante altos, si tenemos en cuenta la composición muestral. Dado que el porcentaje de empresas fracasadas en las diferentes submuestras es bastante bajo en comparación con el de empresas no fracasadas, es normal que las tasas de aciertos para el primer grupo no sean tan altas. En todo caso, todos los porcentajes son superiores al 50%, llegando incluso a más del 80%, como puede observarse respecto al sector de la construcción.

in the service industries, as well as current liabilities, solvency issues are also significant.

Once the prediction models were developed, firms in each industry subsample were classified as either failed or non-failed using the information provided by the significant variables in each discriminant function. The classification results are shown in Table 8. As it can be observed, they are quite similar in the three industries, but the best results are achieved in the building one, both in each firms group and in the total subsample.

In the three industries, good classification results are achieved, with a very high percentage of well classified non-failed firms, especially in the building industry, where the hit rate is 100%. Regarding the failed firms, the correct classification percentages are quite high, if we take into account the sample composition. Provided the percentage of failed firms in the subsamples is quite low in comparison to the non-failed firms one, it is normal that the correct classification results among the former group are not so good. In any case, all the rates are higher than 50%, even achieving more than 80%, as it can be observed regarding the building industry.

Tabla 8. Resultados de clasificación / *Table 8. Classification results*

Empresas / *Firms*	SECTOR / INDUSTRY		
	Industria / *Manufacturing*	Construcción / *Building*	Servicios / *Service*
Fracasadas / *Failed*	59,09%	83,33%	70,59%
No fracasadas / *Non-failed*	94,92%	100%	95,81%
Total	85,19%	97,85%	93,97%

Por consiguiente, los ratios financieros, junto con información no financiera tanto relativa a la empresa como externa a la misma, son útiles como variables para predecir el fracaso empresarial. La inclusión de estos tipos de información en los modelos de predicción del fracaso mediante la aplicación del análisis discriminante lleva a buenos resultados de clasificación en los tres sectores analizados, mostrando que este método predictivo es adecuado en el campo del fracaso empresarial.

4. CONCLUSIONES

Nuestro estudio se ha centrado en el análisis del fracaso empresarial por sectores, con el fin de determinar los factores diferenciadores del fracaso en cada sector. Para ello, hemos aplicado uno de los métodos estadísticos más utilizados en este campo, el análisis discriminante, que ha mostrado su capacidad para predecir el fracaso en muestras correspondientes a diferentes periodos y países. Debido a la robustez de la técnica, permite obtener buenos resultados de clasificación, incluso cuando no se cumplen las hipótesis exigidas para su aplicación.

Para seleccionar la muestra de estudio, nos hemos centrado en las empresas con domicilio social en Castilla y León cuya información estaba disponible en la base de datos SABI utilizada para recoger los datos, correspondientes tanto a variables financieras como no financieras relativas a la empresa, además de información externa. La muestra seleccionada aplicando un procedimiento mixto aleatorio y no aleatorio fue dividida en tres sectores principales: industria, construcción y servicios.

Therefore, financial ratios together with non-financial both firm-related and external information are very useful as variables in order to predict business failure. The inclusion of all these kinds of information in prediction models by the application of a discriminant analysis leads to good classification results in the three analysed industries, showing that this predictive method is appropriate in business failure field.

4. CONCLUSIONS

Our study has focused on a cross-industry analysis of business failure, with the aim of identifying the differential failure factors in each industry. In order to do that, one of the main statistical methods in this field has been applied: discriminant analysis, which has shown its ability to predict business failure in samples corresponding to different periods and countries. Because of the robustness of this technique, it allows obtaining good classification results, even though the required hypotheses for its application are not achieved.

To select the study sample, we focused on firms with head offices in the region of Castile and León (Spain) whose information was available in the database SABI which was used to collect the data referring to both financial and non-financial firm-related variables, together with external information. The selected sample, applying a combined random and non-random procedure, was divided into three main industries: manufacturing, building and service.

Antes de obtener los modelos discriminantes para predecir el fracaso empresarial en cada uno de esos sectores, aplicamos previamente un ACP, con el fin de reducir el número de ratios financieros y la información externa seleccionada como variables potencialmente explicativas para nuestro estudio.

Para resumir las variables externas, en los tres sectores se obtuvieron dos factores, definidos como "actividad económica general" y "nivel de precios". Respecto a la información financiera, algunos factores extraídos fueron comunes a todos los sectores: estructura de recursos, capacidad para devolver deudas, liquidez y circulante. Pero también definimos otros factores específicos de cada sector, como la rentabilidad económica en el sector de la construcción, el margen de beneficios en el de servicios o los fondos propios, tanto en la industria como en los servicios.

Por lo que se refiere a los resultados de la predicción, tanto las variables externas como las no financieras relativas a la empresa resultaron significativas en las funciones discriminantes para cada sector, lo que confirma la importancia de este tipo de información en la predicción del fracaso empresarial. Los aspectos financieros también resultaron significativos en los tres modelos, en los que el más importante fue la capacidad de la empresa para generar recursos internamente. Otros aspectos significativos para evitar el fracaso fueron los siguientes: los fondos propios, en la industria; la rentabilidad y aspectos de circulante, en la construcción y la solvencia, en el sector servicios.

Toda la información contenida en las funciones discriminantes fue utilizada para clasificar las empresas de cada sector como fracasadas o no fracasadas. Los resultados de clasificación fueron buenos en los tres sectores, aunque los

Before developing the discriminant models to predict business failure in each of those industries, a previous PCA was applied, in order to reduce the number of financial ratios and external information selected as potentially explanatory variables for our study.

To summarize the external variables, two factors defined as 'general economic activity' and 'price level' were obtained in the three industries. Regarding the financial information, some extracted factors were common to all the industries: liability structure, ability to return debts, liquidity and current position. But some other industry-specific factors were defined, such as economic profitability in the building industry, profit margin in the service ones or equity in both the manufacturing and the service industries.

With regard to the prediction results, both external and non-financial firm-related variables became significant in the discriminant functions for each industry, which confirms the importance of this kind of information in predicting business failure. Financial issues were also significant in the three models, being the most important one the firm's ability to generate internally resources. Other significant issues in order to avoid failure are the following: equity, in the manufacturing industry; profitability and current issues in the building one, and solvency, in the service industries.

All the information in the discriminant functions was used to classify the firms in each industry as failed or non-failed. The classification results were good in the three industries, although the highest hit rates were achieved in the building one.

porcentajes más altos se alcanzaron en el de la construcción. En todo caso, los resultados fueron mejores en el grupo de empresas sanas que en el de fracasadas, lo que puede explicarse por la composición de cada submuestra, con un mayor porcentaje de ese tipo de empresas.

In any case, the results were better regarding the non-failed group than the failed one, which can be explained because of each subsample composition, with a higher percentage of that kind of firms.

BIBLIOGRAFÍA / REFERENCES

Altman, E.I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, *XXIII(*4), 589-609.

Altman, E.I. (1984). The success of business failure models: An international survey. *Journal of Banking & Finance*, *8*, 171-198.

Altman, E.I., Marco, G. y Varetto, F. (1994). Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience). *Journal of Banking & Finance*, *18*, 505-529.

Balcaen, S. y Ooghe, H. (2006). 35 years of studies on business failure: An overview of the classical statistical methodologies and their related problems. *The British Accounting Review*, *38*(1), 63-93.

Beaver, W.H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, Supplement to vol. 4: Empirical research in accounting: Selected studies, 71-111.

Blum, M. (1974). Failing company discriminant analysis. *Journal of Accounting Research*, Spring, 1-25.

Charitou, A., Neophytou, E. y Charalambous, C. (2004). Predicting corporate failure: Empirical evidence for the UK. *European Accounting Review*, *13*(3), 465-497.

Chava, S. y Jarrow, R.A. (2004). Bankruptcy prediction with industry effects. *Review of Finance*, *8*, 537-569.

Cybinski, P. (2001). Description, explanation, prediction – the evolution of bankruptcy studies? *Managerial Finance*, *27*(4), 29-44.

Deakin, E.B. (1972). A discriminant analysis of predictors of business failure. *Journal of Accounting Research*, spring, 161-179.

Dimitras, A.I., Slowinski, R., Susmaga, R. y Zopounidis, C. (1999). Business failure prediction using rough sets. *European Journal of Operational Research*, *114*(2), 263-280.

Dimitras, A.I., Zanakis, S.H. y Zopounidis, C. (1996). A survey of business failures with an emphasis on prediction methods and industrial applications. *European Journal of Operational Research*, *90*(3), 487-513.

Edmister, R.O. (1972). An empirical test of financial ratio analysis for small business failure prediction. *Journal of Financial and Quantitative Analysis*, *7*, march, 1477-1493.

Elam, R. (1975). The effect of lease data on the predictive ability of financial ratios. *The Accounting Review*, *50*(1), 25-43.

Izan, H.Y. (1984). Corporate distress in Australia. *Journal of Banking and Finance*, *8*, 303-320.

Jones, F.L. (1987). Current techniques in bankruptcy prediction. *Journal of Accounting Literature*, *6*, 131-164.

Junta de Andalucía (2004). *Tejido empresarial y factores de éxito. Una aproximación al caso andaluz*. Sevilla: Servicio de Asesoría Técnica y Publicaciones. Consejería de Economía y Hacienda.

Keasey, K. y Watson, R. (1991). Financial distress prediction models: A review of their usefulness. *British Journal of Management*, *2*(2), 89-102.

Laffarga Briones, J., Martín Marín, J.L. y Vázquez Cueto, M.J. (1985). El análisis de la solvencia en las instituciones bancarias: Propuesta de una metodología y aplicaciones a la Banca española. *ESIC-MARKET*, *48*, abril-junio, 51-73.

Laitinen, E.K. (1991). Financial ratios and different failure processes. *Journal of Business, Finance & Accounting*, *18*(5), 649-673.

Laitinen, E.K. y Laitinen, T. (1998). Misclassification in bankruptcy prediction in Finland: Human information processing approach. *Accounting, Auditing & Accountability Journal*, *11*(2), 216-244.

Lang, L.H.P. y Stulz, R.M. (1992). Contagion and competitive intra-industry effects of bankruptcy announcements. *Journal of Financial Economics*, *32*, 45-60.

Lincoln, M. (1984). An empirical study of the usefulness of accounting ratios to describe levels of insolvency risk. *Journal of Banking and Finance*, *8*, 321-340.

McDonald, B. y Morris, M.H. (1984). The statistical validity of the ratio method in financial analysis: An empirical examination. *Journal of Business Finance & Accounting*, *11*(1), 89-96.

Mora Enguídanos, A. (1994). Los modelos de predicción del fracaso empresarial: una aplicación empírica del logit. *Revista Española de Financiación y Contabilidad*, *XXIII*(78), 203-233.

Odom, M.D. y Sharda, R. (1992). A neural network model for bankruptcy prediction. En, R.R. Trippi y E. Turban (Eds.), *Neural networks in finance and investing. Using artificial intelligence to improve real-world performance* (pp. 177-185). Cambridge: Probus Publishing Company.

Ohlson, J.A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, *18*(1), 109-131.

Palepu, K.G. (1986). Predicting takeover targets. A methodological and empirical analysis. *Journal of Accounting and Economics*, *8*, 3-35.

Peel, M.J., Peel, D.A. y Pope, P.F. (1986). Predicting corporate failure - some results for the UK corporate sector. *Omega, International Journal of Management Science*, *14*(1), 5-12.

Pina Martínez, V. (1989). La información contable en la predicción de la crisis bancaria 1977-1985. *Revista Española de Financiación y Contabilidad*, *XVIII*(58), 309-338.

Platt, H.D. y Platt, M.B. (1990). Development of a class of stable predictive variables: The case of bankruptcy prediction. *Journal of Business Finance & Accounting*, *17*(1), 31-51.

Platt, H.D. y Platt, M.B. (1991). A note on the use of industry-relative ratios in bankruptcy prediction. *Journal of Banking and Finance*, *15*(6), 1183-1194.

Ravi Kumar, P. y Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review. *European Journal of Operational Research*, *180*(1), 1-28.

Román Martínez, I., de la Torre Martínez, J.M. y Zafra Gómez, J.L. (2001). Análisis sectorial de la predicción del riesgo de insolvencia: Un estudio empírico. *XI Congreso AECA "Empresa, Euro y Nueva Economía"*. Madrid, 26-28 de septiembre.

Smith, M. y Liou, D.K. (2007). Industrial sector and financial distress. *Managerial Auditing Journal*, *22*(4), 376–391.

Taffler, R.J. (1982). Forecasting company failure in the UK using discriminant analysis and financial ratio data. *Journal of the Royal Statistical Society*, Series A – Statistics in Society, *145*(3), 342-358.

Taylor, J.D. (1997/1998). Cross-industry differences in business failure rates: Implications for portfolio management. *Commercial Lending Review*, *13*(1), 36-46.

Theodossiou, P.T. (1991). Alternative models for assessing the financial condition of business in Greece. *Journal of Business, Finance & Accounting*, *18*(5), 697-720.

Zmijewski, M.E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research*, Supplement to vol. *22*, Studies on Current Econometric Issues in Accounting Research, 59-82.

Pecvnia Monográfico 2012, pp. 85-110

DISTRIBUCIÓN DE LAS TRANSFORMACIONES LINEALES DE LOS RESIDUOS MÍNIMOS CUADRADOS *STUDENTIZADOS* INTERNAMENTE / *DISTRIBUTION OF LINEAR TRANSFORMATIONS OF INTERNALLY STUDENTIZED LEAST SQUARES RESIDUALS*

Seppo Pynnönen[1]
sjp@uwasa.fi

University of Vaasa (Finland)

Resumen

Los residuos de regresión por mínimos cuadrados ordinarios tienen una distribución que depende de un parámetro escalar. El término "*Studentización*" se utiliza comúnmente para describir una cantidad U dependiente de un parámetro de escala dividida por una estimación de escala S, de forma que el ratio resultante, U/S, sigue una distribución que no tiene el inconveniente del parámetro de escala desconocido. La *Studentización* externa hace referencia a un ratio en que el numerador y el denominador son independientes, mientras que la *Studentización* interna se refiere al ratio en que ambos son dependientes. La ventaja de la *Studentización* interna es que puede utilizarse cualquier estimador de escala común, mientras que en la *Studentización* externa, cada residuo es obtenido por un estimador de escala diferente, con el fin de alcanzar la independencia. Con errores de regresión normales, la distribución conjunta de un conjunto arbitrario (linealmente independiente) de residuos *Studentizados* internamente está bien documentada. Sin embargo, en algunas aplicaciones una combinación lineal de residuos internamente *Studentizados* puede resultar útil. Sus limitaciones han sido bien documentadas, pero la distribución no parece haberse derivado en la literatura. Este trabajo contribuye a la literatura existente, en el sentido de obtener la distribución conjunta de una transformación arbitraria lineal de residuos de regresión por mínimos cuadrados ordinarios internamente *Studentizados* con distribución esférica de error. Todas las principales versiones de los residuos de regresión internamente *Studentizados* que se han utilizado comúnmente en la literatura son casos especiales de la transformación lineal.

Palabras clave: Transformación de Borel de residuos *Studentizados*; Residuos normados; Distribución esférica; Distribución elíptica.

[1] Department of Mathematics and Statistics, University of Vaasa, P.O.Box 700, FI-65101, Vaasa, Finland.

Abstract

Ordinary least squares regression residuals have a distribution that is dependent on a scale parameter. The term 'Studentization' is commonly used to describe a scale parameter dependent quantity U divided by a scale estimate S such that the resulting ratio, U/S, has a distribution that is free of from the nuisance unknown scale parameter. *External* Studentization refers to a ratio in which the nominator and denominator are independent, while *internal* Studentization refers to a ratio in which these are dependent. The advantage of the internal Studentization is that typically one can use a single common scale estimator, while in the external Studentization every single residual is scaled by different scale estimator to gain the independence. With normal regression errors the joint distribution of an arbitrary (linearly independent) subset of internally Studentized residuals is well documented. However, in some applications a linear combination of internally Studentized residuals may be useful. The boundedness of them is well documented, but the distribution seems not be derived in the literature. This paper contributes to the existing literature by deriving the joint distribution of an arbitrary linear transformation of internally Studentized residuals from ordinary least squares regression with spherical error distribution. All major versions of commonly utilized internally Studentized regression residuals in literature are obtained as special cases of the linear transformation.

Keywords: Borel transformation of Studentized residuals; Normed residuals; Spherical distribution; Elliptical distribution.

1. INTRODUCCIÓN

Las transformaciones de residuos juegan un papel clave en los diagnósticos de regresión. Por consiguiente, las propiedades de las distribuciones de los residuos subyacentes son inherentes a las inferencias estadísticas eficientes sobre la calidad del modelo. Los residuos *Studentizados* se consideran útiles en el análisis de datos atípicos, en particular (por ejemplo, Chatterjee and Hadi, 1988, Sec. 4.2.1). Como consecuencia, hay un continuo interés en estudiar las propiedades estadísticas de diferentes formas de residuos *Studentizados*; entre otros podemos citar como ejemplos Abrahamse y Koerts (1971), Beckman y Trussel (1974), Chatterjee y Hadi (1988), Díaz-García y Gutiérrez-Jáimez (2006, 2007), Pynnönen (2012).

1. INTRODUCTION

Transformations of residual play a key role in regression diagnostics. Therefore, the distributional properties of the underlying residuals are eminent to efficient statistical inferences about the quality of the model. Studentized residual are considered useful in analysis of outliers, in particular (e.g., Chatterjee and Hadi, 1988, Sec. 4.2.1). As a consequence there is a continuous interest to study the statistical properties of different forms of Studentized residuals; examples are among others Abrahamse and Koerts (1971), Beckman and Trussel (1974), Chatterjee and Hadi (1988), Díaz-García and Gutiérrez-Jáimez (2006, 2007), Pynnönen (2012).

Este trabajo continúa esta línea de investigación y deduce, explícitamente, la distribución de una transformación lineal arbitraria de residuos de una regresión múltiple con errores elípticos que han sido interna y externamente *Studentizados*. En principio, la mayoría de resultados presentados en este trabajo pueden obtenerse como casos especiales del de regresión multivariante abordado en Pynnönen (2012). Sin embargo, muchos de esos resultados no son tan obvios y por ello se incluyen dentro de la estructura de regresión múltiple.

Para este objetivo, consideremos una modelo de regresión con n observaciones:

$$y = X\beta + u \qquad (1)$$

donde X es una matriz no estocástica de orden $n \times p'$ con rango $p \leq p' < n$, y es un vector n-dimensional de respuestas observables, β es un vector p'-dimensional de parámetros de pendiente, y u es un vector n-dimensional de errores homoscedásticos no observables que siguen una distribución de contorno esférico con matriz de varianzas-covarianzas proporcional a $\sigma_u^2 I_n$, donde $\sigma_u^2 > 0$ es el parámetro de escala de la varianza e I_n es la matriz identidad de orden $n \times n$.

Los residuos por mínimos cuadrados ordinarios (MCO) vienen dados por:

$$\begin{aligned} \hat{u} &= Q_y \\ &= Q_u \end{aligned} \qquad (2)$$

donde

$$Q = I_n - X(X'X)^{-}X' \qquad (3)$$

This paper continues this research and explicitly derives the distribution of an arbitrary linear transformation of internally and externally Studentized residuals from a multiple regression with elliptical errors. In principle most of the results presented in this paper can be obtained as special cases from multivariate regression dealt with in Pynnönen (2012). However, many of the results are not that obvious and motivates us to reconstruct them within the multiple regression framework.

For the purpose, consider a regression model with n observations

$$y = X\beta + u \qquad (1)$$

where X is an $n \times p'$ nonstochastic matrix with rank $p \leq p' < n$, y is an n-vector of observable responses, β is a p'-vector of slope parameters, and u is an n-vector of unobservable homoscedastic errors that follow some spherical contoured distribution with variance-covariance matrix proportional to $\sigma_u^2 I_n$, where $\sigma_u^2 > 0$ is the scalar variance parameter and I_n is the $n \times n$ identity matrix.

The ordinary least squares (OLS) residuals are given by

$$\begin{aligned} \hat{u} &= Q_y \\ &= Q_u \end{aligned} \qquad (2)$$

where

$$Q = I_n - X(X'X)^{-}X' \qquad (3)$$

es una matriz simétrica idempotente de orden $n \times n$ con rango $n - p$ en la que la prima denota la matriz traspuesta y $(X'X)^-$ es la inversa generalizada de $X'X$.

La *Studentización* es un término común utilizado para describir la división de un estadístico dependiente de un parámetro de escala, U, por una estimación de escala S de forma que la distribución del ratio resultante U/S ya no depende de los parámetros de escala (véase, por ejemplo, Margolin 1977). Por lo general, U y S se obtienen de los mismos datos, en cuyo caso el ratio U/S se denomina internamente *Studentizado* si U y S son dependientes y externamente *Studentizado* si son independientes (véase Cook y Weisberg, 1982: 18).

En la regresión por mínimos cuadrados los residuos internamente *Studentizados* se definen como:

$$\tilde{r}_i = \frac{\hat{u}_i}{s\sqrt{q_{ii}}} \qquad (4)$$

donde $\hat{u}_i$ es el componente *i*-ésimo del vector $\hat{u}$, $s = \sqrt{\hat{u}'\hat{u}/(n-p)}$, y q_{ii} es el elemento *i*-ésimo de la diagonal principal de la matriz Q. Incluso con errores distribuidos normalmente, si bien $\tilde{r}_i$ es el conocido ratio de una variable aleatoria normalmente distribuida y la raíz cuadrada de una variable aleatoria *chi*-cuadrado, el resultado final no es una variable aleatoria distribuida según una *t*. La razón es que el numerador y el denominador no son independientes debido al hecho de que $\hat{u}_i^2 < \hat{u}'\hat{u}$ para todo $i = 1,\ldots, n$. Además, puesto que $q_{ii} \geq 1 + 1/n$ (p.e., Cook y Weisberg, 1982: 12), a diferencia de una variable aleatoria que sigue una distribución t y que supone todos los valores reales, se verifica que $\tilde{r}_i^2 \leq n - p$.

is an $n \times n$ symmetric idempotent matrix with rank $n - p$ in which the prime denotes the matrix transposition and $(X'X)^-$ is a generalized inverse of $X'X$.

Studentization is a common term used to describe division of a scale parameter dependent statistic, say U, by a scale estimate S such that the distribution if the resulting ratio U/S is free from the nuisance scale parameters (see e.g. Margolin, 1977). Typically U and S are derived from the same data in which case the ratio U/S is called internally Studentized if U and S are dependent and externally Studentized if they are independent (see Cook and Weisberg, 1982: 18).

In the least squares regression the internally Studentized residuals are defined as

$$\tilde{r}_i = \frac{\hat{u}_i}{s\sqrt{q_{ii}}} \qquad (4)$$

where $\hat{u}_i$ is the *i*th component of the vector $\hat{u}$, $s = \sqrt{\hat{u}'\hat{u}/(n-p)}$, and q_{ii} is the *i*th diagonal element of the matrix Q. Even with normally distributed errors, although $\tilde{r}_i$ is the familiar ratio of a normally distributed random variable and a square root of a scaled *chi*-square random variable, the end result is not a *t*-distributed random variable. The reason is that the nominator and the denominator are not independent due to the fact that $\hat{u}_i^2 < \hat{u}'\hat{u}$ for all $i = 1,\ldots, n$. Furthermore, because $q_{ii} \geq 1 + 1/n$ (e.g., Cook and Weisberg, 1982: 12), it follows that unlike a *t*-distributed random variable that assumes all real values, $\tilde{r}_i^2 \leq n - p$.

Stefansky (1972), Ellenberg (1973) y Díaz-García y Gutiérrez-Jáimez (2007) obtuvieron la distribución conjunta de un conjunto arbitrario (no singular) de los residuos internamente *Studentizados* definidos en (4). Beckman and Trussell (1974) obtuvieron la distribución del estadístico t para un valor arbitrario $\tilde{r}_i$. Pynnönen (2012) obtuvo la distribución de los resultados referidos para una transformación lineal arbitraria de residuos internamente *Studentizados* de una regresión multivariante con errores elípticos. El presente trabajo obtiene las distribuciones conjuntas de una transformación arbitraria lineal no singular de residuos interna y externamente *Studentizados*, de las que son casos especiales las distribuciones conjuntas de un conjunto de residuos interna y externamente *Studentizados*. El trabajo también muestra que las inferencias relativas a las transformaciones lineales de residuos (*Studentizados*) son, de hecho, un problema de variable omitida en regresión. Como hemos señalado, la mayoría de estos resultados son casos especiales de los presentados en Pynnönen (2012). Sin embargo, debido a que la regresión univariante es un procedimiento de modelización estadística mucho más popular en análisis empíricos aplicados y porque los resultados de la regresión multivariante no siempre pueden trasladarse de forma sencilla al caso univariante, creemos que la discusión de los resultados en el contexto de la regresión múltiple está suficientemente justificada.

2. PRINCIPALES RESULTADOS

Partimos de la siguiente definición para la familia de las distribuciones de contorno esféricos (p.e. Kariya y Eaton, 1977),

Definición 1 Un vector aleatorio u de orden $n \times 1$ sigue una distribución de contorno esférico si

$$Hu \overset{d}{=} u \tag{5}$$

Stefansky (1972), Ellenberg (1973), and Díaz-García and Gutiérrez-Jáimez (2007) derived the joint distribution of an arbitrary (nonsingular) subset of the internally Studentized residuals defined in (4). Beckman and Trussell (1974) derive the distribution for a t-statistic for an arbitrary single $\tilde{r}_i$. Pynnönen (2012) derived the distribution of related results of an arbitrary linear transformation of internally Studentized residuals of multivariate regression with elliptical errors. The present paper derives the joint distributions of an arbitrary non-singular linear transformation of internally and externally Studentized residuals of which the joint distributions of a subset of internally and externally Studentized residuals are special cases. The paper also shows that inference regarding linear transformations of (Studentized) residuals is factually an omitted variable problem in regression. As noted above, most of these results are factually special cases of those given in Pynnönen (2012). However, because univariate regression is vastly more popular statistical modeling devise in applied empirical analyses and because the results from the multivariate regression may not always be straightforward to translate to the univariate counterpart, we think that discussion of the result in the context of the multiple regression is well warranted.

2. MAIN RESULTS

We use the following definition for the family of spherical contoured distributions (e.g. Kariya and Eaton, 1977).

Definition 1 A $n \times 1$ random vector u is spherical contoured distributed if

$$Hu \overset{d}{=} u \tag{5}$$

donde $H \in O(n)$ siendo

$$O(n) = \{H(n \times n) matrix : H'H = I\}$$

el grupo de matrices ortogonales de orden $n \times n$ y "$\overset{d}{=}$" significa que los dos vectores aleatorios siguen la misma distribución. Denotamos la familia de distribuciones esféricas por $S(n)$, y denotamos $u \in S(n)$ para indicar que la distribución de la variable aleatoria u pertenece a la familia de distribuciones esféricas.

Una propiedad importante de las distribuciones esféricas (y más general, de las distribuciones elípticas) es que todas las distribuciones marginales son esféricas (elípticas) (p.e. Muirhead, 1982: 34).

Para el operador "$\overset{d}{=}$" utilizamos el siguiente resultado. Si las variables aleatorias x e y siguen la misma distribución, es decir, $x \overset{d}{=} y$, entonces para cualquier función de Borel f, se verifica que $f(x) \overset{d}{=} f(y)$. Para su comprobación, veáse Anderson y Fang (1990).

Sea

$$C_n = \{x \in \Re^n : \| x \| = 1\} \qquad (6)$$

la esfera unitaria en el espacio euclídeo n-dimensional $\Re^n$, donde $\| x \| = \sqrt{x'x}$ es la norma euclídea. Entonces, la distribución uniforme es la única distribución en C_n que es invariante bajo $O(n)$ (p.e., Kariya y Eaton, 1977).

where $H \in O(n)$ with

$$O(n) = \{H(n \times n) matrix : H'H = I\}$$

the group of $n \times n$ orthogonal matrices, and "$\overset{d}{=}$" means that the two random vectors have the same distributions. We denote the family of spherical distributions by $S(n)$, and denote $u \in S(n)$ to mean that the distribution of the random variable u belongs to the family of spherical distributions.

An important property of spherical distributions (and more generally of elliptical distributions) is that all marginal distributions are spherical (elliptical) (e.g., Muirhead, 1982, p. 34).

For the operator "$\overset{d}{=}$" we utilize the following important result. If random variables x and y have the same distribution, i.e., $x \overset{d}{=} y$ then for any Borel function f, $f(x) \overset{d}{=} f(y)$. For a proof, see Anderson and Fang (1990).

Let

$$C_n = \{x \in \Re^n : \| x \| = 1\} \qquad (6)$$

be the unit sphere in the n-dimensional Euclidian space $\Re^n$, where $\| x \| = \sqrt{x'x}$ is the Euclidian norm. Then the uniform distribution is the unique distribution on C_n that is invariant under $O(n)$ (e.g., Kariya and Eaton, 1977).

Lema 1 Para cualquier $u \in S(n)$ con $P(u=0) = 0$, la variable aleatoria normada

$$\frac{u}{\|u\|} \overset{d}{=} u \tag{7}$$

donde U es un vector aleatorio que sigue una distribución uniforme en la esfera C_n.

Comprobación. Véase Kariya y Eaton (1977, teorema 2.1).

Proposición 1 Una importancia particular del lema 1 es que, debido a que $N(0, I_n)$, la distribución conjunta de n variables aleatorias normales estándar independientes, es un miembro de $S(n)$, la unicidad de la distribución uniforme en la esfera C_n implica que para cualquier $u \in S(n)$, se verifica:

$$\frac{u}{\|u\|} \overset{d}{=} \frac{z}{\|z\|} \tag{8}$$

donde $z \sim N(0, I_n)$. Esto es, podemos estudiar las propiedades de la distribución de varias tranformaciones de los vectores aleatorios normados $u/\|u\|$ con $u \in S(n)$ en términos del ratio $z/\|z\|$, donde $z \sim N(0, I_n)$.

Definición 2 Los residuos normalizados

$$r = \frac{\hat{u}}{\|\hat{u}\|} \tag{9}$$

se denominan residuos normalizados internamente (*INRs: internally normalized residuals*), donde $\hat{u}$, definido en la ecuación (2), es un vector de residuos por mínimos cuadrados de una regresión con errores que siguen una distribución esférica.

Lemma 1 For any $u \in S(n)$ with $P(u = 0) = 0$, the normed random variable

$$\frac{u}{\|u\|} \overset{d}{=} u \tag{7}$$

where U is a uniformly distributed random vector on the sphere C_n.

Proof. See Kariya and Eaton (1977, Theorem 2.1).

Remark 1 A particular importance of Lemma 1 is that because $N(0, I_n)$, the joint distribution of n independent standard normal random variables, is a member of $S(n)$, the uniqueness of the uniform distribution on the sphere C_n implies that for any $u \in S(n)$

$$\frac{u}{\|u\|} \overset{d}{=} \frac{z}{\|z\|} \tag{8}$$

where $z \sim N(0, I_n)$. That is, we can study the distributional properties of various transformations of the normed random vector $u/\|u\|$ with $u \in S(n)$ in terms of the ratio $z/\|z\|$, where $z \sim N(0, I_n)$.

Definition 2 The normalized residuals

$$r = \frac{\hat{u}}{\|\hat{u}\|} \tag{9}$$

are called internally normalized residuals (*INRs*), where $\hat{u}$ defined in equation (2) is a vector of least squares residuals from a regression with spherical distributed errors.

Los siguientes lemas son versiones univariantes de Pynnönen (2012).

Lema 2 (Pynnönen 2012, Lema 1) Bajo el supuesto de que los errores $u \in S(n)$ con $P(u = 0) = 0$ en la regresión $y = X\beta + u$ definida en la ecuación (1)

$$r = \frac{\hat{u}}{\| \hat{u} \|} \overset{d}{=} \frac{v}{\| v \|} \tag{10}$$

donde

$$v = Qz \tag{11}$$

con $z \sim N(0, I_n)$ y $Q = I_n - X(X'X)^{-}X'$ está definido en (3).

Demostración: Utilizando la ecuación (2) y considerando que para una cantidad fija Q la transformación $Q / \| Q \|$ es continua (a.e.) en $\Re^n$ y, por tanto, una función de Borel, tenemos

The following Lemmas are univariate versions of Pynnönen (2012).

Lemma 2 (Pynnönen (2012), Lemma 1) Under the assumption that the errors $u \in S(n)$ with $P(u = 0) = 0$ in the regression $y = X\beta + u$ defined in equation (1),

$$r = \frac{\hat{u}}{\| \hat{u} \|} \overset{d}{=} \frac{v}{\| v \|} \tag{10}$$

where

$$v = Qz \tag{11}$$

with $z \sim N(0, I_n)$ and $Q = I_n - X(X'X)^{-}X'$ is defined in (3).

Proof: Using equation (2) and noting that for fixed Q the transformation $Q / \| Q \|$ continuous (a.e.) on $\Re^n$ and hence a Borel function, we have

$$\frac{\hat{u}}{\| \hat{u} \|} = \frac{Qu}{\| Qu \|} = \frac{Qu / \| u \|}{\| (Qu / \| u \|) \|} \overset{d}{=} \frac{Qz / \| z \|}{\| (Qz / \| z \|) \|} = \frac{Qz}{\| Qz \|} = \frac{v}{\| v \|} \tag{12}$$

donde $z \sim N(0, I_n)$ y $v = Qz$. Esto completa la comprobación del lema.

Proposición 2 El lema 2 muestra explícitamente que *r* definido en la ecuación (9) no depende del parámetro de escala σ_u^2 y, por consiguiente, es una transformación *Studentizada* de los residuos.

Proposición 3 El resultado en el lema 2 es fundamental en el sentido de que todos los problemas inferenciales en la regresión $y = X\beta + u$ que podrían resolverse en términos de funciones (de Borel) de residuos internamente normalizados, aunque no dependiendo de la normalidad de $u \in S(n)$, comparten exactamente las

where $z \sim N(0, I_n)$ and $v = Qz$. This completes the proof of the lemma.

Remark 2 Lemma 2 shows explicitly that *r* defined in (9) does not depend on the scale parameter σ_u^2, and thus indeed is a Studentizing transformation of the residuals.

Remark 3 The result in Lemma 2 is fundamental in the sense that all inference problems in regression $y = X\beta + u$ that can be worked out in terms of (Borel) functions of the internally normalized residuals, although not depending on the normality of $u \in S(n)$,

mismas propiedades estadísticas que si u estuviera normalmente distribuido. Ejemplos son las distribuciones nulas de los estadísticos estándar t y F en regresión (véase también Chmielewski, 1981). Esto se puede ver fácilmente considerando la hipótesis lineal general de la que pueden obtenerse como casos especiales los tests para un coeficiente simple relacionados con el estadístico t. Supongamos, por simplicidad, que X es de rango completo. Expresando la hipótesis lineal general como

$$H_0 : R\beta = q \tag{13}$$

donde R es una matriz conocida de orden $k \times p$ de rango k y q es un vector conocido de orden $k \times 1$. El estadístico F es de la forma (p.e. Johnston y DiNardo, 1997: 97)

$$F = \frac{(\hat{u}_R'\hat{u}_R - \hat{u}'\hat{u})/k}{\hat{u}'\hat{u}/(n-p)} \tag{14}$$

donde $\hat{u}_R$ son los residuos de mínimos cuadrados bajo las restricciones de la hipótesis fijada en (13). Bajo estas restricciones, el estadístico F puede escribirse como

$$F = \frac{u'Q_R u/k}{u'Qu/(n-p)} \tag{15}$$

donde $u \in S(n)$, Q_R es una matriz simétrica idempotente de rango k, Q es la matriz idempotente simétrica de rango $(n-p)$ definida en la ecuación (3) y $Q_R Q = 0$. Por tanto, utilizando el mismo método que en la ecuación (12), obtenemos

$$F \overset{d}{=} \frac{z'Q_R z/k}{z'Qz/(n-p)} \tag{16}$$

share exactly the same statistical properties as if u were normally distributed. Examples are the null distributions of the standard t and F-statistics in regression (see also Chmielewski, 1981). This is easily seen by considering the general linear hypothesis from which single coefficient tests with related t-statistics can be obtained as special cases. Assume for simplicity that X is of full rank. Write the general linear hypothesis as

$$H_0 : R\beta = q \tag{13}$$

where R is a $k \times p$ known matrix of rank k and q is a $k \times 1$ known vector. The F-statistic is of the form (e.g. Johnston and DiNardo, 1997: 97)

$$F = \frac{(\hat{u}_R'\hat{u}_R - \hat{u}'\hat{u})/k}{\hat{u}'\hat{u}/(n-p)} \tag{14}$$

where $\hat{u}_R$ are the least squares residuals under the restrictions of the hypothesis in (13). Under these restrictions the F-statistic can be written in the form

$$F = \frac{u'Q_R u/k}{u'Qu/(n-p)} \tag{15}$$

where $u \in S(n)$, Q_R is a symmetric idempotent matrix of rank k, Q is the symmetric rank $(n-p)$ idempotent matrix given in equation (3), and $Q_R Q = 0$. Thus, using the same method as in equation (12), we obtain

$$F \overset{d}{=} \frac{z'Q_R z/k}{z'Qz/(n-p)} \tag{16}$$

donde $z \sim N(0, I_n)$, que implica que el lado derecho de la ecuación (16) sigue una distribución F con k y $n - p$ grados de libertad. Por tanto, con los errores siguiendo una distribución esférica general, la distribución nula del estadístico F en (14) es $F(k; n - p)$.

El interés de este trabajo se centra en las transformaciones lineales de los residuos internamente normados provenientes de regresiones con errores que siguen una distribución esférica general. Por el lema 2, si M es una matriz $m \times n$, entonces

$$Mr \overset{d}{=} Mv / \| v \| \tag{17}$$

donde $v = Qz$ con $z \sim N(0, I_n)$. Esto es, de nuevo los resultados de la distribución de Mr pueden derivarse en términos de variables aleatorias independientes distribuidas según una normal de media cero y varianza unitaria.

Lema 3 (Pynnönen 2012, Lema 2) Sea M una matriz de orden $m \times n$. Entonces

$$Mr \overset{d}{=} \tilde{M}\tilde{z} / \| \tilde{z} \| \tag{18}$$

donde $\tilde{z} \sim N(0, I_{n-p})$ y

$$\tilde{M} = MH_{n-p} \tag{19}$$

es una matriz de orden $m \times (n - p)$ en la que H_{n-p} es una matriz $n \times (n - p)$ que contiene los vectores propios de los autovalores unitarios de la matriz simétrica idempotente Q, tal que

$$\tilde{M}\tilde{M}' = MQM' \tag{20}$$

Demostración: La comprobación es análoga a la de Pynnönen (2012), Lema 2.

where $z \sim N(0, I_n)$, which implies that the right hand side of (16) has the F-distribution with k and $n - p$ degrees of freedom. Thus, with the general spherical distributed errors, the null distribution of the F-statistic in (14) is $F(k; n - p)$.

The interest of this paper is in the linear transformations of the internally normed residuals stemming from regressions with general spherical distributed errors. By Lemma 2 if M is an $m \times n$ matrix, then

$$Mr \overset{d}{=} Mv / \| v \| \tag{17}$$

where $v = Qz$ with $z \sim N(0, I_n)$. That is, again the distribution results of Mr can be derived in terms of independent zero mean and unit variance normal distributed random variables.

Lemma 3 (Pynnönen (2012), Lemma 2) Let M be an $m \times n$ matrix. Then

$$Mr \overset{d}{=} \tilde{M}\tilde{z} / \| \tilde{z} \| \tag{18}$$

where $\tilde{z} \sim N(0, I_{n-p})$ and

$$\tilde{M} = MH_{n-p} \tag{19}$$

is an $m \times (n - p)$ matrix in which H_{n-p} is an $n \times (n - p)$ matrix containing the eigenvectors of the unit eigenvalues of the symmetric idempotent matrix Q, such that

$$\tilde{M}\tilde{M}' = MQM' \tag{20}$$

Proof: Proof is analogous to that of Pynnönen (2012), Lemma 2.

Podríamos proceder con los resultados de este lema para obtener todos los resultados que siguen. Sin embargo, elegimos utilizar los resultados de este lema más adelante y seguir en términos de la matriz v definida en la ecuación (11). El motivo de esta elección es que algunos de los resultados intermedios que siguen pueden ser de interés en el análisis de los residuos brutos (es decir, no normalizados) de las regresiones con errores normales.

Lema 4 (Pynnönen 2012, Lemma 4) Como en la ecuación (11), sea $v = Qz$, donde $z \sim N(0, I_n)$ y sea M una matriz (no estocástica) de orden $m \times n$ con $r = rank(M) < n - p$, entonces

$$V_M = v'v - v'M'(MQM')^{-}Mv \qquad (21)$$

y

$$U_M = v'M'(MQM')^{-}Mv \qquad (22)$$

se distribuyen independientemente como $\chi^2(n-p-r)$ y $\chi^2(r)$, respectivamente.

Demostración: La comprobación es análoga a la de Pynnönen (2012), Lema 4.

Corolario 1

$$\frac{v'M'(MQM')^{-}Mv}{v'v} \leq 1 \qquad (23)$$

Demostración: Como $V_M \geq 0$ en (21), resulta la desigualdad (23), que completa la comprobación.

Corolario 2 Sea M una matriz de orden $m \times n$, entonces

$$r'M'(MQM')^{-}Mr \leq 1 \qquad (24)$$

We could proceed with the results of this lemma to obtain all the results what follow. However, we choose to utilize the results of this lemma only later and proceed in terms of v defined in equation (11). The motivation of this choice is that some of the intermediate distributional results what follow may be of interest in analysis of raw residuals (i.e., non-normalized) from regressions with normal errors.

Lemma 4 (Pynnönen (2012), Lemma 4) As in equation (11), let $v = Qz$, where $z \sim N(0, I_n)$ and let M be an $m \times n$ (nonstochastic) matrix with $r = rank(M) < n - p$, then

$$V_M = v'v - v'M'(MQM')^{-}Mv \qquad (21)$$

and

$$U_M = v'M'(MQM')^{-}Mv \qquad (22)$$

are independently distributed as $\chi^2(n-p-r)$ and $\chi^2(r)$, respectively.

Proof: Proof is parallel to Pynnönen (2012), Lemma 4.

Corollary 1

$$\frac{v'M'(MQM')^{-}Mv}{v'v} \leq 1 \qquad (23)$$

Proof: Because $V_M \geq 0$ in (21), the inequality (23) follows, which completes the proof.

Corollary 2 Let M be an $m \times n$ matrix, then

$$r'M'(MQM')^{-}Mr \leq 1 \qquad (24)$$

Demostración: Por el lema 2 $V_M \geq 0$, lo que implica que $I_n - M'(MQM')^{-}M)$ es semidefinida positiva. Por tanto,

$$r'(I_n - M'(MQM')^{-}M)r \geq 0$$

Además, por la definición de *r* en (9), *r'r* = 1, lo que implica el resultado del corolario en (24).

Lema 5 Bajo los supuestos del lema 4, V_M y Mv son independientes.

Demostración: Considerando $Mv = MQz$, la ecuación (21) sería $V_M = (\tilde{Q}z)'(\tilde{Q}z)$, donde $\tilde{Q} = Q - QM'(MQM')^{-}MQ$ es una matriz simétrica idempotente. Entonces, la multiplicación directa y las propiedades de las inversas generalizadas implican que $\tilde{Q}QM' = 0$. Esto significa que la variable aleatoria $\tilde{Q}z$ de V_M y MQz están incorreladas, que junto con la normalidad de *z* implica que V_M y Mz son independientes, lo que completa la comprobación del lema.

Con los resultados de los lemas anteriores, podemos obtener los principales resultados de esta sección respecto a las propiedades de la distribución de una transformación lineal de residuos internamente *Studentizados*.

Teorema 1 Suponiendo el modelo de regresión lineal en (1) con errores distribuidos esféricamente, $u \in S(n)$, consideremos una transformación lineal arbitraria

$$r_M = Mr \qquad (25)$$

de residuos internamente normalizados *r* definidos en la ecuación (9), donde M es una matriz de orden $m \times n$ tal que MQM' es definida positiva. Entonces, para

Proof: By Lemma 2 $V_M \geq 0$, which implies that $I_n - M'(MQM')^{-}M)$ is positive semi-definite. Thus,

$$r'(I_n - M'(MQM')^{-}M)r \geq 0$$

and by the definition of *r* in (9), *r'r* = 1, which imply the result in (24) of the corollary.

Lemma 5 Under the assumptions of Lemma 4, V_M and Mv are independent.

Proof. Write $Mv = MQz$ and in (21) $V_M = (\tilde{Q}z)'(\tilde{Q}z)$, where $\tilde{Q} = Q - QM'(MQM')^{-}MQ$ is a symmetric idempotent matrix. Then direct multiplication and the properties of generalized inverses imply $\tilde{Q}QM' = 0$. This implies that the defining random variable $\tilde{Q}z$ of V_M and MQz are uncorrelated, which together with the normality of *z* imply that V_M and Mz are independent, completing the proof of the lemma.

With the results of the above Lemmas we can derive the main results of this section regarding the distributional properties of a linear transformation of internally Studentized residuals.

Theorem 1 Assuming the linear regression model in (1) with spherically distributed errors, $u \in S(n)$, consider an arbitrary linear transformation

$$r_M = Mr \qquad (25)$$

of the internally normalized residuals *r* defined in equation (9), where M is an $m \times n$ matrix such that MQM' is positive definite. Then for $m < n - p$ the joint

$m < n - p$ la distribución conjunta del vector aleatorio r_M de orden $m \times 1$ es

distribution of the $m \times 1$ random vector r_M is

$$f_{r_M}(x) = c_{n-p,m} \mid MQM' \mid^{-1/2} \left(1 - x'(MQM')^{-1}x\right)^{\frac{1}{2}(n-p-m)-1} \tag{26}$$

para $x'(MQM')^{-1}x \leq 1$, donde

for $x'(MQM')^{-1}x \leq 1$, where

$$c_{n-p,m} = \frac{\Gamma[(n-p)/2]}{\pi^{m/2}\Gamma[(n-p-m)/2]} \tag{27}$$

$$c_{n-p,m} = \frac{\Gamma[(n-p)/2]}{\pi^{m/2}\Gamma[(n-p-m)/2]} \tag{27}$$

y $\Gamma(\cdot)$ es la función Gamma.

and $\Gamma(\cdot)$ is the Gamma function.

Demostración: Bajo el supuesto de no singularidad, MQM' es definida positiva, existe la inversa $(MQM')^{-1}$ y $r = m$, es decir, el rango de la matriz es m. El límite, $x'(MQM')^{-1}x \leq 1$, sigue del colorario 2. Por el lema 2, $r = v/\|v\|$, donde v es un vector aleatorio normal. Esto implica que $Mr/\|r\| \overset{d}{=} Mv/\|v\|$ Por tanto, encontrando la distribución de

Proof. Under the nonsingularity assumption MQM' is positive definite, the inverse $(MQM')^{-1}$ exists, and $r = m$, i.e., the rank of the matrix is m. The bounds, $x'(MQM')^{-1}x \leq 1$, follow from Corollary 2. By Lemma 2, $r = v/\|v\|$, in which v is a normal random vector. This implies $Mr/\|r\| \overset{d}{=} Mv/\|v\|$. Thus, finding the distribution of

$$r_M^v = \frac{Mv}{\|v\|} \tag{28}$$

$$r_M^v = \frac{Mv}{\|v\|} \tag{28}$$

obtenemos la distribución de r_M. Con estos resultados, el resto puede seguirse de forma análoga a Ellenberg (1973). Esto es, debido al resultado de la distribución χ^2 de V_M en el lema 4, la normalidad de $v_M = Mv$ y la independencia de v_M y V_M por el lema 5, su función de densidad conjunta es el producto de sus funciones de densidad, que resulta en

gives the distribution of r_M. With these results, the rest can be proceeded parallel to Ellenberg (1973). That is, due to the χ^2-distribution result of V_M in Lemma 4, normality of $v_M = Mv$, and independence of v_M and V_M by Lemma 5, their joint density is the product of their densities, resulting to

$$f_{v_M,V_M}(u,v) = \frac{1}{\pi^{\frac{1}{2}m}2^{\frac{1}{2}(n-p)} \mid MQM' \mid^{\frac{1}{2}} \Gamma[(n-p-m)/2]} v^{\frac{1}{2}(n-p-m)-1} \times \exp\left\{-\frac{1}{2}[u'(MQM')^{-1}u + v]\right\} \tag{29}$$

Definiendo las siguientes transformaciones

$$x = u/\sqrt{y/(n-p)}$$
$$y = u'(MQM')^{-1}u + v \quad (30\text{-}31)$$

el Jacobiano de la transformación es

$$y^{\frac{1}{2}m} \quad (32)$$

Utilizando los resultados anteriores, la función de densidad conjunta de r_M^v y $s = v'v$ resulta

Define next transformations

$$x = u/\sqrt{y/(n-p)}$$
$$y = u'(MQM')^{-1}u + v \quad (30\text{-}31)$$

The Jacobian of the transformation is

$$y^{\frac{1}{2}m} \quad (32)$$

Using these, the joint density of r_M^v and $s = v'v$ becomes

$$\begin{aligned} f_{r_M^v,s}(x,y) &= y^{\frac{1}{2}m} f_{v_M,V_M}\left(x\sqrt{y/(n-p)}, y - x'(MQM')^{-1}x\right) \\ &= \frac{|MQM'|^{-\frac{1}{2}}}{\pi^{\frac{1}{2}m}\Gamma[(n-p-m)/2]} \\ &\times\left(1-x'(MQM')^{-1}x\right)^{\frac{1}{2}(n-p-m)-1} \\ &\times\frac{1}{2^{(n-p)/2}} y^{\frac{1}{2}(n-p)-1} e^{-\frac{1}{2}y} \end{aligned}$$

Integrando respecto a y se obtiene finalmente la función de densidad marginal de r_M^v, que es de la forma definida en la ecuación (26). Esto completa la comprobación del teorema.

El mencionado teorema determina la función de densidad conjunta en el caso de $m < n - p$. En el límite con $m = n - p = \text{rank}(Q)$ se demuestra que la distribución es uniforme (con respecto a una medida de volumen adecuada). Esto resulta porque por el lema 3

$$r_M = Mr \overset{d}{=} \tilde{M}\tilde{z}/\|\tilde{z}\| \quad (33)$$

donde la matriz $\tilde{M}$ es una matriz cuadrada e invertible. Por tanto, puesto que (por el lema 1) $\tilde{z}/\|\tilde{z}\|$ sigue una distribución uniforme en la esfera C_{n-p} y r_M sigue la misma distribución que la transformación lineal unívoca de $\tilde{z}/\|\tilde{z}\|$, tenemos:

Integrating with respect to y yields finally the marginal density of r_M^v, which is of the form in (26). This completes the proof of the theorem.

The above theorem gives the joint density in the case of $m < n - p$. On the borderline with $m = n - p = \text{rank}(Q)$, the distribution proves to be uniform (w.r.t suitable volume measure). This is because by Lemma 3

$$r_M = Mr \overset{d}{=} \tilde{M}\tilde{z}/\|\tilde{z}\| \quad (33)$$

where the matrix $\tilde{M}$ is a square matrix and invertible. Thus, because (by Lemma 1) $\tilde{z}/\|\tilde{z}\|$ is uniformly distributed on the sphere C_{n-p} and r_M has the same distribution as the one-to-one linear transformation of $\tilde{z}/\|\tilde{z}\|$, we have

Teorema 2 Bajo los supuestos del teorema 1, cuando $m = n - p$, Mr se distribuye uniformemente en la esfera $\{x \in \Re^{n-p} : x'(MQM')^{-1}x = 1\}$.

Con estos resultados, también obtenemos los siguientes resultados secundarios relativos a la distribución uniforme en la esfera C_n.

Corolario 3 Sea U un vector aleatorio n-dimensional que sigue una distribución uniforme en la esfera C_n. Entonces, la distribución conjunta de una transformación lineal arbitraria

$$u_M = MU \tag{34}$$

donde M es una matriz $m \times n$ con rango $m < n$ tal que la función de densidad es:

Theorem 2 Under the assumptions of Theorem 1, when $m = n - p$, Mr is uniformly distributed on the sphere $\{x \in \Re^{n-p} : x'(MQM')^{-1}x = 1\}$.

With these results we obtain also the following side results regarding uniform distribution on the sphere C_n.

Corollary 3 Let U be a random n-vector that is uniformly distributed on the sphere C_n. Then the joint distribution of an arbitrary linear transformation

$$u_M = MU \tag{34}$$

where M is an $m \times n$ matrix with rank $m < n$ is such that the density is

$$f_{u_M}(x) = c_{n,m} \mid MM' \mid^{-1/2} \left(1 - x'(MM')^{-1}x\right)^{\frac{1}{2}(n-m)-1} \tag{35}$$

para $x'(MM')^{-1}x \le 1$, $m < n$ y $c_{n,m}$ es el definido en la ecuación (27). Para $m = n$ la distribución es uniforme en la esfera $\{x \in \Re^n : x'(MM')^{-1}x = 1\}$.

Casos particulares son la distribución conjunta de los márgenes m-variables $(m < n)$ U_m de U que se obtienen seleccionando, por ejemplo $M = (I_m : 0_{n-m})$. Esto resulta en

for $x'(MM')^{-1}x \le 1$, $m < n$, and $c_{n,m}$ is defined equation (27). For $m = n$ the distribution is uniform on the sphere $\{x \in \Re^n : x'(MM')^{-1}x = 1\}$.

Particular special cases are the joint distribution of m-variate $(m < n)$ margins U_m of U that are obtained by selecting for example $M = (I_m : 0_{n-m})$. This results to

$$f_{u_m}(x) = \frac{\Gamma[n/2]}{\pi^{m/2}\Gamma[(n-m)/2]}(1 - x'x)^{\frac{1}{2}(n-m)-1} \quad ,, \quad x'x \le 1 \tag{36}$$

obtenido en Eaton (1981: 392). En el siguiente epígrafe discutiremos aplicaciones más concretas relativas a diversos aspectos inferenciales de la regresión basada en MCO. Sin embargo, antes de eso, abordamos algunos resultados relativos a residuos externamente *Studentizados.*

derived for example in Eaton (1981: 392). In the next section we will discuss more concrete applications related to various aspects OLS based regression inference. Before that we, however, deal with some results related to *externally* Studentized residuals by noting first:

Proposición 4. El lema 5 junto con el lema 2 y el resultado en (8) implican que

$$\frac{M\hat{u}}{\sqrt{\hat{V}_M}} \overset{d}{=} \frac{Mv}{\sqrt{V_M}} \tag{37}$$

donde $\hat{V}_M = \hat{u}'\hat{u} - \hat{u}M'(MQM')^{-}M\hat{u}$. El numerador y el denominador en el lado derecho del ratio, $Mv = V_M$, son independientes. Además, de forma análoga a la comprobación del lema 2, es fácil observar que la distribución de $M\hat{u}/\sqrt{\hat{V}_M}$ es de nuevo independiente del parámetro de escala σ_u^2. Así, una definición natural para los residuos externamente *Studentizados* de una transformación lineal $M\hat{u}$ de residuos sería

$$e_M = M\hat{u}/\sqrt{\hat{V}_M} \tag{38}$$

Puesto que en la ecuación (37), el numerador se distribuye normalmente y es independiente de la raíz cuadrada de la variable *Chi*-cuadrado en el denominador, el ratio sigue una distribución que es una constante múltiple de la distribución t multivariante con una matriz de covarianzas proporcional a MQM' y número de grados de libertad, por el lema 4, igual a $n - p - r$ con $r = \text{rank}(M) = n - p$.

3. APLICACIONES

3.1. Distribución conjunta de las clases de residuos *Studentizados*

En primer lugar, notemos que todas las principales clases de residuos definidos en la literatura pueden obtenerse como casos especiales de la ecuación (25). En particular, todos los resultados implícitos no son dependientes de la normalidad de los errores, $u \in S(n)$, del modelo de regresión definido en (1).

Remark 4 Lemma 5 together with Lemma 2 and the result in (8) imply that

$$\frac{M\hat{u}}{\sqrt{\hat{V}_M}} \overset{d}{=} \frac{Mv}{\sqrt{V_M}} \tag{37}$$

where $\hat{V}_M = \hat{u}'\hat{u} - \hat{u}M'(MQM')^{-}M\hat{u}$. The nominator and the denominator in the right hand side ratio, $Mv = V_M$, are independent. Furthermore, paralleling the proof of Lemma 2, it is easy to see that the distribution of $M\hat{u}/\sqrt{\hat{V}_M}$ is again independent of the scale parameter σ_u^2. Thus, a natural definition for externally Studentized residuals of a linear transformation $M\hat{u}$ of the residuals would be

$$e_M = M\hat{u}/\sqrt{\hat{V}_M} \tag{38}$$

Because in (37) the nominator is normally distributed and independent of the square root of the *Chi*-squared variable in the denominator, the ratio has the distribution that is a constant multiple of multivariate *t*-distribution with a covariance matrix proportional to MQM' and degrees of freedom by Lemma 4 equal to $n - p - r$ with $r = \text{rank}(M) = n - p$.

3. APPLICATIONS

3.1. Joint distribution of classes of Studentized residuals

We note first that all the major classes of residuals defined in literature can be obtained as special cases of (25). In particular all the implied results are not dependent on normality of the errors, $u \in S(n)$, of the regression model in (1).

Por ejemplo, consideremos además de los residuos internamente *Studentizados* definidos en la ecuación (4), otras formas de residuos discutidos p.e. en Chatterjee y Hadi (1988) y Lloynes (1979):

Residuos normalizados:

$$\hat{u}_i / \sqrt{\hat{u}'\hat{u}} \qquad (39)$$

Residuos estandarizados:

$$\hat{u}_i / s \qquad (40)$$

donde $s = \sqrt{u'u/(n-p)}$.

Residuos de Abrahamse-Koerts (normalizados):

$$B'\hat{u} / \sqrt{\hat{u}'\hat{u}} \qquad (41)$$

donde B es una matriz de orden $n \times n$ definida en Abrahamse and Koerts (1971), cumpliendo $\hat{u}'BB'\hat{u} = \hat{u}'\hat{u}$.

Consideremos un subconjunto arbitrario linealmente independiente $I_m = \{i_1, i_2, \ldots, i_m\} \subset \{1, \ldots, n\}$, $m \le n$ de las clases anteriores de residuos. Puede observarse fácilmente que cada uno de ellos es un caso especial de la transformación lineal definida en la ecuación (25). Para nuestro objetivo, definamos M_I como una matriz de orden $m \times n$ en la que cada fila $j = 1, \ldots, m$ es un vector de orden $1 \times n$ con elemento $i_j = 1$ y ceros en el resto, $i_j \in I_m$. Además, denotemos $D^{-1/2}$ a una matriz diagonal de orden $n \times n$ con elementos $(q_{11})^{-1/2}, \ldots, (q_{nn})^{-1/2}$. Entonces, un conjunto, I_m, de residuos internamente *Studentizados*, definidos en la ecuación (4), se obtiene definiendo en la ecuación (25), $M = (n-p)^{1/2} D^{-1/2} M_I$; un conjunto de residuos normalizados, definidos en la ecuación (39), se obtiene definiendo $M = M_I$; un conjunto de residuos

For example, consider in addition to the internally Studentized residuals defined in equation (4), other forms of residuals discussed e.g. in Chatterjee and Hadi (1988) and Lloynes (1979):

Normalized residuals:

$$\hat{u}_i / \sqrt{\hat{u}'\hat{u}} \qquad (39)$$

Standardized residuals:

$$\hat{u}_i / s \qquad (40)$$

where $s = \sqrt{u'u/(n-p)}$.

Abrahamse-Koerts residuals (normalized):

$$B'\hat{u} / \sqrt{\hat{u}'\hat{u}} \qquad (41)$$

where B is an $n \times n$ matrix defined in Abrahamse and Koerts (1971), satisfying $\hat{u}'BB'\hat{u} = \hat{u}'\hat{u}$.

Consider an arbitrary linearly independent subset $I_m = \{i_1, i_2, \ldots, i_m\} \subset \{1, \ldots, n\}$, $m \le n$ of the above classes of residuals. It is easily seen that each of these is a special case of the linear transformation defined in equation (25). For the purpose, define M_I as an $m \times n$ matrix in which each row $j = 1, \ldots, m$ is a $1 \times n$ vector with element $i_j = 1$ and zeros elsewhere, $i_j \in I_m$. Furthermore, let $D^{-1/2}$ denote an $n \times n$ diagonal matrix with elements $(q_{11})^{-1/2}, \ldots, (q_{nn})^{-1/2}$. Then a set, I_m, of internally Studentized residuals, defined in equation (4), is obtained by defining in equation (25), $M = (n-p)^{1/2} D^{-1/2} M_I$, a set of normalized residuals, defined in equation (39), is obtained by defining $M = M_I$, a set of standardized residuals, defined in equation (40), is obtained by defining $M = (n-p)^{1/2} M_I$, and a set of

estandarizados, definidos en la ecuación (40), se obtiene definiendo $M = (n-p)^{1/2} M_I$; y un conjunto de residuos de Abrahamse-Koerts, definidos en la ecuación (41), se obtiene definiendo $M = M_I B'$.

3.2. Inferencia estadística

3.2.1. Variables omitidas

En particular, seleccionando $m = 1$ de forma que M se convierte en un vector fila, el teorema 1 implica que la función de densidad de la distribución de una combinación lineal (no singular) arbitraria $r_m = m'\hat{u} / \| \hat{u} \|$, donde m es un vector de orden $n \times 1$ de números reales que verifican $m'Qm > 0$, es

$$f_{r_m}(x) = \frac{\Gamma[(n-p)/2](m'Qm)^{-1/2}}{\Gamma[(n-p-1)/2]\Gamma[1/2]} \left(1 - \frac{x^2}{mQ'm}\right)^{\frac{1}{2}(n-p-1)-1} \tag{42}$$

para $|x| \leq \sqrt{m'Qm}$. Por consiguiente, $r_m^2 / m'Qm$ sigue una distribución Beta con parámetros 1/2 and $(n - p - 1)/2$. Fijando el componente i-ésimo en m igual a 1 y el resto de elementos igual a cero, lleva a la función de densidad para un simple residuo internamente *Studentizado* $\tilde{r}_i = \hat{u}_i / s\sqrt{q_{ii}}$ definido en la ecuación (4) donde $s = \sqrt{\hat{u}\hat{u}/(n-p)}$,

$$f_{r_i}(r) = \frac{\Gamma[(n-p)/2]}{\Gamma[(n-p-1)/2]\Gamma[1/2]\sqrt{n-p}} \left(1 - \frac{r^2}{n-p}\right)^{\frac{1}{2}(n-p-1)-1} \tag{43}$$

$|r| \leq \sqrt{(n-p)}$.

De nuevo, $r_i^2/(n-p)$ sigue una distribución Beta con parámetros 1/2 and $(n - p - 1)/2$. Por tanto, las distribuciones de residuos simples internamente *Studentizados* y sus

Abrahamse-Koerts residuals are obtained, defined in equation (41), by defining $M = M_I B'$.

3.2. Statistical inference

3.2.1. Omitted variables

In particular, selecting $m = 1$ such that M becomes a row vector, Theorem 1 implies that the density function of the distribution of an arbitrary (nonsingular) linear combination $r_m = m'\hat{u} / \| \hat{u} \|$, where m is an $n \times 1$ vector of real numbers satisfying $m'Qm > 0$, is

for $|x| \leq \sqrt{m'Qm}$. Thus, $r_m^2 / m'Qm$ follows a Beta distribution with parameters 1/2 and $(n - p - 1)/2$. Setting the ith component in m equal to 1 and all others equal to zero gives the density function for a single internally Studentized residual $\tilde{r}_i = \hat{u}_i / s\sqrt{q_{ii}}$ defined in equation (4), where $s = \sqrt{\hat{u}\hat{u}/(n-p)}$,

Again, $r_i^2/(n-p)$ follows a Beta distribution with parameters 1/2 and $(n - p - 1)/2$. Thus, the distributions of single internally Studentized residuals and their arbitrary

combinaciones lineales arbitrarias pertenecen a la misma familia de distribuciones de forma que a través de simples transformaciones están idénticamente distribuidos como una distribución Beta con parámetros 1/2 and $(n-p-1)/2$. Como es obvio, esto facilita las inferencias basadas en residuos individuales o en sus combinaciones lineales. Además, como se muestra más adelante, la situación puede resultar más fácil introduciendo una transformación que lleva a una ditribución *t* común.

De hecho, si *studentizamos* $m'\hat{u}$ de la misma forma que $\tilde{r}_i$ en la ecuación (4) definiendo $\tilde{r}_m = m'\hat{u}/s\sqrt{q_m}$, donde $s=\sqrt{\hat{u}'\hat{u}/(n-p)}$ y $q_m = m'Qm$, podemos escribir

linear combinations belong to the same family of distributions such that through simple transformations they are identically distributed as the Beta distribution with parameters 1/2 and $(n-p-1)/2$. This obviously facilitates inference based on individual residuals or their linear combinations. Moreover, as is shown below, the situation can be further facilitated by introducing a transformation that leads to a common *t*-distribution.

In fact, if we studentize $m'\hat{u}$ in a manner of $\tilde{r}_i$ in equation (4) by defining $\tilde{r}_m = m'\hat{u}/s\sqrt{q_m}$, where $s=\sqrt{\hat{u}'\hat{u}/(n-p)}$ and $q_m = m'Qm$, we can write

$$t_m = \tilde{r}_m\left(\frac{n-p-1}{n-p-\tilde{r}_m^2}\right)^{\frac{1}{2}} = \frac{m'\hat{u}/\sqrt{m'Qm}}{\sqrt{\hat{V}_m/(n-p-1)}} \overset{d}{=} \frac{m'v/\sqrt{m'Qm}}{\sqrt{V_m/(n-p-1)}} \qquad (44)$$

donde, $\hat{V}_m = \hat{u}'\hat{u}-(m'\hat{u})^2/(m'Qm)$, $V_m = v'v-(m'v)^2/(m'Qm)$ y $v = Qz$ con $z \sim N(0, I_n)$. Así, t_m es una variable aleatoria que sigue una distribución *t* con $n-p-1$ grados de libertad en virtud de los lemas 4 y 5 y que $m'v/\sqrt{m'Qm} \sim N(0,1)$. De nuevo, un caso especial es un residuo simple, en cuyo caso la segunda relación en la ecuación (44) puede escribirse como

where $\hat{V}_m = \hat{u}'\hat{u}-(m'\hat{u})^2/(m'Qm)$, $V_m = v'v-(m'v)^2/(m'Qm)$ and $v = Qz$ with $z \sim N(0, I_n)$. Thus, t_m is a *t*-distributed random variable with $n-p-1$ degrees of freedom by virtue of Lemma 4, Lemma 5, and that $m'v/\sqrt{m'Qm} \sim N(0,1)$. Again, a special case of this is a single residual, in which case the second relation in the middle of (44) can be written as

$$t_i = \frac{\hat{u}_i}{s_{(i)}\sqrt{q_{ii}}} \qquad (45)$$

donde

$$t_i = \frac{\hat{u}_i}{s_{(i)}\sqrt{q_{ii}}} \qquad (45)$$

where

$$s_{(i)}^2 = \frac{(n-p)s^2-\hat{u}_i^2/q_{ii}}{n-p-1} \qquad (46)$$

es la media cuadrática residual de una muestra de la que se ha eliminado la observación *i*-ésima de la regresión (p.e.,

$$s_{(i)}^2 = \frac{(n-p)s^2-\hat{u}_i^2/q_{ii}}{n-p-1} \qquad (46)$$

is the residual means square from a sample with the *i*th observation removed from the regression (e.g., Beckman and Trussell,

Beckman y Trussell, 1974). Por la ecuación (44), la relación entre el residuo externamente *Studentizado*, t_i, y el residuo internamente *Studentizado*, $\tilde{r}_i$, definido en la ecuación (4), es

$$t_i = \tilde{r}_i \left(\frac{n-p-1}{n-p-\tilde{r}_i^2} \right)^{\frac{1}{2}} \qquad (47)$$

(c.f.r. Cook y Weisberg 1982: 20). Los residuos individuales se utilizan normalmente como estadísticos de diagnóstico para la comprobación del modelo, buscando datos atípicos y observaciones influyentes (véase Cook y Weisberg, 1982). El resultado final es que una función lineal (no singular) arbitraria de residuos internamente *Studentizados* puede transformarse en un estadístico t con $n - p - 1$ grados de libertad. Por tanto, utilizando la sencilla transformación recogida en (44), más que confiando en la distribución Beta, puede utilizarse la conocida distribución t en las correspondientes inferencias estadísticas.

Si m en la definición de $\widetilde{r}_m$ en el estadístico definido en (44) es un vector de observaciones de la variable omitida en la regresión $y = X\beta + u$, t_m es un estadístico t para contrastar la significación de la omisión. Esto puede generalizarse inmediatamente a matrices de forma que las transformaciones lineales de los residuos normalizados pueden utilizarse para contrastar variables omitidas en una regresión. De forma más precisa, denotemos como Z m variables adicionales de la regresión en (1) tal que

$$y = X\beta + Z\gamma + e \qquad (48)$$

donde γ es un vector m-dimensional con los coeficientes de pendiente adicionales y $e \in S(n)$.

La hipótesis nula a contrastar es:

$$H_0 : \gamma = 0 \qquad (49)$$

1974). By equation (44), the relationship between the externally Studentized residual, t_i, and the internally Studentized residual, $\tilde{r}_i$, defined in equation (4) is

$$t_i = \tilde{r}_i \left(\frac{n-p-1}{n-p-\tilde{r}_i^2} \right)^{\frac{1}{2}} \qquad (47)$$

(c.f. Cook and Weisberg (1982), p. 20). Individual residuals are typically used as diagnostic tools for the model checking, testing for outliers and influential observations (see, Cook and Weisberg, 1982). The end result is that an arbitrary (nonsingular) linear function of internally Studentized residuals can be transformed to a t-statistic with $n - p - 1$ degrees of freedom. Thus, utilizing the simple transformation in (44), rather than relying on the Beta distribution the familiar t-distribution can be used instead in the related statistical inference.

If m in the definition of $\widetilde{r}_m$ in statistic (44) is an observation vector of an omitted variable from the regression $y = X\beta + u$, t_m is a t-statistic for testing the significance of the omission. This generalizes immediately to matrices such that linear transformations of the normalized residuals can be utilized as such for testing omitted variables from a regression. More precisely, let Z denote m additional variables of the regression in (1) such that

$$y = X\beta + Z\gamma + e \qquad (48)$$

where γ is an m-vector of the additional slope coefficients and $e \in S(n)$.

The null hypothesis to be tested is

$$H_0 : \gamma = 0 \qquad (49)$$

Dados los residuos $\tilde{u}$ procedentes de la regresión $y = X\beta + u$, el contraste de la hipótesis (49) puede basarse en la transformación lineal definida en (25). Eso es así porque, en el caso general, por el lema 2

Given residuals $\tilde{u}$ from the regression $y = X\beta + u$, the testing for hypothesis (49) can be based on linear transformation in (25). This is because, in the general case, by Lemma 2

$$t_M^2 = \frac{r'M'(MQM')^{-1}Mr/m}{(r'r - r'M'(MQM')^{-1}Mr)/(n-p-m)} \overset{d}{=} \frac{v'M'(MQM')^{-1}Mv/m}{v'(I - M'(MQM')^{-1}M)v/(n-p-m)} \tag{50}$$

donde las normas $\|v\|$ se han anulado. Por el lema 4 el último ratio es un cociente entre dos variables aleatorias independientes que siguen una distribución *chi*-cuadrado con m y $n-p-m$ grados de libertad, respectivamente, implicando que t_M^2 sigue una distribución F con m y $n-p-m$ grados de libertad. Esto es,

where the norms $\|v\|$ have been canceled out. By Lemma 4 the last ratio is a ratio of two independent *chi*-squared random variables with degrees of freedom m and $n-p-m$, respectively, implying that t_M^2 is F-distributed with m and $n-p-m$ degrees of freedom. That is,

$$t_M^2 \sim F(m, n-p-m) \tag{51}$$

Por tanto, fijando $M = Z'$, y notando que las normas $\|\hat{u}\|$ vuelven a anularse, obtenemos

Thus, setting $M = Z'$, and noting that the norms $\|\hat{u}\|$, again cancel out, we obtain

$$t_Z^2 = \frac{\hat{u}'Z(Z'QZ)^{-1}Z'\hat{u}/m}{\hat{u}'(I - Z(Z'QZ)^{-1}Z')\hat{u}/(n-p-m)} \tag{52}$$

que, por la ecuación (50), sigue una distribución F con m y $n-p-m$ grados de libertad bajo la hipótesis nula. Utilizando álgebra de matrices resulta que:

which by (50) is F-distributed with m and $n-p-m$ degrees of freedom under the null hypothesis. Using little matrix algebra shows that

$$t_Z^2 = \frac{(\hat{u}'\hat{u} - \hat{e}'\hat{e})/m}{\hat{e}'\hat{e}/(n-p-m)} \tag{53}$$

es decir, el habitual estadístico F para contrastar variables omitidas, donde $\hat{e}$ es el vector de residuos por MCO de la regresión en (48).

i.e., the usual F-statistic for testing omitted variables, where $\hat{e}$ is the vector of OLS residuals from regression (48).

Es de notar que todas las hipótesis lineales del tipo $R\beta = q$ definido en la ecuación (13) pueden convertirse de nuevo en el problema de variables omitidas con el estadístico F definido en (52). Así, los resultados desarrollados con anterioridad muestran desde otro ángulo la robustez de los habituales estadísticos *t* y F en el análisis de regresión, en el sentido de que sus distribuciones nulas son independientes de la normalidad de $u \in S(n)$.

3.2.2. Medidas de sensibilidad

Díaz-García *et al*. (2007) proponen varias generalizaciones de las distancias de Cook multivariantes para detectar una o más observaciones influyentes. A continuación demostramos las posibilidades de utilizar los resultados obtenidos en este trabajo en dicha dirección.

Sea $y = X\beta + u$ la regresión inicial como está definida en la ecuación (1) y consideremos la regresión

$$y = X\beta + m\gamma + e \qquad (54)$$

donde m es el vector de observaciones de una simple variable explicativa adicional. Entonces, podría ser de interés mediar la influencia de la omisión de la variable m en la estimación de β. Una medida apropiada es la distancia de Cook.

En lo que sigue podemos suponer sin pérdidad de generalidad que en la regresión definida en (1) $\text{rank}(X) = p$ e $y \sim N(0, \sigma_u^2 I_n)$. Utilizando entonces los resultados del epígrafe 3.2.1 y siguiendo las deducciones de Díaz-García *et al*. (2007) hasta la ecuación (10), obtenemos la distancia de Cook

$$D_m = \frac{1}{qs^2}(\hat{\beta} - \hat{\beta}_{(m)})'(X'X)(\hat{\beta} - \hat{\beta}_{(m)}) \quad (55)$$

It is notable that all linear hypotheses of the form $R\beta = q$ defined in equation (13) can be returned to the above omitted variables problem with F-statistic of the form (52). Thus the above results show from another angle the robustness of the usual t and F-statistics in regression analysis in the sense that their null-distributions are independent of the normality of $u \in S(n)$.

3.2.2. Sensitivity measures

Díaz-García *et al*. (2007) propose several generalizations to multivariate Cook's distances to detect one or more influential observations. Below we demonstrate possibilities to utilize results derived in this paper in that direction.

Let the initial regression be $y = X\beta + u$ as defined in equation (1) and consider the regression

$$y = X\beta + m\gamma + e \qquad (54)$$

where m is the observation vector of a single additional explanatory variable. Then it might be of interest to measure the influence of the omission of variable m on the estimate of β. An appropriate measure is the Cook distance.

In what follows we can assume without loss of generality that in regression (1) $\text{rank}(X) = p$ and $y \sim N(0, \sigma_u^2 I_n)$. Utilizing then results in Section 3.2.1 and following the derivations up to equation (10) in Díaz-García *et al*. (2007), we get the Cook distance

$$D_m = \frac{1}{qs^2}(\hat{\beta} - \hat{\beta}_{(m)})'(X'X)(\hat{\beta} - \hat{\beta}_{(m)}) \quad (55)$$

donde $\hat{\beta}$ es el estimador MCO de β en la regresión inicial con m omitido; $\hat{\beta}_{(m)}$ es el estimador MCO de β cuando m está incluido, y $s^2 = \hat{u}'\hat{u}/(n-p) = \| \hat{u} \|^2 /(n-p)$ es la varianza residual de la regresión incial con m eliminado. Por tanto, D_m mide la influencia o sensibilidad de la omisión en las estimaciones de β. Con el resultado para matrices particionadas, obtenemos una expresión equivalente (Díaz-García *et al*. 2007, ecuación (9)) a la diferencia explícita de las estimaciones debida a la omisión como

$$\hat{\beta} - \hat{\beta}_{(m)} = (X'X)^{-1} X'mm'\hat{u} / q_m \qquad (56)$$

con $q_m = m'Qm = (I_n - X(X'X)^{-1}X')m$. Sustituyendo (56) en (55) y operando resulta:

where $\hat{\beta}$ is the OLS estimator of β from the initial regression with m omitted, $\hat{\beta}_{(m)}$ is the OLS estimator of β when m is included, and $s^2 = \hat{u}'\hat{u}/(n-p) = \| \hat{u} \|^2 /(n-p)$ is the residual variance from the initial regression with m omitted. Thus, D_m measures the influence or sensitivity of the omission on the estimates of β. With result for partitioned matrices, we obtain an analog to (Díaz-García *et al*. (2007), equation (9)) the explicit difference of the estimates due to the omission as

$$\hat{\beta} - \hat{\beta}_{(m)} = (X'X)^{-1} X'mm'\hat{u} / q_m \qquad (56)$$

with $q_m = m'Qm = (I_n - X(X'X)^{-1}X')m$. Using (56) in (55) and arranging terms yields

$$D_m = \frac{(n-p)}{qq_m^2} m' \frac{\hat{u}}{\| \hat{u} \|} \frac{\hat{u}'}{\| \hat{u} \|} mm'X(X'X)^{-1}X'm \qquad (57)$$

$$= \frac{(n-p)(m'm - q_m}{qq_m^2} q_m^2 \qquad (58)$$

donde $r_m = m'\hat{u} / \| \hat{u} \|$. Se observa inmediatamente que si m es ortogonal a las columnas de X, $q_m = m'm$ y $D_m = 0$. Utilizando la expresión (42) y su relación con la distribución Beta presentada en la página 102, tenemos:

where $r_m = m'\hat{u} / \| \hat{u} \|$. It is immediately observed that if m is orthogonal to columns of X, $q_m = m'm$ and $D_m = 0$. Utilizing (42) and its relation to the Beta distribution discussed on page 102, we get

$$D_m \sim a_m Beta\left(\frac{1}{2}, \frac{n-p-1}{2}\right) \qquad (59)$$

donde

$$a_m = \frac{(n-q)(m'm - q_m)}{qq_m} \qquad (60)$$

Si m es un vector de coordenadas con valor uno en la posición *i*-ésima y cero en el resto, obtenemos la distancia de Cook, una versión multivariante de la propuesta y debatida en Díaz-García *et al*. (2007).

$$D_m \sim a_m Beta\left(\frac{1}{2}, \frac{n-p-1}{2}\right) \qquad (59)$$

where

$$a_m = \frac{(n-q)(m'm - q_m)}{qq_m} \qquad (60)$$

If m is a coordinate vector with one in the *i*th position and zeros elsewhere we obtain the Cook distance, a multivariate version of which is proposed and discussed in Díaz-García *et al*. (2007).

3.3. Algunas aplicaciones especiales

Los resultados de este trabajo respecto a las distribuciones pueden aportar mayor comprensión en el campo de estudios de eventos en economía financiera (una excelente revisión puede consultarse en Campbell, Lo, and MacKinlay, 1997, Cap. 4). El tradicional enfoque paramétrico está basado en residuos *Studentizados* externos. Sin embargo, un enfoque no paramétrico basado en sumas de rango del tipo Wilcoxon, sugerido por Corrado (1989) para contrastar las rentabilidades de un único periodo, y ampliado en Campbell y Wasley (1993) para contrastar las sumas de rango multi-día acumuladas, está incrementando su popularidad. Los estadísticos de contraste utilizados son del tipo de residuos internamente *Studentizados*, donde el numerador y el denominador no son independientes. Por consiguiente, la teoría sobre las distribuciones desarrollada en este trabajo puede utilizarse fácilmente en el examen de las propiedades de las distribuciones de los estadísticos de contraste respectivos. Kolari y Pynnönen (2011) desarollan un estudio de evento para contrastar la estrategia de rentabilidades anormales acumuladas con contrastes de rango, donde la teoría de la distribución asintótica se obtiene utilizando los resultados de variables aleatorias internamente normalizadas. Luoma y Pynnönen (2010) y Luoma (2011) son otros ejemplos, donde los resultados anteriores se utilizan en la obtención de distribuciones asintóticas de las versiones de contrastes de rango y signo discutidos en sus trabajos.

4. CONCLUSIONES

Este trabajo obtiene la distribución conjunta de una transformación lineal general de residuos internamente

3.3. Some special applications

The distribution results of this paper can give additional insight in the field of event studies in financial economics (an excellent review is in Campbell, Lo, and MacKinlay, 1997, Ch. 4). The traditional parametric approach is based on external Studentized residuals. However, a non-parametric approach based on Wilcoxon-type rank sums, suggested by Corrado (1989) for testing single period returns, and extended in Campbell and Wasley (1993) for testing cumulative multi-day rank sums, is gaining increasingly popularity. The used test statistics are obviously of the type of internally Studentized residuals, where the nominator and denominator are not independent. Thus, the distribution theory developed in this paper can be readily utilized in examination of the distributional properties of the related test statistics. Kolari and Pynnönen (2011) develop an event study testing strategy of cumulative abnormal returns with rank tests, where the asymptotic distribution theory is derived by utilizing the results of internally normalized random variables. Luoma and Pynnönen (2010) and Luoma (2011) are other examples, where the above results are utilized in deriving the asymptotic distributions of the versions of rank and sign tests discussed in their papers.

4. CONCLUSIONS

This paper derives the joint distribution of a general linear transformation of internally Studentized residual from a general linear regression. Other types of residuals, commonly used in practical applications, can be easily obtained as special cases by

Studentizados procedentes de una regresión lineal general. Otros tipos de residuos, utilizados comúnmente en aplicaciones prácticas, pueden obtenerse fácilmente como casos especiales definiendo de forma apropiada la transformación lineal.

Las distribuciones de conjuntos arbitrarios además de las distribuciones marginales de residuos simples se obtienen como casos especiales de la distribución general mediante la definición de transformaciones lineales de un modo adecuado. Este trabajo también discute algunas aplicaciones potenciales en las que los resultados pueden utilizarse de forma inmediata.

defining the linear transformation appropriately.

The distributions of arbitrary subsets as well as marginal distributions of single residuals are obtained as special cases from the general distribution by defining linear transformation in a suitable manner. The paper discusses also some potential applications in which the results can be readily applied.

BIBLIOGRAFÍA/REFERENCES

Abrahamse, A.P.J. y Koerts, J. (1971). New estimators in regression analysis. *Journal of the American Statistical Association, 66*, 71-74.

Anderson, T.W. y Kai-Tai Fang (1990). On the theory of multivariate elliptically contoured distributions and their applications. In T.W. Anderson & Kai-Tai Fang (Eds.), *Statistical inference in elliptically contoured and related distributions* (pp. 1-23). New York: Allerton Press Inc.

Beckman, R.J. y Trussell, H.J. (1974). The distribution of an arbitrary studentized residual and the effects of updating in multiple regression. *Journal of the American Statistical Association, 69*, 199-201.

Campbell, C.J. y Wasley, C.E. (1993). Measuring security price performance using NASDAQ returns. *Journal of Financial Economics, 33*, 73-92.

Campbell, J.Y., Lo, A.W. y Craig MacKinlay, A. (1997). *The econometrics of financial markets*. Princeton, NJ: Princeton University Press.

Chatterjee, S. y Hadi, A.S. (1988). *Sensitivity analysis in linear regression*. New York: Wiley.

Chmielewski, A.K. (1981). Elliptically symmetric distributions: A review and bibliography. *International Statistical Review, 49*, 67-74.

Corrado, C.J. (1989). A nonparametric test for abnormal security price performance in event studies. *Journal of Financial Economics, 23*, 385-395.

Cook, D.R. y Weisberg, S. (1982). *Residuals and influence in regression*. Chapman and Hall.

Díaz-García, J.A. y Gutiérrez-Jáimez, R. (2006). The distribution of the residual from a general elliptical multivariate linear model. *Journal of Multivariate Analysis, 97*, 1829-1841.

Díaz-García, J.A. y Gutiérrez-Jáimez, R. (2007). The distribution of residuals from a general elliptical linear model. *Journal of Statistical Planning and Inference, 137*, 2347-2354.

Díaz-García, J.A., Gutiérrez-Jáimez, R. y Alvarado-Castro, V.M. (2007). Exact distributions for sensitivity analysis in linear regression. *Applied Mathematical Sciences, 1*, 1083-1100.

Eaton, M.L. (1981). On the projections of isotropic distributions. *Annals of Statistics, 9*, 391-400.

Ellenberg, J.K. (1973). The joint distribution of the standardized least squares residuals from a general linear regression. *Journal of the American Statistical Association, 68*, 941-943.

Johnston, J. y DiNardo, J. (1997). *Econometric methods* (4th ed.). Singapore: McGraw-Hill.

Kariya, T. y Eaton, M.L. (1977). Robust tests for spherical symmetry. *Annals of Statistics, 5*, 206-215.

Kolari, J. y Pynnönen, S. (2011). Nonparametric rank tests for event studies. *Journal of Empirical Finance, 18*, 953-971.

Luoma, T. (2011). *Nonparametric event study tests for testing cumulative abnormal returns*. Acta Wasaensia.

Luoma, T. y Pynnönen, S. (2010). Testing for cumulative abnormal returns in event studies with the rank tests, *Working Paper*, University of Vaasa (Submitted).

Lloynes, R.M. (1979). A note on Prescott's upper bound for normed residuals. *Biometrika, 66*, 387-389.

Margolin, B.H. (1977). The distribution of internally studentized statistics via Laplace transform inversion. *Biometrika, 64*, 573-582.

Muirhead, R.J. (1982). *Aspects of multivariate analysis*. New York: Wiley.

Pynnönen, S. (2012). Distribution of an arbitrary linear transformation of internally Studentized residuals of multivariate regression with elliptical errors. *Journal of Multivariate Analysis, 107*, 40-52.

Stefansky, W. (1972). Rejecting outliers in fractional designs. *Technometrics*, *14*, 469-479.

Pecvnia Monográfico 2012, pp. 111-136

EL PAPEL DE LA ESTADÍSTICA EN LA METODOLOGÍA SEIS SIGMA. UNA PROPUESTA DE ACTUACIÓN EN SERVICIOS SANITARIOS / *THE KEY ROLE OF STATISTICAL METHODS IN SIX-SIGMA: A PROPOSAL OF IMPLEMENTATION IN HEALTH CARE SERVICES*

Carmen Huerga Castro[1]
carmen.huerga@unileon.es
Julio I. Abad González[1]
julio.abad@unileon.es
Pilar Blanco Alonso[1]
pilar.blanco@unileon.es

Universidad de León

Resumen

La metodología seis sigma es un programa de mejora continua de la calidad que, en base a hechos y datos, persigue reducir errores y avanzar hacia altos objetivos de calidad. Ofrece un enfoque estructurado, analítico y racional para el establecimiento de proyectos de mejora acordes con los objetivos planteados. Si bien la popularidad del seis sigma se deriva de su aplicación en los procesos productivos del sector industrial, cada vez está más extendida su aplicación en el sector servicios y, por ende, en los servicios sanitarios donde la "satisfacción del cliente" adquiere una relevancia vital.
La aplicación de seis sigma requiere el uso de un amplio abanico de herramientas estadísticas, de hecho, el término sigma representa la desviación típica de una distribución y es el factor clave para conocer la variabilidad de la misma. Por ello, en este trabajo señalamos las herramientas más apropiadas en cada etapa o fase de implementación del seis sigma (definir, medir, analizar, mejorar y controlar) y presentamos una propuesta de aplicación en un servicio sanitario.

Palabras clave: Seis sigma; Control estadístico de la calidad; Servicio sanitario.

Abstract

Six-Sigma is a strategy for continuous quality improvement based on facts and data that attempts to reach higher quality standards and lower number of defects. Six-Sigma provides a structured,

[1] Facultad de Ciencias Económicas y Empresariales, Departamento de Economía y Estadística, Área de Estadística e Investigación Operativa. Universidad de León, Campus de Vegazana, 24071-León (España).

analytic and rational approach that allows the implementation of quality improvement projects according to the planned objectives. Although its current popularity is mainly due to its widespread implementation in the industrial sector, it is also being increasingly used in the services sector, such as in health care services, where the customer's satisfaction has an even more crucial relevance.
Six-Sigma involves the use of a wide range of statistical tools; in fact, the term *sigma* means standard deviation, which is a key measure of the distribution's variability. In this paper the statistical tools more suitable for each phase of Six-Sigma's adoption are presented as well a proposal of its adoption in health care services.

Keywords: Six-Sigma; Statistical quality control; Health care services.

1. INTRODUCCIÓN

Seis sigma se inicia en Motorola a finales de la década de 1980 como una estrategia de mejora de la calidad, adoptada para enfrentarse a una crisis en la calidad de sus productos. Las aportaciones de Bill Smith y Mikel Harry impulsaron una metodología que requiere el uso de herramientas estadísticas, y cuyo objetivo consiste en eliminar la variabilidad de los procesos y producir los resultados esperados, con el mínimo número de defectos, bajos costes y máxima satisfacción del cliente. Esta filosofía contrasta con la forma tradicional de asegurar la calidad, basada únicamente en la inspección final. A partir de esta propuesta inicial, que llevó a Motorola a ganar el prestigioso premio de calidad Baldrige en 1988, la metodología seis sigma ha evolucionado, pasando de ser una simple herramienta de medición de defectos, a una filosofía de mejora de calidad que se basa en hechos y datos y encuentra en la Estadística valiosas herramientas de análisis. Seis sigma se apoya en medidas objetivas y pone la atención en los resultados a corto plazo para conseguir mejoras a más largo plazo.

El nivel sigma está directamente relacionado con la cantidad de oportunidades de defecto que puede presentar un producto o servicio. De hecho, el objetivo del seis sigma es lograr procesos que no tengan más de 3,4 defectos por millón de oportunidades.

1. INTRODUCTION

Six-Sigma originated at Motorola Inc. in the late 1980s as a quality improvement strategy designed to deal with a quality crisis in their products. Bill Smith and Mike Harry's contributions fostered a strategy that, by using statistical tools, aimed to reduce processes variability and achieve the expected results with minimum defects, low costs and maximum customers' satisfaction. Six-Sigma contrasts the traditional way of quality assurance, just based on a final inspection. From Motorola's initial proposal, awarded in 1988 with the prestigious Malcolm Baldrige National Quality Award, Six-Sigma has evolved and become a philosophy of continuous quality improvement that, based on objective measures and using different statistical tools, focus on short-term results in order to achieve long-term improvements. But Six-Sigma is closely related to the number of nonconformities per million opportunities of a product or service. In fact, Six-Sigma's goal is to achieve processes with a level of performance equating 3.4 nonconformities per million opportunities.

Es evidente que las consecuencias que un defecto puede acarrear, difieren según el tipo de producto o servicio, y aunque no necesariamente todos los procesos deben operar a nivel seis sigma, en las empresas que prestan servicios de salud, buscar un nivel de calidad seis sigma debería ser un objetivo primordial.

La atención sanitaria puede entenderse como un proceso de producción. Cuando un paciente entra en un sistema sanitario, desencadena un proceso de producción que debe satisfacer sus necesidades tanto como sea posible (Vuori, 1996: 13). Según afirma Vuori es posible identificar algunos componentes básicos del proceso de la atención al paciente: realización de la historia clínica, diagnóstico, tratamiento y seguimiento, aunque el contenido real de estos componentes depende del paciente y sus necesidades.

La metodología seis sigma en el área de la salud comenzó en Estados Unidos a finales de la década de 1990. En 1998 el Dr. Chassin publicó un artículo titulado "Is Health Carey Ready for Six Sigma Quality?" donde expone un análisis de las causas subyacentes de los problemas de calidad y sugiere la adopción de métodos seis sigma dentro de un marco global para el cambio. Encontramos también propuestas de aplicación de la metodología seis sigma en hospitales chilenos (Ramírez, Pinto, Serpell y Enberg, 2007) en un enfoque donde se plantean las ventajas y consecuencias de incorporar esta metodología en el sector servicios y más concretamente en el de la salud. Otras aplicaciones interesantes muestran las posibilidades para reducir errores médicos (Buck, 2001) y del mismo modo en los hospitales alemanes se ha implementado esta metodología en casos concretos del departamento de enfermería (Van den Heuvel, Doess y Vermaat, 2004). Por su parte, Jiménez et al. (2007) la aplican en la mejora de la logística sanitaria.

Obviously, the consequences of nonconformities depend on the kind of products or services, and maybe performing at a Six-Sigma level could not be possible in every process; however, achieving this quality level should be a main goal for the health care companies.

In fact, health care could be considered as a production process: when a patient contacts the health care system, a production process oriented to meet their needs as much as possible starts up (Vuori, 1996: 13). According to Vuori, it is possible to identify some basic components in the patient attendance process: medical record writing, diagnosis, treatment and follow-up, though the actual content of these components depend on the patient and their needs.

It was in the late 1990s when Six-Sigma started to be implemented in Health Care companies in the United States. An analysis of the underlying causes for medical errors and a proposal for the implementation of Six-Sigma methods in a global improvement framework is presented in 1998 Chassin's paper *Is health care ready for Six Sigma quality?* Ramirez, Pinto, Serpell and Enberg (2007) discuss the advantages and consequences derived from the adoption of Six-Sigma methods in the services sector in general, and in health care services in particular, and suggest its adoption in Chilean Hospitals. Buck (2001) focuses on the possibilities of Six-Sigma in reducing medical errors, while Van den Heuvel, Doess and Vermaat (2004) describe its implementation in the nursery department of a Dutch hospital, and Jiménez et al. (2007) propose their adoption to improve healthcare logistics.

Esto demuestra que las organizaciones que prestan sus servicios en el sector de la salud no son ajenas a los principios del seis sigma y buscan nuevas formas para mejorar sus procesos y reducir la insatisfacción de los pacientes, aunque ello requiere el compromiso de todos los agentes implicados (médicos, enfermeras, administradores, personal de apoyo, etc.). La puesta en marcha de un plan de trabajo seis sigma en un servicio de salud lleva consigo una buena planificación e infraestructura. El punto de partida es el reconocimiento y la localización de los problemas para saber qué proyectos hay que abordar y cómo llevarlos a cabo. En esta tarea, la metodología que se propone ofrece un marco eficaz y un nuevo enfoque perfectamente aplicable en la asistencia sanitaria.

Presentamos a continuación los conceptos y las etapas que integran esta metodología, así como la potencia de las técnicas estadísticas en la consecución de los objetivos de calidad mediante una propuesta de actuación en el ámbito sanitario.

2. LA METODOLOGÍA SEIS SIGMA: CONCEPTO, CARACTERÍSTICAS Y FASES

Desde el punto de vista estadístico sigma representa la desviación típica (σ) de un conjunto de datos, es decir la dispersión respecto al valor medio. En el contexto que nos ocupa sigma representa la variación existente en un proceso en relación con las especificaciones o requerimientos establecidos. La magnitud de sigma está relacionada directamente con el número de unidades defectuosas (que no cumplen las especificaciones), de modo que si la desviación es pequeña habrá pocos valores fuera de las especificaciones.

All this proves that companies and institutions in the health care sector are not alien to Six-Sigma principles, i.e. they search new ways of improving their processes and increase their patients' satisfaction, and that requires the commitment of the whole organization (physicians, nurses, managerial and administrative staff, etc.).

Moreover, the implementation of the Six-Sigma methodology in health services requires both planning and infrastructure. It should starts by detecting and spotting where are the problems in order to determine which projects should be developed and how to do it. For this purpose, this methodology provides with an effective framework and a new approach which could be perfectly applied in the health care area.

In the following sections, the concepts and phases integrating the Six-Sigma methodology are presented as well as an example of implementation in the health care context that illustrates the key role played by the statistical methods in the quality improvement.

2. SIX-SIGMA: CONCEPT, FEATURES AND PHASES

In statistical terms, sigma (σ) represents the standard deviation of a data set, i.e. the dispersion of those data from their mean or average. In a production process, sigma stands for the actual variation of a process in terms of the established specifications or requirements. Sigma size is directly related to the number of nonconformities (those product units that do not meet the specifications), i.e. the lower sigma is, the less nonconformities will be.

En la terminología de control de calidad, al hablar de seis sigma se está midiendo el número de sigmas que se incluyen dentro del intervalo definido por los límites de especificación superior e inferior; cuando sigma es pequeño mayor es el número de las mismas que caben dentro de las especificaciones y, en consecuencia, menor es el número de unidades defectuosas.

A partir de este concepto estadístico se ha desarrollado toda una filosofía de calidad enfocada a la mejora continua, mediante el análisis de los procesos y la puesta en marcha de métodos adecuados para medir y controlar su funcionamiento. Así, cuando se decide poner en marcha la metodología seis sigma lo que se pretende es minimizar defectos hasta tener como máximo 3,4 defectos por millón de oportunidades. Este objetivo se logra reduciendo la variación existente en los procesos, de modo que sean más predecibles y los productos o servicios que originan sean mejores.

Podemos resumir las características principales de esta metodología en las siguientes:

- Su objetivo es lograr productos y servicios de calidad
- Establece como prioridad al cliente.
- Se basa en hechos y datos, diseñando un esquema para recogerlos y analizarlos.
- Sus principios son aplicables tanto en los procesos productivos como en los servicios.
- Requiere la implicación, participación y compromiso de todo el personal implicado.

La puesta en marcha de la metodología seis sigma se lleva a cabo mediante el ciclo DMAMC (Definir, Medir, Analizar, Mejorar y Controlar).

In quality control, the term Six-Sigma refers to the number of standards deviations that are included between the lower and upper specification limits; if sigma is small, then a high number of standard deviations can be included between those limits and therefore, a lower number of nonconformities will be produced.

Based on this statistical concept, a global philosophy, focused on the continuous quality improvement through the analysis, measuring, monitoring and control of the performance of the processes, has been developed. In fact, Six-Sigma's goal is minimizing the number of defects and keeping it under a level of 3.4 nonconformities per million opportunities. This goal is achieved by reducing the inherent variation of the processes so that they become more predictable and the resulting quality of the products or services improves.

The main features of Six-Sigma strategy could be summarized in the following:

- It aims to achieve products and services with a high standard of quality.
- It is customer-oriented.
- Since it is based on facts and data, it provides with schemes to collect and analyse them.
- It can be applied both to products or services.
- It demands the implication, participation and commitment of the whole staff of the company.

The phases in applying Six-Sigma methodology are often referred to as Define-Measure-Analyse-Improve-Control (DMAIC):

Definir. En esta primera etapa se plantea el problema, se especifica el objetivo o meta que se pretende alcanzar, y se identifican los elementos que intervienen en el proyecto.

Medir. En segundo lugar, se obtiene información sobre la situación actual del proceso que se evalúa, con el fin de detectar las casusas reales de los problemas.

Analizar. A partir de los datos, y usando métodos estadísticos, se procede a su análisis e interpretación.

Mejorar. Decidir y diseñar las acciones de mejora que hay que implementar para atacar las causas de los problemas de modo que el proceso alcance los resultados esperados.

Controlar. Realizar un seguimiento de las acciones de mejora y comprobar sus resultados.

3. TÉCNICAS ESTADÍSTICAS EN LAS ETAPAS DEL SEIS SIGMA

En cualquiera de las fases mencionadas se presentan problemas que requieren la utilización de la Estadística y de sus técnicas. En realidad, seis sigma se sustenta en los principios de calidad como la trilogía de Juran, la metodología de Deming (medir, analizar, mejorar y controlar), las siete herramientas de Ishikawa, el diseño robusto de Taguchi, los gráficos de control de Shewhart, los estudios de capacidad, los postulados de "cero defectos" de Crosby, así como en métodos estadísticos más o menos complejos del diseño experimental, del análisis de regresión, etc. El uso de estas herramientas en combinación con el software estadístico disponible en la actualidad resulta imprescindible para solucionar los problemas de calidad y avanzar en su mejora.

Define: the first step is to define the problem to be addressed, specifying the goals that are aimed, and identifying the elements involved in the project.

Measure: in the second step, information about the current situation of the evaluated process is gathered in order to detect the root causes of defects.

Analyse: in the third step, the data gathered in the previous step are analysed and interpreted by means of statistical tools.

Improve: in the fourth step, improvement actions are determined and designed in order to significantly reduce the defect levels of the process.

Control: in the last step, the improvement actions that have been adopted are followed up in order to check if their expected results have been actually achieved.

3. STATISTICAL TOOLS IN THE SIX-SIGMA PHASES

In any of the phases previously mentioned, situations requiring the application of statistical techniques may appear. In fact, Six-Sigma actually stands on quality principles such as Juran trilogy, Deming cycle (Plan-Do-Study-Adjust), Ishikawa seven basic tools of quality, Taguchi robust design method, Shewhart control charts, capability studies, Crosby zero defects postulate, as well as other statistical methods such as design of experiments or regression analysis. Therefore, in order to solve quality-related problems and promote quality improvements, the use of these tools by means of statistical software currently available becomes essential.

Cada problema puede plantearse mediante la siguiente relación:

$$Y = f(X_1, X_2, \ldots\ldots, X_n)$$

donde *Y* es la variable dependiente que se asocia con el resultado esperado mientras que los factores que influyen en el mismo son las variables independientes $X_1, X_2, \ldots\ldots, X_n$. Pero no todos los factores tienen la misma influencia e importancia, de ahí la necesidad de disponer de herramientas que permitan distinguir entre los muchos triviales y los pocos vitales.

En la metodología que propugna seis sigma se encuentran las bases para encontrar dicha ecuación. Para ello combina diferentes técnicas estadísticas con otras que no lo son estrictamente, como por ejemplo el QFD[2] (Quality Function Deployment) y el FMEA[3] (Failure Mode and Effects Analysis).

Sin ánimo de ser exhaustivos, la Tabla 1 recoge distintas técnicas utilizadas en la metodología seis sigma. Aunque las herramientas de tipo estadístico están presentes en prácticamente todas las fases, es evidente que en la fase de análisis resulta imprescindible el uso de la Estadística. Por otro lado, conviene señalar que una misma técnica o herramienta se puede utilizar en distintas fases.

Any problem could be represented through the following relation:

$$Y = f(X_1, X_2, \ldots\ldots, X_n)$$

where *Y* is the dependent variable represents a performance measure of the output and the influencing factors are represented by the independent variables $X_1, X_2, \ldots\ldots, X_n$. However, since not all the factors have the same influence or importance, it is vital to separate the vital few from the trivial many by using the appropriate methods.

Therefore, this equation can be determined by means of Six-Sigma tools, which include not only statistical methods but also other techniques such as Quality Function Deployment (QFD)[4] or Failure Mode and Effects Analysis (FMEA)[5], being the most relevant those presented in Table 1. Although the statistical tools are used in almost every Six-Sigma phase (in fact, the same tool can be used in more than one phase), it is obvious that their use is particularly essential in the analysis phase.

[2] El QFD tiene como objetivo trasladar las necesidades de los clientes a requisitos y características de calidad creando procesos que puedan contribuir al aseguramiento de estas características. Se puede traducir como Despliegue de la Función de Calidad.

[3] El FMEA se utiliza para identificar, evaluar y prevenir los fallos y sus efectos en un producto o servicio. De este modo permite determinar qué características hay que controlar. Se puede traducir como Análisis Modal de Fallos y Efectos.

[4] QFD aims to transform customers' needs into quality features and requirements by developing processes capable of assuring these features.

[5] FMEA is used to identify, evaluate and prevent defects and their consequences in products or services and, eventually, to determine which features must be controlled.

Tabla 1

FASE	CONCEPTO	HERRAMIENTAS
Definir	Identificar procesos sobre los que actuar, detallando las necesidades de los clientes	-QFD -Diagrama de flujo -FMEA
Medir	Desarrollar plan de recogida de datos. Obtener información y medir las características de calidad	-Plantillas recogida de datos -Muestreo estadístico -FMEA -Brainstorming
Analizar	Analizar la información para determinar las causas de los problemas	-Histograma -Gráfico de Pareto -Diagrama causa-efecto -Diagrama de dispersión -Análisis de regresión -Series temporales -Pruebas de hipótesis -Gráficos de control -Estudios de capacidad y nivel sigma -ANOVA
Mejorar	Buscar e implementar acciones de mejora	-Brainstorming -Diseño de experimentos -FMEA
Controlar	Comprobar y mantener la mejora. Establecer plan de control	-Gráficos de control -Análisis de capacidad -Determinación del nivel sigma del proceso

Table 1

PHASE	CONCEPT	TOOLS
Define	Identify core processes and define customer requirements	-QFD -Flux diagram -FMEA
Measure	Gather information and measure quality features following a previously designed scheme	-Data collection grids -Statistical sampling -FMEA -Brainstorming
Analyse	Analyse the collected data in order to determine the causes of the defects	-Histogram -Pareto chart -Cause-effect diagram -Scatter plot -Regression analysis -Time series analysis -Hypothesis tests -Control charts -Capability analysis and sigma quality level calculation -ANOVA
Improve	Design and implement improvement actions	-Brainstorming -Design of experiments -FMEA
Control	Check and ensure the improvements are sustained by establishing control plans	-Control charts -Capability analysis -Calculation of the process' sigma quality level

4. DETERMINACIÓN DEL NIVEL SIGMA

En el contexto que nos ocupa la evaluación de los procesos se realiza en base a los niveles de sigma: desde el nivel 1σ hasta el nivel 6σ. Trataremos en este apartado de explicar cómo se traduce el nivel de calidad seis sigma en 3,4 defectos por millón.

Generalmente las características de calidad estudiadas en los procesos se ajustan a un modelo probabilístico normal y, por lo tanto, los datos originados se distribuyen respecto a un valor central (media) con una dispersión que se mide mediante la desviación típica; cuanto más pequeña sea la desviación más centrado y fiable será el proceso y más leptocúrtica será la distribución.

Por otro lado, hay que tener en cuenta que el ámbito de la calidad se entiende que una unidad, producto o servicio es defectuoso si está fuera de los límites de especificación. Estos límites son aquellos entre los que pueden oscilar los valores individuales de la característica de calidad para que el producto sea considerado como aceptable. Son determinados por la dirección, los diseñadores del producto o la normativa legal vigente, y se pueden establecer de forma bilateral o unilateral, de acuerdo con un valor objetivo y unos límites que no se pueden superar. En este sentido, proporcionan una región de variabilidad fuera de la cual las unidades producidas no son válidas.

Los límites de especificación también se conocen como "límites de tolerancia" y por ello pueden expresarse de la forma siguiente:

Límites de especificación =
valor objetivo ± tolerancia

4. SIGMA QUALITY LEVEL CALCULATION

In this context, processes can be evaluated by means of the calculation of their sigma quality level which ranges from 1σ level to 6σ level. In what follows, the relation between 6σ level and 3.4 nonconformities per million opportunities will be addressed.

When analysing a process, the quality features of interest usually follow a normal distribution, therefore the collected data are distributed around a central value (mean) with a certain spread or dispersion that can be measured through the standard deviation.

On the other hand, a unit, product or service will be considered as nonconformity when any of its quality features falls out of the specification limits. These limits, within which variations of quality features are accepted, are determined by the managerial board, the product designers or de current legal regulations. These limits are usually expressed in terms of a target value and a tolerance, according to the following expression: Specification limits = target ± tolerance, i.e.:

LSE= valor objetivo + tolerancia = VO + tolerancia
LIE= valor objetivo – tolerancia = VO - tolerancia

El porcentaje de unidades defectuosas se obtiene entonces calculando la probabilidad de obtener valores fuera de las especificaciones.

Se definen, además, los límites naturales del proceso, o límites de variación natural, como aquellos entre los que se mueve el proceso sin que sea posible mejorarlo. Dichos límites abarcan prácticamente la totalidad de la producción (se admite que contienen el 99,73% de la misma). Si la característica que se controla se distribuye de forma $N(\mu, \sigma)$ los límites de tolerancia natural se sitúan a una distancia de 3σ por encima y por debajo de la media, es decir, vienen dados por: $\mu \pm 3\sigma$.

La comparación entre la variabilidad natural y la variabilidad exigida por las especificaciones se realiza mediante el estudio de la capacidad. La forma común de expresar la capacidad es en términos de índices o medidas adimensionales, que cuantifican el comportamiento del proceso teniendo en cuenta los parámetros del mismo y las especificaciones del producto. Uno de los primeros índices de capacidad se atribuye a Juran y está definido de la forma siguiente:

$$C_p = \frac{LSE - LIE}{6\sigma}$$

Supuesto que el parámetro de interés es la media del proceso y que dicha media coincide con el punto medio del intervalo de especificación, la capacidad de un proceso indica su rendimiento cuando opera bajo control y la posibilidad que tiene de producir dentro de las especificaciones, es decir el índice C_p mide la capacidad potencial del proceso.

Lower specification limit (LSL) = target – tolerance
Upper specification limit (USL) = target + tolerance

Therefore, the percentage of nonconformities can be determined by computing the probabilities of getting values out of the specification limits.

Let us introduce the concept of *natural tolerance limits for the process*, namely the inherent variation range of a process that cannot be reduced. When the quality feature to be controlled is normally distributed, considering as natural tolerance limits a range of three standard deviation (σ) on either side of the mean (μ), i.e. $\mu \pm 3\sigma$, will include almost the whole production of the process (exactly 99.73% of it).

The comparison of this natural tolerance range with the customer tolerance range leads to the study of the capability of a process. A measure of it, alleged to Juran, is the so called process capability index:

$$C_p = \frac{USL - LSL}{6\sigma}$$

Under the assumption that the specification target (the mid-point of the specification range) equates the process mean, this index C_p serves as both an indicator of the process performance when operating on target and also as a measure of the potential capability of the process (the possibility of producing within the specification range):

- Si C_p>1, el proceso es capaz: cuanto mayor es el índice, más capaz será el proceso de producir dentro de las especificaciones.
- Si C_p=1, el proceso es estrictamente capaz: el porcentaje de unidades que no cumplen las especificaciones es sólo del 0,27% (27 de cada 10000 unidades), pero cualquier cambio en la media o en la dispersión incrementaría dicho porcentaje.
- Si C_p<1, el proceso no es capaz: en el proceso se obtienen más de 27 por cada 10000 unidades que no cumplen las especificaciones.

Los índices de capacidad tienen un papel similar a los niveles de calidad (medidos en cantidad de sigmas) ya que ambos tratan de traducir en números la relación entre el funcionamiento del proceso y las exigencias y especificaciones de los productos o servicios. De este modo, la diferencia entre las especificaciones, dividido por el valor de la desviación típica, proporciona el número de sigmas que comprende dicha diferencia. Por tanto, el nivel de calidad, expresado como nivel "**k sigma**", se obtiene dividiendo la mitad del intervalo de especificación entre la desviación típica (véase Figura 1). Desarrollamos a continuación la situación para distintos niveles de calidad.

Proceso centrado: La media del proceso coincide con el valor objetivo (VO: valor medio del intervalo de especificación).

Nivel de calidad 3 sigma. En este caso caben 6 desviaciones típicas (6σ) en el intervalo definido por los límites de especificación y el rendimiento del proceso es del 99,73%.

$$\begin{aligned} P(\mu - 3\sigma < X < \mu + 3\sigma) &= P(-3 < Z < 3) \\ &= 0{,}9986 - 0{,}0013 \\ &= 0{,}9973 \end{aligned}$$

- If C_p>1, the process is capable: the larger this index is, the more capable of producing output within the specification limits is the process.
- If C_p=1, the process just meets the specification limits: any shift in the mean or the spread will result in more nonconformities than 27 per 10,000 units.
- If C_p<1, the process is incapable: the output contains more than 27 nonconformities per 10,000 units.

Process capability index is closely related to sigma quality level: both aim to translate into figures the ratio between the process performance and the product or service requirements. In fact, almost the same elements are considered in the calculation of sigma quality level which can be defined as the number of standard deviations included in the tolerance:

$$SQL = \frac{\text{Tolerance}}{\sigma} = \frac{\frac{1}{2}(USL - LSL)}{\sigma} = \frac{USL - LSL}{2\sigma}$$

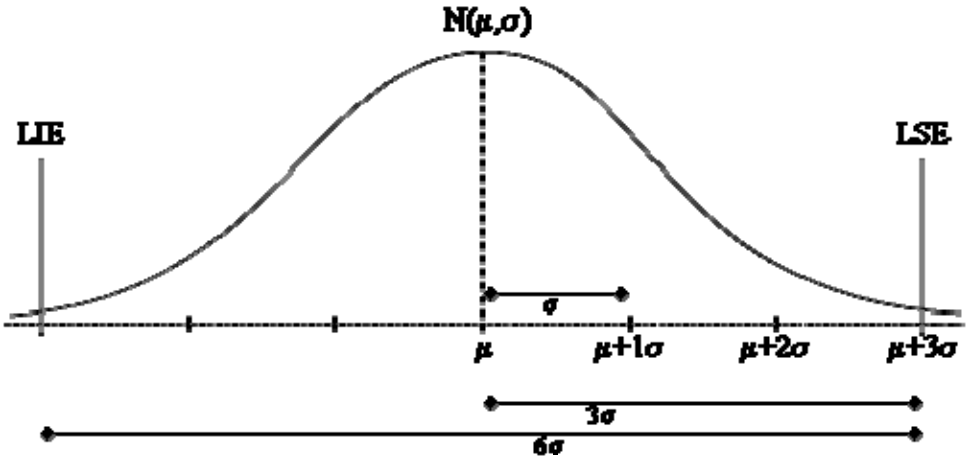
Some examples of different sigma quality levels are commented below:

Process operating on target: the target value (mid-point of the specification range) equates the actual mean of the process.

3 sigma quality level: six standard deviations (6σ) are included in the specification range and the process yield is 99.73% (see Figure 1).

$$\begin{aligned} P(\mu - 3\sigma < X < \mu + 3\sigma) &= P(-3 < Z < 3) \\ &= 0.9986 - 0.0013 \\ &= 0.9973 \end{aligned}$$

Figura 1/*Figure 1*



La proporción de unidades fuera de las especificaciones es de 0,0027 es decir, 27 de cada 10000 unidades o 2700 partes por millón (PPM), lo que supone 1350 PPM fuera de cada límite de especificación. En este caso el índice $C_p=1$.

Nivel de calidad 4 sigma. En este caso el rendimiento del proceso es:

$$P(\mu - 4\sigma < X < \mu + 4\sigma) = 0{,}99993$$

La proporción de unidades fuera de las especificaciones es 0,000063 (63 PPM). En este caso $C_p= 1{,}33$ pues caben 8 desviaciones típicas dentro de las especificaciones.

Nivel de calidad 6 sigma. Significa que en el intervalo definido por las especificaciones, $\mu \pm 6\sigma$, caben 12 desviaciones y el porcentaje de unidades fuera de las especificaciones sería de 0,000000002 lo que equivale a 0,002 PPM. En este caso $C_p= 2$.

The proportion of units that fall out of the specification range is 0.27%, 27 per 10,000 units or 2,700 parts per million (ppm), since 1,350 ppm exceed the upper specification limit and other 1,350 ppm fail to exceed the lower specification limit. In this case the process capability index $C_p=1$.

4 sigma quality level: eight standard deviations (8σ) can be included in the specification range and the process yield is:

$$P(\mu - 4\sigma < X < \mu + 4\sigma) = 0.99993$$

The proportion of units that fall out of the specification range is 63 ppm and the process capability index $C_p=1.33$.

6 sigma quality level: in this case, twelve standard deviations (12σ) can be included in the specification range $\mu \pm 6\sigma$, thus the process capability index $C_p=2$ and the process fallout is 0.002 ppm.

Es evidente que el planteamiento anterior no coincide con los 3,4 PPM que proclama la metodología seis sigma. Ello es debido a que la media μ no siempre coincide con el VO, es decir se admite que, debido a factores aleatorios, el proceso puede estar descentrado y desplazarse hasta ±1,5σ respecto del valor objetivo[6].

Proceso descentrado: Cuando el proceso está descentrado $+1{,}5\sigma$ (de igual forma se razona si el proceso está descentrado $-1{,}5\sigma$), para un **nivel de calidad 3 sigma** habría **66811 PPM** fuera de las especificaciones. Si la característica de calidad estudiada sigue una distribución normal con media $\mu' = \mu + 1{,}5\sigma$, la probabilidad de que la característica estudiada esté dentro del intervalo definido por las especificaciones sería:

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = \{X \sim N(\mu', \sigma)\} = P(-4{,}5 < Z < 1{,}5) = 0{,}9331894$$

En consecuencia, el porcentaje de unidades fuera de las especificaciones es de 0,0668106 que se traduce en 66810,6 PPM.

De la misma manera, un **proceso 6 sigma** supone un rendimiento del 99,99966% y un porcentaje máximo de unidades fuera de las especificaciones de 0,0000034 (**3,4 PPM**) como se observa a continuación:

$$P(\mu - 6\sigma < X < \mu + 6\sigma) = \{X \sim N(\mu', \sigma)\} = P(-7{,}5 < Z < 4{,}5) = 0{,}9999966$$

Por otro lado, cuando el proceso está descentrado y la media del mismo no coincide con el punto medio entre las

Obviously, this result is much lower than the alleged Six-Sigma fallout (which is 3.4 ppm). This difference is due to the fact that the mean of the process do not always equate the target value because of random factors[7] and, therefore, the process may not be operating on target.

Process not operating on target: when the difference between the target value μ (mid-point of the specification range) and the actual mean of the process equates $+1.5\sigma$ (the same reasoning applies if it is -1.5σ), the fallout for a 3 sigma quality level will be 66,811 ppm. This is proved in what follows: if the quality feature under study is normally distributed with a mean $\mu' = \mu + 1.5\sigma$, the probability of fallout will be:

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = \{X \sim N(\mu', \sigma)\} = P(-4.5 < Z < 1.5) = 0.93319$$

Thus, the proportion of units that fall out of the specification range will be 66,811 ppm.

In the same vein, the yield of a 6 sigma quality level process not operating on target is 99.99966% and its fallout is 3.4 ppm:

$$P(\mu - 6\sigma < X < \mu + 6\sigma) = \{X \sim N(\mu', \sigma)\} = P(-7.5 < Z < 4.5) = 0.9999966$$

Moreover, when the process is not operating on target a different process capability index is used:

[6] En el modelo 6 sigma de Motorola se parte de la hipótesis de que, a largo plazo, la media se desplazará, por razones de variabilidad, hasta 1,5 sigma.

[7] Motorola's Six-Sigma model is based on the assumption that, in the long-term, the mean of the process can move up or down up to 1.5σ from the target value.

especificaciones, se suele definir un nuevo índice de capacidad de la forma siguiente:

$$C_{pk} = mín\left\{\frac{\mu - LIE}{3\sigma}, \frac{LSE - \mu}{3\sigma}\right\}$$

Los índices C_p y C_{pk} coinciden cuando el proceso está centrado, pero C_{pk} es menor que C_p en caso contrario. Si el proceso está descentrado de modo que $\mu' = \mu + 1{,}5\sigma$ y el intervalo de especificación está situado en $\mu \pm 3\sigma$ se obtiene

$$C_{pk} = mín\left\{\frac{\mu' - LIE}{3\sigma}, \frac{LSE - \mu'}{3\sigma}\right\} = mín\{1{,}5; 0{,}5\} = 0{,}5$$

Del mismo modo, si el intervalo de especificación es $\mu \pm 6\sigma$, el índice $C_{pk} = 1{,}5$.

La Tabla 2 presenta, para distintos niveles de calidad, los índices de capacidad, el valor de PPM y el rendimiento, cuando el proceso está centrado y cuando está descentrado ± 1,5σ.

$$C_{pk} = min\left\{\frac{\mu - LIE}{3\sigma}, \frac{LSE - \mu}{3\sigma}\right\}$$

C_p and C_{pk} are equal when the process is operating on target, but C_{pk} is lower than C_p otherwise. If the process is operating off target by a difference of $+1.5\sigma$ ($\mu' = \mu + 1.5\sigma$) and the specification range is set to $\mu \pm 3\sigma$:

$$C_{pk} = min\left\{\frac{\mu - LIE}{3\sigma}, \frac{LSE - \mu}{3\sigma}\right\} = mín\{1.5\,, 0.5\} = 0.5$$

Likewise, if the specification range is set to $\mu \pm 6\sigma$, the index $C_{pk} = 1.5$.

In Table 2, different sigma quality levels and their corresponding capability indices, fallouts and yields are presented both considering the process is operating on target and off target by a difference of $\pm 1.5\sigma$.

Tabla 2/*Table 2*

Nivel sigma *SQL*	Proceso centrado / *Process on target*			Proceso descentrado en ±1,5σ *Process off target by ±1,5σ*		
	C_p	PPM defectuosas *Fallout* (ppm)	Rendimiento *Yield*	C_{pk}	PPM defectuosas *Fallout* (ppm)	Rendimiento *Yield*
1,5	0,5	133.614	86,6386%	0,000	501.350	49,8650%
2	0,66	45.500	95,4500%	0,167	308.770	69,1230%
3	1	2.700	99,7300%	0,500	66.811	93,3189%
4	1,33	63	99,9937%	0,833	62.10	99,3790%
5	1,66	0,57	99,999943%	1,167	233	99,977%
6	2	0,002	99,9999998%	1,500	3,4	99,99966%

El nivel sigma se puede determinar tanto a partir del valor de PPM como del rendimiento del proceso utilizando las siguientes funciones de Microsoft® Excel:

Nivel sigma = −INV.NORM.ESTAND (PPM/10^6)+1,5
Nivel sigma = INV.NORM.ESTAND (RENDIMIENTO)+1,5

Donde el valor −INV.NORM.ESTAND(PPM/10^6) ≡ INV.NORM.ESTAND(RENDIMIENTO) se conoce como valor Z de referencia. La mayoría del software estadístico calcula este valor Z de referencia, que indica el nivel sigma del proceso cuando está centrado. El nivel sigma cuando el proceso está descentrado se puede calcular de manera sencilla simplemente sumando 1,5 al valor Z de referencia.

5. SEIS SIGMA EN UN SERVICIO SANITARIO

La aplicación de la metodología seis sigma en un servicio de atención sanitaria es perfectamente posible si entendemos que, en general, un servicio es un proceso medible sobre el que se puede actuar. A pesar de la complejidad de un servicio como el que presentamos, en seis sigma encontramos pautas de actuación que nos ayudarán a tomar decisiones y a plantear líneas de mejora.

Tomaremos como referencia un Centro de Salud en el que hay un equipo de Atención Primaria compuesto por médicos de familia, pediatras (0-14 años), enfermeras, una matrona, un trabajador social, el servicio de extracción para análisis clínicos y el personal administrativo. La revisión de las encuestas de satisfacción, que se realizan periódicamente, permite detectar los fallos más significativos en la prestación del servicio sanitario. La información recogida

Moreover, sigma quality levels can be computed from both the process fallout (ppm) and yield by using any of the following Microsoft® Excel functions:

SQL = − NORM.S.INV(fallout/10^6)+1.5
SQL = NORM.S.INV(yield)+1.5

Where the value −NORM.S.INV (fallout/10^6)≡NORM.S.INV (yield) can be referred to as *benchmark* Z-score. Most statistical software computes this benchmark Z-score, that indicates the sigma quality level for a process on target, instead of the sigma quality level when the process is operating off target. However, the latter can be easily calculated by just adding 1.5 to the benchmark Z-score.

5. SIX-SIGMA IMPLEMENTATION IN A HEALTHCARE SERVICE

In what follows, an example for the implementation of this methodology in a healthcare service is described. This implementation is perfectly possible bearing in mind that any service is a measurable process that may be monitored and controlled. Though its complexity, six-sigma provides with rules that guide the decision making process and the planning of future actions.

A primary health medical group (that includes family physicians, paediatricians, nurses, one midwife, one social worker, clinical laboratory technologists and administrative staff) will be taken as a reference. The results of the customer satisfaction survey that is carried out periodically in this healthcare centre is taken as the starting point of the improvement process.

en dichas encuestas se estructura en los siguientes apartados:

- Accesibilidad (localización del Centro de Salud y facilidad para conseguir cita)
- Tiempos de espera
- Información recibida
- Trato y atención
- Cuidados: seguimiento y coordinación
- Profesionalidad del personal

Uno de los aspectos que se pretende mejorar, en la medida de lo posible, es "*el tiempo de espera hasta entrar en la consulta*" (los encuestados afirman que el tiempo medio de espera se sitúa alrededor de los 20 minutos). Para ello se decide formar un grupo de trabajo en el que participe todo el personal involucrado: médicos, enfermeras, auxiliares, personal administrativo, etc. Dado que resulta complejo evaluar el tiempo de espera de los pacientes (sobre todo si el tiempo se evalúa a partir de la opinión de los mismos), se opta por identificar y analizar qué factores influyen en el retraso. De este modo, en la ecuación $Y = f(X_{1,}, X_2, \ldots\ldots, X_n)$ la variable dependiente (Y) se asocia con el "*tiempo de espera"*, mientras que las variables independientes $X_{1,}, X_2, \ldots\ldots, X_n$ serán los factores que influyen, en mayor o menor medida, y cuya importancia trataremos de determinar.

De acuerdo con la metodología seis sigma, para estudiar este problema se describen las fases del ciclo DMAMC.

Definir: el objetivo que se persigue es reducir el retraso sobre la hora de cita y programar adecuadamente las citas.

Medir: en esta fase se recopila información relativa a la situación actual, con el fin de abordar y profundizar en el problema. Siempre que sea posible, los datos se preparan para ser tratados estadísticamente.

This survey shows that one of the indicators that should be improved is the time patients spend waiting for their turn (according to those surveyed, the average waiting time is about 20 minutes). Therefore, a working group is formed including all the personnel involved (physicians, nurses, and auxiliary and administrative staff) in order to identify those factors determining the patients' waiting time. Thus, in the equation $Y = f(X_{1,}, X_2, \ldots\ldots, X_n)$ the dependent variable *Y* represents the *wait time*, while the independent variables $X_{1,}, X_2, \ldots\ldots, X_n$ are the determining factors whose influence or importance is being assessed.

According to Six-Sigma methodology, this problem should be addressed by following the DMAIC phases:

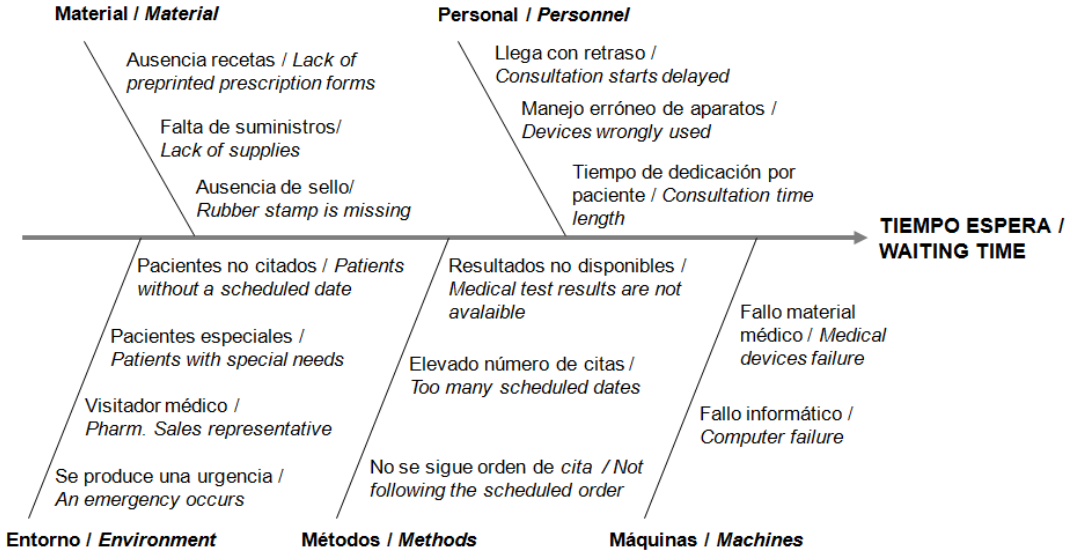
Define: the main goal aimed is to reduce the time the patients spend waiting for their turn by increase the punctuality and improving the scheduling procedure.

Measure: in this phase, information about the current situation of the evaluated process is gathered in order to detect the root causes of the problem. Thus, a brainstorming is used in order to identify the wait time's explaining factors.

Como punto de partida se realiza un brainstorming para identificar las causas del retraso. Mediante un diagrama causa-efecto se agrupan estas causas en cinco grupos: Personal, Métodos, Material, Entorno y Máquinas y se elabora un diagrama causa efecto (diagrama de Ishikawa) como el que aparece en la Figura 2. Todos los motivos que se apuntan en el gráfico son situaciones habituales en el desarrollo de una consulta. Evidentemente la clasificación realizada no es exhaustiva y puede, y debe, revisarse si aparecen nuevas condiciones. Por otra parte, el diagrama causa efecto recoge las causas más importantes del problema, pero debe entenderse que dichas causas son sólo causas potenciales. En consecuencia, es necesario recoger datos para confirmar que las relaciones causa efecto realmente existen.

These factors are clustered in five groups (material, personnel, environment, methods and machines) and are represented in an Ishikawa cause-effect diagram (see Figure 2). All the elements shown in the diagram are usual situations that potentially may cause the problem; however, data should be gathered to confirm that the cause-effect relation actually exists.

Figura 2. Diagrama causa-efecto / *Figure 2. Ishikawa-effect diagram*



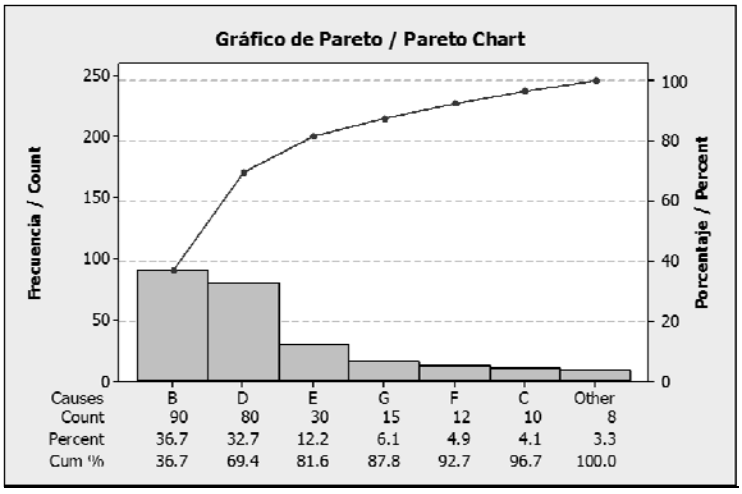
Con el fin de disponer de información fiable, se selecciona una muestra de 20 días y se anotan las incidencias que se producen durante el desarrollo de la consulta, así como la frecuencia de cada una de ellas. También se registra el número de pacientes citados cada día (que se traslada diariamente de la administración a la consulta), y se contabiliza el número de pacientes que asisten a consulta sin tener cita así como el número de pacientes que fallan. A partir de los datos recogidos se realiza un estudio estadístico aplicando las herramientas adecuadas.

Analizar: en primer lugar se construye un gráfico de Pareto (Figura 3) con el fin de detectar las situaciones más relevantes sobre las que centrar la atención. Hay tres tipos de incidencias que acumulan alrededor del 80% de las frecuencias: "*tiempo de dedicación por paciente*", "*llegada de pacientes no citados*" y "*no se respeta el orden de cita*". Evidentemente, los dos últimos motivos están estrechamente relacionados y se condicionan mutuamente.

Therefore, a 20-day sample is selected and the frequency of each of these situations is counted. It is also registered the number of patients dated, the number of patients that attend to the Health Service without a previously scheduled date, and the number of patients that miss a scheduled date. All this information will be analysed by means of the appropriate statistical tools in the next step.

Analyse: first of all, a Pareto chart (see Figure 3) is used to visualise which are the most frequent situations that increase waiting times. The following three situations account for almost 80%: *consultation time length*, *patients without a scheduled date*, and *not following the scheduled patients order*, being the two latter tightly related and mutually conditioned.

Figura 3/*Figure 3*



A: Retraso comienzo consulta / *A: Consultation starts delayed* B: Tiempo dedicación por paciente / *Consultation time length* C: Manejo erróneo de aparatos / *Devices wrongly used* D: Pacientes no citados / *Patients without a scheduled date*	E: No se respeta el orden / *Not following the scheduled patients order* F: Resultados no disponibles / *Medical test results are not available* G: Se produce una urgencia / *An emergency occurs* H: Otros / *Others*

Por otra parte se analizan las variables "*número de pacientes citados*", "*número de pacientes no citados" y "número de pacientes que no asisten"* y se deduce *"total pacientes atendidos"*. Para estas variables se realiza un análisis descriptivo que se recoge en la Tabla·3. De acuerdo con estos datos, se citan diariamente una media de 44,5 pacientes y se atiende una media de 50,8 pacientes por día. Además, ambas medias son representativas pues su dispersión relativa es muy baja (CV<8%).

On the other hand, the variables *number of patients dated* (DATED), *number of patients without a date* (NON-DATED), number of patients that miss a scheduled date (MISSING), as well as the total number of attended patients (ATTENDED) are summarized (see Table 3).

According to this data, the average number of patients dated per day is 44.5 while the average number of patients attended per day is 50.8. Both measures are highly representative since the relative dispersions of both variables are very low (CV<8%).

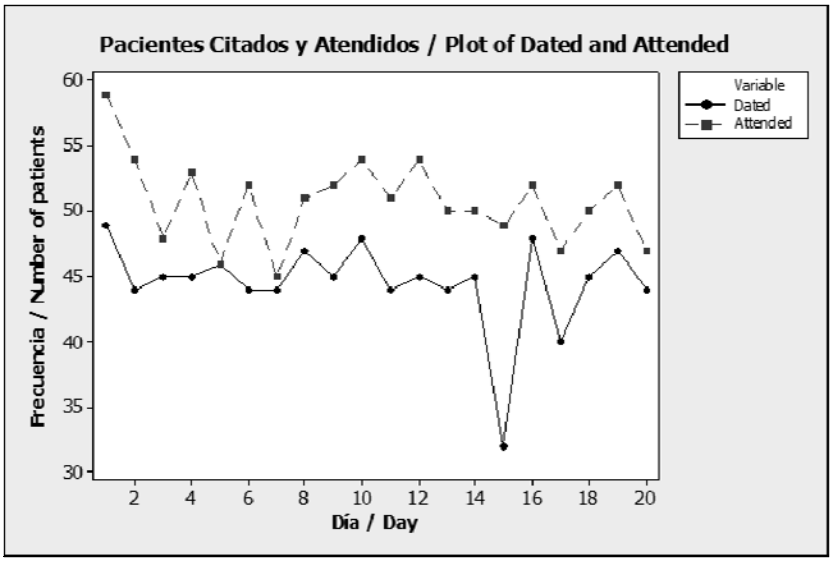
Tabla 3/*Table 3*

Variable	Media *Mean*	Desviación estándar *Standard deviation*	Coeficiente de Variación (CV) *Coefficient of Variation (CV)*	Mínimo *Minimum*	Mediana *Median*	Máximo *Maximum*
Citados *Dated*	44,550	3,546	7,96 %	32,000	45,000	49,000
No citados *Non dated*	9,000	3,584	39,82 %	5,000	9,000	21,000
No asisten *Missing*	2,750	2,489	90,53 %	0,000	2,000	8,000
Atendidos *Attended*	50,800	3,302	6,50 %	45,000	51,000	59,000

Mediante el test de normalidad de Anderson-Darling podemos deducir que la distribución de la variable "*pacientes atendidos diariamente*" se ajusta a una distribución normal (p=0,557). Por el contrario, la hipótesis nula de normalidad para la variable "*pacientes citados*" debe ser rechazada ($p\leq$0,005). La evolución conjunta de ambas variables se recoge en el gráfico de la Figura 4. Dicho gráfico pone de manifiesto que, prácticamente todos los días, acuden a la consulta más pacientes de los citados. Esta circunstancia influye en gran medida en el retraso y en el aumento del tiempo de espera, pues, dependiendo del tratamiento que se de a dichos pacientes, se puede alterar el orden fijado en las citaciones.

The Anderson-Darling normality test is conducted on both variables and the results indicate that the *number of attended patients* (ATTENDED) could be considered as normally distributed (p=0.557). On the contrary, the null hypothesis of normality for the *number of patients dated* (DATED) should be rejected (p<0.005). After plotting both variables in a line graph (see Figure 4), it becomes evident that, almost every day, the number of attended patients is higher than the number of those dated. Depending on how the patients without date are dealt with, this situation may increase greatly the waiting times of those dated.

Figura 4/*Figure 4*



Es obvio, entonces, que la llegada de pacientes sin cita previa (salvo las urgencias), es un factor que afecta al correcto desarrollo de la consulta. Podemos asimilar esta situación a la de un proceso en el que se controla el porcentaje de unidades defectuosas (los pacientes que acuden sin cita) y construir un *gráfico de control p.* De este modo podemos estudiar y controlar si el porcentaje de pacientes no citados (calculado en relación al total de los atendidos cada día) se mantiene dentro de los límites estadísticos (situados a distancia de *k sigma* respecto de la media estimada). Los gráficos de control constituyen una de las herramientas más importantes del control estadístico de procesos. Su finalidad es comprobar si un proceso opera bajo control y diferenciar la variabilidad natural (inherente al proceso) de la variabilidad no natural (debida a causas especiales). Cuando un proceso está en control estadístico es estable y previsible a lo largo del tiempo.

Thus, since the patients who have not a scheduled date interfere in the normal healthcare service, they could be controlled in the same vein as the industrial processes' nonconformities are. Control charts, which are one of the most important tools in statistical process control (SPC), allow to determine whether a process is currently under control, and also to distinguish between variation in a process resulting from common causes and variation resulting from special causes. One type of these charts, the *p-chart*, is used to monitor and control whether the proportion of nonconformities (in our case, the ratio of the number of patients without a date to the total number of attended patients per day) falls between the control limits (drawn at *k* sigma from the mean of the proportions).

En la Figura 5 se representa un gráfico de control *p* con límites de control 2 sigma y 3 sigma. La observación correspondiente al día 15 aparece fuera de los límites de control (situación que puede deberse a la presencia de alguna causa "especial"). Se decide repetir el gráfico recalculando los límites de control una vez eliminada la información correspondiente a ese día. El gráfico resultante indica que el proceso está en control estadístico (Figura 6) y la línea central muestra que un 16,44% de los pacientes atendidos acudieron sin cita previa. Además, los porcentajes diarios representados en el gráfico se mantienen entre los límites de control de forma aleatoria, sin mostrar ningún patrón que sugiera un comportamiento especial en determinados días. Estos límites de control no son constantes, ya que tampoco lo es el número de pacientes atendidos cada día, y nos pueden servir de referencia para monitorizar y vigilar la situación futura.

In Figure 5, a p-chart with 2 sigma and 3 sigma control limits is plotted. There is just one observation, the one that corresponds to day number 15, that falls outside the outer control limits (possibly due to a "special cause").

If these special causes of variation can be identified and eliminated, it is normal to recalculate the chart limits with this observation omitted from the calculations and to scrutinize the revised chart (see Figure 6). The resulting chart indicates that the process is under control and the central line shows that 16.44% attended patients did not have a scheduled date. Moreover, all the data plotted in the chart fall randomly between the control limits and apparently do not follow any trend or special pattern. These control limits, which are not constant since the number of attended patients per day is also variable, can be taken as references to monitor the future situation.

Figura 5. Porcentaje de pacientes no citados
Figure 5. Proportion of patients without date

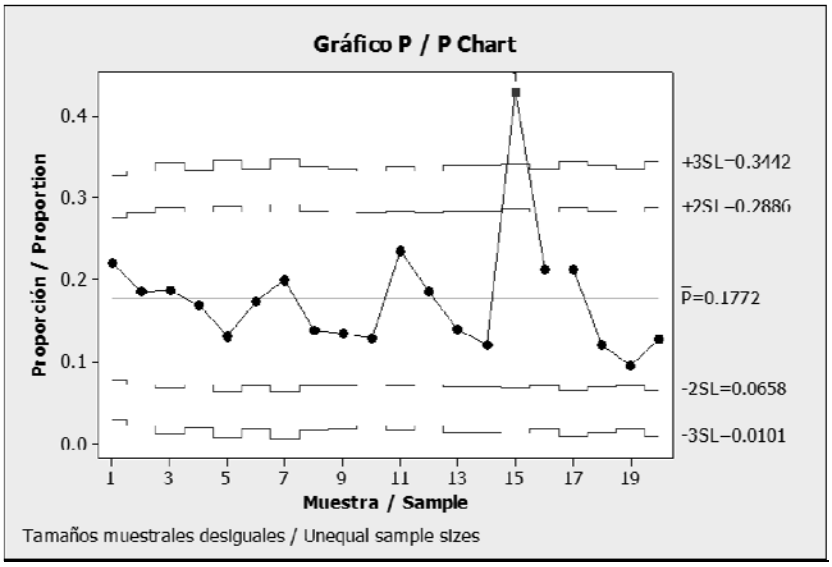
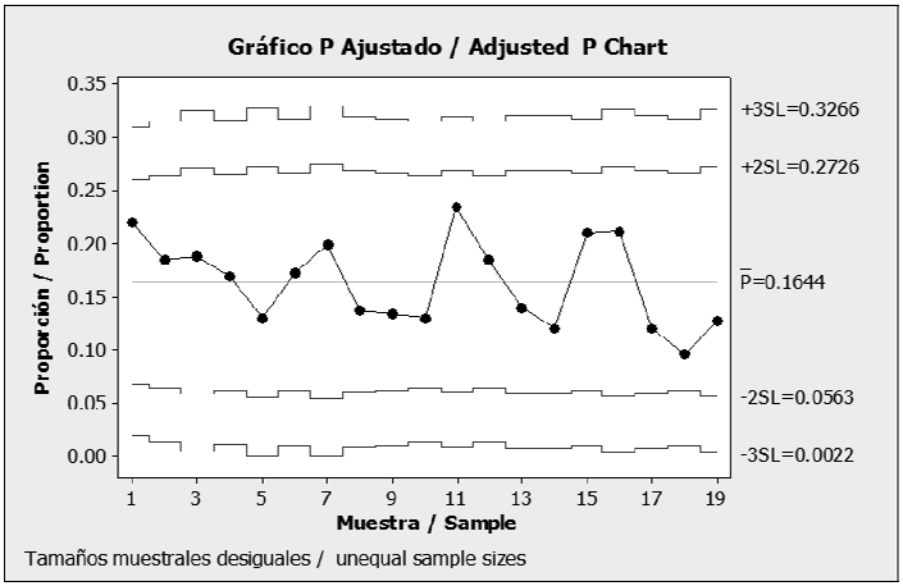


Figura 6. Porcentaje de pacientes no citados (Ajustado)
Figure 6. Proportion of patients without date (ajusted)



Los resultados anteriores ofrecen una visión más clara del problema planteado. El retraso en el tiempo de espera de los pacientes respecto a la hora de cita, se debe más a factores ajenos que a la propia programación de las citas. Es evidente que el número de citas asignadas cada día se programa en función de las horas que el médico debe estar en la consulta, y en el tiempo dedicado por paciente. Sin embargo, dado que la presencia de pacientes no citados supone un porcentaje del 16,44% se debe tener en cuenta este aspecto a la hora de programar las citas. Entendiendo la atención al paciente como un proceso afectado por este factor (que influye en el tiempo de espera), se puede calcular el nivel sigma del proceso. Para ello se estudia la capacidad del proceso correspondiente a una situación binomial y se obtienen los resultados que aparecen en la Tabla 4.

All the previous results provide with a clearer vision of the addressed problem. The length of the waiting times are mainly due to other factors different from the scheduling of dates itself. Obviously, the number of dated scheduled per day are assigned depending on the physician's working time and the consultation time length per patient. However, the planning should also consider the patients who do not have a scheduled date since they account for 16.44% of the total attended.

In order to determine the sigma quality level of the process a capability analysis corresponding to a binomial situation is conducted (see the results in Table 4).

Tabla 4. Análisis de capacidad del proceso binomial: Resumen (nivel de confianza 95%) / *Table 4. Binomial process capability analysis: Summary (confidence level: 95%)*

% Defectuosos / *Nonconformities* (%)	16,44
IC inferior / *Confidence interval (lower limit)*	14,16
IC superior / *Confidence interval (upper limit)*	18,93
Def PPM / *Nonconformities* (ppm)	164426
IC inferior / *Confidence interval (lower limit)*	141597
IC superior / *Confidence interval (upper limit)*	189319
Z del proceso / *Benchmark Z-score*	0,9764
IC inferior / *Confidence interval (lower limit)*	0,8804
IC superior / *Confidence interval (upper limit)*	1,0732

Si el desarrollo de las consultas diarias no se viera afectada por la presencia de los pacientes que acuden sin cita, mejoraría la puntualidad en la atención pues, cuanto menor sea el porcentaje de dichos pacientes, mayor será el nivel sigma del proceso. En el caso que analizamos el nivel sigma se sitúa en 2,48 (valor Z del proceso + 1,5 = 0,9764+1,5), lejos del objetivo de un nivel de calidad 6 sigma.

Mejorar: las acciones de mejora deberían centrarse en establecer prioridades en la atención: en primer lugar y, salvo que se presente una urgencia, prestar atención a los pacientes con cita previa. De este modo estos pacientes serían atendidos a la hora asignada. Aunque esta forma de proceder suele ser habitual en las consultas, también es cierto que, en ocasiones, se mezcla el orden de llegada con el orden de cita, lo que ocasiona problemas y protestas. Por otro lado, es tarea del personal sanitario del centro de Atención Primaria concienciar a los pacientes de la necesidad de pedir cita para mejorar el funcionamiento de las consultas (salvo que la consulta sea urgente).

Should the daily consult schedule be not altered by the presence of patients without a date, the punctuality would increase and the waiting time would be reduced. Thus, the lower is the proportion of patients without a date, the higher the sigma quality level (SQL) of the process will be. For the current data, SQL=Benchmark Z-score+1.5= 2.48, which is pretty far from the target of 6 sigma quality level.

Improve: the improvement actions should be focused on ensuring the fulfilment of the priority rule: unless an emergency occurs, the patients who have a scheduled date will be the first consulted, and those patients without a date will be consulted afterwards. Following this rule should guarantee that the patients dated are consulted on their scheduled time. On the other hand, the primary health healthcare service staff should raise public awareness on the need of having a scheduled date to improve the whole service.

Controlar: se debería realizar un seguimiento para comprobar si mejora la puntualidad en la atención. Posteriormente habría que recoger de nuevo información para estudiar si la mejora ha sido efectiva. Para ello, las herramientas estadísticas vuelven a ser imprescindibles.

6. CONCLUSIONES

La metodología seis sigma es perfectamente aplicable a la asistencia sanitaria, pues constituye una metodología estructurada que combina distintas herramientas estadísticas y proporciona un nueva perspectiva para mejorar la atención al paciente. Mediante seis sigma se identifican los factores que intervienen en un proceso para que, al conocerlos y tratarlos, sea posible diseñar un servicio estándar, siempre que se mantengan las condiciones, y realizar modificaciones o adoptar medidas si estas condiciones cambian.

En este trabajo hemos querido demostrar que las herramientas estadísticas resultan imprescindibles en la implantación de seis sigma, porque aportan un enfoque analítico y contrastado para abordar los problemas que se presentan. Por otro lado, hemos expuesto el razonamiento estadístico en el que se sustenta el cálculo del nivel sigma y su relación con los índices de capacidad.

La puesta en marcha de esta metodología en un servicio sanitario no es una tarea individual, por el contrario requiere la colaboración de todas las personas implicadas en el proceso. Además, tal y como hemos planteado, no es necesario recurrir a complejas herramientas estadísticas para resolver los problemas, pues existen técnicas sencillas muy adecuadas para analizar los datos y extraer conclusiones.

Control: A follow-up procedure should be implemented to monitor the improvement in the patient waiting times: new data should be regularly gathered to check, by means of the appropriate statistical tools, if there has been an actual, significant, and stable improvement in punctuality.

6. CONCLUSIONS

Six-sigma methodology could be perfectly implemented in healthcare services since it is based on a structured methodology that combines different statistical tools and provides with a new approach for the service improvement. By means of six-sigma, the factors that potentially may interfere in a process are identified, analysed and controlled in order to design a standard service under certain conditions and, if these conditions change, make the appropriate changes to readapt the process to the new situation.

The aim of this paper has been to show how the statistical tools are essential when it comes to implement six-sigma methodology since they provide with an analytic and proved approach to any problem. Moreover, the statistical rationale behind the computation of the sigma quality level and its relation to the capability indexes has been also addressed.

The implementation of six-sigma methodology cannot be an individual task; on the contrary it requires the cooperation of all those involved in the patient care process. But, on the other hand, as it has been shown in our example, it does not require the usage of complex statistical

Su utilización en esta propuesta nos ha permitido identificar los factores que intervienen en el aumento del tiempo de espera, para trabajar sobre el que más influye: la llegada de pacientes no citados. La información analizada revela que un 16,44% de los pacientes que se atienden cada día acuden a la consulta sin cita previa, situación que afecta al correcto desarrollo de la misma. Se debería tratar de minimizar el efecto de estos pacientes no citados, bien prestándoles atención al final de la consulta, bien exigiendo disponer de cita (salvo en casos urgentes).

Como reflexión final, si el objetivo que se persigue es mejorar la satisfacción del paciente, hay que determinar cuáles son las características o variables que condicionan esta satisfacción. Pero ello supone un cambio de actitud y un liderazgo que impulse su puesta en marcha y consiga el compromiso de todos. Solo si existe ese compromiso se podrán detectar los fallos y, en consecuencia, plantear alternativas de mejora.

tools since there are simpler techniques that allow analysing the data and getting to conclusions. As an example of their implementation in healthcare service, a specific situation was presented (excessive wait time length), various factors that interfere in the process were identified, then those more influential were indicated (16.44% patients have not a scheduled date) and, finally, improvement actions were proposed (ensuring the fulfilment of the priority rules).

Finally, it could be said that the aspects or variables that influence that satisfaction should be identified in order to improve the patient satisfaction. But doing so demands a change in attitudes and a leadership to foster the implementation of this methodology and to obtain the commitment of everyone involved. Only if that commitment exists, the problems will be detected and, as a consequence, alternative actions to fix them will be proposed.

BIBLIOGRAFÍA/REFERENCES

Antony, F., Kumar, M. y Cho, B. (2007). Six sigma in service organisations: Benefits, challenges and difficulties, common myths empirical observations and success factors. *International Journal of Quality & Reliability Management, 24*(3), 294-311.

Buck, C. (2001). Application of Six Sigma to reduce medical errors. *Proceedings of the 55th Annual Quality Congress of the American Society for Quality*.

Carey, R.G. y Lloyd, R.C. (2001). *Measuring quality improvement in healthcare. A guide to statistical process control applications*. Ed. ASQ.

Chassin, M. (1998). Is health care ready for six sigma quality? *The Milbank Quarterly, 76*(4), 565-591.

Henderson, G.R. (2006). *Six sigma. Quality improvement with minitab*. Chichester: John Wiley & Sons.

Jiménez, V. *et. al*. (2007). Aplicación de las técnicas Lean-Seis Sigma para mejorar la logística sanitaria. *Primer Congreso de Logística y Gestión de la Cadena de Suministro,* Zaragoza.

Mariño Navarrete, H. (2005). ¿Calidad Seis Sigma para el sector salud? *Rev. Vía Salud,* Ene-Mar, *31*, 17-22.

Palacios, J.L. (2003). Función y aplicabilidad del control estadístico de calidad en los servicios sociales. *Cuadernos de Trabajo Social, 16*, 29-48.

Ramírez Valdivia, M.T., Pinto de la Sota Navarro, S.A., Serpell Bley, A. y Enberg, L. (2007). ¿Seis Sigma en hospitales chilenos? *Rev. OIKOS*, *11*(24), 31-46.

Tennant, R., Mohammed, M.A., Coleman, J.J. y Martin, V. (2007). Monitoring patients using control charts: A systematic review. *International Journal for Quality in Health Care; 19*(4), 187-194.

Valdivia Pérez, A., Arteaga Pérez, L., Escortell Mayor, E., Monge Corella, S. y Villares Rodríguez, J.E. (2009). Análisis de las reclamaciones en atención primaria mediante el control estadístico de procesos. *Revista de Calidad Asistencial, 24*(4), 155-161.

Van den Heuvel, J., Doess, R.J.M.M. y Vermaat, M.B. (2004). Six sigma in a Dutch hospital: Does it work in the nursing department?" *Quality and Reliability Engineering International*, *20*, 419-426.

Vuori, H.V. (1996). *El control de calidad en los servicios sanitarios. Conceptos y metodología.* Barcelona: Ed. Masson S.A.

Pecvnia Monográfico 2012, pp. 137-164

EVOLUCIÓN DE LOS PRECIOS DE VIVIENDA Y DE SUELO URBANO EN ESPAÑA / *EVOLUTION IN PRICES OF HOUSING AND URBAN SOIL IN SPAIN*

María Gómez Riocerezo[1]

Ministerio de Fomento

Resumen

En este trabajo se analiza la evolución que han experimentado los precios de la vivienda y del suelo urbano a partir de los datos aportados por las estadísticas disponibles en la Subdirección General de Estudios Económicos y Estadísticas del Ministerio de Fomento. El período temporal considerado comprende desde el año 2004 hasta el primer trimestre de 2012, corroborando el ajuste del que han sido objeto dichos precios desde finales de 2007.

Palabras clave: Vivienda; Suelo urbano; Precios; España.

Abstract

In this paper evolution in prices of housing and urban soil is analysed by using the data included in the statistics available in the *Subdirección General de Estudios Económicos y Estadísticas* (Department of Economic Studies and Statistics) belonging to the *Ministerio de Fomento* (Ministry of Public Works). The study period goes from 2004 to the first quarter of 2012, in order to confirm the adjustment these prices have suffered from the end of 2007.

Keywords: Housing; Urban soil; Prices; Spain.

1. INTRODUCCIÓN

Durante el período 1997-2007 han tenido lugar intensas subidas de los precios de la vivienda en la mayor parte de los países

1. INTRODUCTION

Between 1997 and 2007, most developed countries have seen a considerable increase in housing prices,

[1] Jefe de Área. Subdirección General de Estudios Económicos y Estadísticas. Ministerio de Fomento.

desarrollados, coincidiendo con los tipos de interés más bajos de los últimos cincuenta años y con unas mejorías sensibles en las restantes condiciones de financiación (plazo, relación préstamo/valor). Los precios inmobiliarios más altos han impulsado un crecimiento mayor de las economías más afectadas por dicho proceso, apoyado en el mayor gasto en vivienda y en consumo familiar. El crecimiento citado ha estado acompañado, además, por un fuerte aumento de los niveles de endeudamiento familiar, por lo que la evolución incrementista de los precios de la vivienda ha contribuido a aumentar, en general, el problema de acceso a la misma por parte de los ciudadanos.

España es uno de los países con un acusado crecimiento de los precios de la vivienda. El principal elemento diferencial de España con los restantes países desarrollados en este proceso ha sido la particular intensidad que ha supuesto, hasta fechas bien recientes, el volumen de nueva construcción que en la actualidad, y por razones de todos conocidas, ha sufrido un notable cambio de tendencia.

El ajuste del sector de la vivienda, iniciado en 2007, se precipitó de manera brusca en 2008, como consecuencia, por una parte, de la crisis financiera mundial, y por otra, de factores internos como son: el deterioro de la economía española, la falta de financiación y el agotamiento del modelo de crecimiento basado en la construcción. A principios del año 2010, la información disponible muestra que el mercado de la vivienda ha entrado en una fase de mayor estabilidad.

La Subdirección de Estudios Económicos y Estadísticas del Ministerio de Fomento elabora, entre otras, las Estadísticas de Precios de Vivienda y de Precios de Suelo. Con la información obtenida en dichas

a period when interest rates have been the lowest in the last 50 years and a time when important improvements in other financing conditions (term, relationship loan/price) have also taken place. Higher prices of real estates have fostered a greater growth in the economies more affected by this process, as a result of a higher expenditure in housing purchase and domestic economy. This growth has also meant a considerable increase in households in debt, so the incremental evolution in housing prices has contributed to amplify the problem of people to purchase a dwelling.

Spain is one of the countries which has observed a huge increase in housing prices. The main difference between Spain and the rest of developed countries has been the special focus on new buildings in the process, which nowadays and due to well-known reasons has suffered a considerable change in its trend.

The adjustment in the housing industry, which began in 2007, took place sharply in 2008; on the one hand, as a result of the worldwide financial crisis, and, on the other hand, due to internal factors, such as the deterioration of the Spanish economy, the lack of financing and the end of a growth model based on building. From the beginning of 2010 onwards, the available information shows that the housing market has entered a greater stability stage.

The *Subdirección de Estudios Económicos y Estadísticas* (Department of Economic Studies and Statistics) belonging to the *Ministerio de Fomento* (Ministry of Public Works) prepares, among others, the Statistics of Housing and Soil Prices. The data included in those statistics will allow us to check how the prices in housing

estadísticas podremos comprobar, a continuación, cómo han evolucionando el sector de la vivienda y el sector del suelo desde el año 2004.

Con el fin de conseguir nuestro cometido procederemos, en primer término a exponer brevemente la metodología que utilizan dichas estadísticas, para elaborar los precios de vivienda y de suelo.

2. ESTADÍSTICA DE PRECIOS DE VIVIENDA

La Estadística de Precios de Vivienda tiene como principal objetivo estimar el precio de la vivienda del mercado inmobiliario en España tanto a nivel provincial como regional, lo que nos permitirá conocer su evolución trimestral y anual.

Para dicha estimación, clasificamos la vivienda en función de su acceso al mercado inmobiliario en vivienda libre y vivienda protegida. Se utiliza otro criterio diferenciador como es que la vivienda sea nueva o de segunda mano.

A tal efecto se considera vivienda libre si la vivienda accede libremente al mercado inmobiliario y vivienda protegida si existen restricciones legales sobre precios, superficies y otras cuestiones referentes al mercado hipotecario, señaladas por ley. En cuanto a la vivienda nueva o de segunda mano, será el año de construcción de la vivienda el que posibilite tal clasificación.

2.1. Recogida de la información

La información se recoge por la Asociación Profesional de Sociedades de Valoración (ATASA), entidad sin fines de lucro que agrupa a las empresas de este sector. ATASA coordina la recogida de la

and urban soil sectors have evolved since 2004.

In order to reach our aim, firstly, we will briefly describe the methodology those statistics apply in order to determine the prices of housing and soil.

2. STATISTICS OF HOUSING PRICES

The aim of this Statistics is to estimate the price of housing in the Spanish real estate market, both at a provincial and regional level, which will allow us to know the quarterly and annual evolution.

To obtain that estimation, housing is divided into free-market housing and state-subsidized housing, according to the way of access to the real estate market.

In that sense, free-market housing is considered to be the housing which enters the real estate market in a free way, whereas the housing is considered to be state-subsidized when there are some legal restrictions regarding prices, areas and other issues related to the real estate market. With regard to new and second-hand housing, the criterion used for the different group is the building year.

2.1. Data collection

Information is collected by the *Asociación Profesional de Sociedades de Valoración* (ATASA) (Professional Association of Property Appraisal Companies), which is a non-profit institution. This association coordinates the collection of information and provides the data included in the databases prepared by the property appraisal companies to the responsible

información y facilita los datos procedentes de las bases de datos de las empresas de valoración de inmuebles a la unidad promotora de la Estadística de Precios de Vivienda en soporte magnético, con un diseño de registro preestablecido y en los plazos acordados. Antes de la remisión de la información, ATASA realiza un trabajo previo de consistencia y validación de datos.

Una vez disponible la información en soporte magnético, se lleva a cabo un análisis exhaustivo de los datos, con el fin de detectar y depurar errores. Se utilizan diferentes técnicas de imputación dependiendo de cada tipo de error.

- La unidad de análisis es la vivienda. A tal efecto, se excluyen las viviendas que en el período de referencia estén en fase de construcción o cuyo uso no sea propio para la residencia de un hogar.
- La población objeto de estudio está formada por todas las viviendas que han sido valoradas por las empresas de tasación en un determinado trimestre. La población es variable trimestralmente, puesto que depende del número de valoraciones efectuadas por las empresas de tasación.
- El ámbito geográfico comprende todo el territorio nacional, incluidas las Ciudades Autónomas de Ceuta y Melilla.
- Esta estadística se lleva a cabo con carácter trimestral.

Variables

Los registros que corresponden a cada una de las tasaciones realizadas, tienen la misma estructura y están compuestos por las siguientes variables:

- Tipo de transacción
- Provincia
- Código provincial
- Municipio

unit of the Statistics of Housing Prices, in a magnetic device, with a predetermined registration design and in the agreed deadlines. Before sending the information, ATASA association carries out a previous activity of data consistency and validation.

Once the information is available in a magnetic device, an exhaustive analysis of data is developed, in order to detect and mend mistakes. Different methods are used, depending on the type of mistake.

- The analysis unit is the housing. In this sense, both housing in a building phase in the reference period and housing whose use is not for a dwelling are excluded.

- The study population is the housing that has been evaluated by the property appraisal companies in a certain quarter. The population is different each quarter since it depends on the number of property appraisals developed by the companies.

- The geographical area is the whole country, including Ceuta and Melilla.

- The statistics is quarterly developed.

Variables

The entries corresponding to each one of the property appraisals have the same structure and are composed of the following variables:

- Type of transaction
- Province
- Province code
- Municipality

- Código municipal
- Código postal
- Fecha de tasación
- Valor de tasación
- Superficie construida
- Año de antigüedad de la vivienda

La variable 'tipo de transacción' especifica la tipología de la vivienda. Existen, como ya hemos avanzado, dos tipos de viviendas perfectamente diferenciadas, viviendas libres y viviendas protegidas.

Los códigos provincial y municipal, de acuerdo con la codificación del Instituto Nacional de Estadística, permiten localizar geográficamente la observación dentro del territorio nacional, es decir, su asignación a una Comunidad Autónoma, provincia y municipio. El código postal permitirá profundizar los estudios en pequeñas áreas.

La 'fecha de tasación' especifica el momento de realización de la tasación y determina si la observación está fuera o no del período de referencia establecido.

El 'valor de tasación' es el valor que la empresa tasadora calcula para cada una de las viviendas.

La 'superficie construida' nos indica los metros cuadrados construidos que tiene la vivienda.

La variable 'año de antigüedad' clasifica a las viviendas como de nueva construcción o de segunda mano. En la estadística se definen como viviendas de nueva construcción aquellas que tienen entre 0 y 1 año de antigüedad, mientras que el resto se clasifican como de segunda mano.

- Municipality code
- Postal code
- Date of appraisal
- Value of appraisal
- Built area
- Housing age

The variable 'type of transaction' indicates the type of housing. As it has been said above, there are two different types of housing: free-market and state-subsidized housing.

The province and municipality codes, according to the Spanish Statistics Institute classification, allow us to geographically locate the unit in the Spanish area, that is, its allocation to a region, province and municipality. The postal code will allow us to study smaller areas in depth.

The 'date of property appraisal' indicates the moment when the appraisal is made and allows us to determine whether the observation is included or not in the reference period.

The 'value of property appraisal' is the value that the property appraisal firm calculates for the housing.

The 'built area' is the measure of square metres built in the housing.

The variable 'housing age' allows us to divide housing into new or second-hand. New building housing is defined in the statistics as housing which is between zero and one year old, whereas the rest is grouped under second-hand housing.

2.2. Estimación de precios

La variable 'tipo de transacción' permite un análisis independiente para cada tipo de vivienda, vivienda libre y vivienda protegida. Esta variable clasifica inequívocamente a las viviendas en uno u otro tipo, generando al mismo tiempo, dos bases de datos. El tratamiento dado a cada base de datos es idéntico y se explica a continuación.

2.2.1. Estimación de precios de vivienda libre

Para cada tipo de vivienda *j* (nueva construcción o de segunda mano) en el estrato *k* de la zona geográfica *z* en el período *t*, se define la variable bidimensional (X,Y) donde X representa el valor de mercado de la vivienda en euros, e Y representa la superficie construida expresada en metros cuadrados.

El precio por metro cuadrado de las viviendas de tipo *j* en el estrato *k* de la zona geográfica *z* en el período *t* vendrá definido por el siguiente cociente:

$$P_{jkz}^{t} = \sum_{i} X_{ijkz}^{t} / n_{jkz}^{t} \Big/ \sum_{i} Y_{ijkz}^{t} / n_{jkz}^{t}$$

cuya expresión simplificada es:

$$P_{jkz}^{t} = \sum_{i} X_{ijkz}^{t} \Big/ \sum_{i} Y_{ijkz}^{t} \quad (1)$$

siendo:

P_{jkz}^{t} el precio por metro cuadrado (*euros/m²*) de las viviendas de tipo *j* en el estrato *k* de la zona geográfica *z* y en el período *t*,

$\sum_{i} X_{ijkz}^{t} / n_{jkz}^{t}$ el valor medio de las viviendas de tipo *j* en el estrato *k* de la zona geográfica *z* y en el período *t*,

2.2. Price estimation

The variable 'type of transaction' allows for an independent analysis for each type of housing: free-market and state-subsidized. This variable unequivocally groups housing into one of both types, which at the same time leads to two different databases. The processing of each database is identical and described below.

2.2.1. Free-market housing price estimation

For each type of housing *j* (new building or second-hand one) in the stratum *k* belonging to the geographical area *z* in the period *t*, a two-dimensional variable (X, Y) is defined, where X indicates the housing market value (in Euros) and Y indicates the built area measured in square metres.

The price for housing *j* in the stratum *k* belonging to the geographical area *z* in the period *t* is expressed by the following formulae:

$$P_{jkz}^{t} = \sum_{i} X_{ijkz}^{t} / n_{jkz}^{t} \Big/ \sum_{i} Y_{ijkz}^{t} / n_{jkz}^{t}$$

whose simplified expression is:

$$P_{jkz}^{t} = \sum_{i} X_{ijkz}^{t} \Big/ \sum_{i} Y_{ijkz}^{t} \quad (1)$$

where:

P_{jkz}^{t} is the price per square metre (*Euros/m²*) of housing *j* in the stratum *k* belonging to the geographical area *z* and in the period *t*,

$\sum_{i} X_{ijkz}^{t} / n_{jkz}^{t}$ is the mean value of housing *j* in the stratum *k* belonging to the geographical area *z* and in the period *t*,

$\sum_i Y^t_{ijkz} \big/ n^t_{jkz}$ la superficie media de las viviendas de tipo *j* en el estrato *k* de la zona geográfica *z* y en el período *t*.

El precio por metro cuadrado de la vivienda tipo *j* en la zona *z* en el período *t* se obtiene como media aritmética ponderada de los precios por metro cuadrado en cada uno de los estratos *k* de las viviendas de tipo *j* en la zona geográfica *z* y en el período *t*:

$$P^t_{jz} = \sum_k \alpha^t_{jkz} P^t_{jkz} \quad (2)$$

La ponderación α^t_{jkz} representa el número de viviendas de tipo *j* del estrato *k* de la zona geográfica *z* en el período *t* sobre el número total de viviendas de tipo *j* de la zona geográfica *z* en el período *t*:

$$\alpha^t_{jkz} = n^t_{jkz} \Big/ \sum_k n^t_{jkz} \quad (3)$$

Donde:

n^t_{jkz} es el número de viviendas de tipo *j* del estrato *k* en la zona geográfica *z* en el período *t*,

$\sum_k n^t_{jkz}$ es el número total de viviendas de tipo *j* en la zona *z* en el período *t*.

La suma de los coeficientes de ponderación es la unidad:

$$\sum_k \alpha^t_{jkz} = 1$$

El precio de vivienda libre por metro cuadrado se define como media aritmética ponderada de los precios de cada tipo de vivienda por metro cuadrado:

$$P^t_z = \sum_j q^t_{jz} P^t_{jz} \quad (4)$$

$\sum_i Y^t_{ijkz} \big/ n^t_{jkz}$ is the mean area of housing *j* in the stratum *k* belonging to the geographical area *z* and in the period *t*.

The square metre price of housing *j* in the geographical area *z* in the period *t* is a weighted arithmetic mean of prices in each one of the strata *k* of housing *j* in the geographical area *z* and in the period *t*:

$$P^t_{jz} = \sum_k \alpha^t_{jkz} P^t_{jkz} \quad (2)$$

The weight coefficient α^t_{jkz} indicates the number of housing *j* in the stratum *k* belonging to the geographical area *z* in the period *t*:

$$\alpha^t_{jkz} = n^t_{jkz} \Big/ \sum_k n^t_{jkz} \quad (3)$$

where:

n^t_{jkz} is the number of housing *j* in the stratum *k* belonging to the geographical area *z* in the period *t*,

$\sum_k n^t_{jkz}$ is the total number of housing *j* in the geographical area *z* in the period *t*.

The addition of the weight coefficients equals one:

$$\sum_k \alpha^t_{jkz} = 1$$

The square metre price for free-market housing is defined as the arithmetic mean of square metre prices for each type of housing:

$$P^t_z = \sum_j q^t_{jz} P^t_{jz} \quad (4)$$

Las ponderaciones vienen definidas por el número de viviendas de tipo *j* en la zona geográfica *z* en el período *t* sobre el número total de viviendas en la zona geográfica *z* en el período *t*:

$$q_{jz}^{t} = n_{jz}^{t} \Big/ \sum_{j} n_{jz}^{t} \quad (5)$$

Donde:

n_{jz}^{t} es el número de viviendas de tipo *j* en la zona geográfica *z* en el período *t*,

$\sum_{j} n_{jz}^{t}$ es el número total de viviendas en la zona geográfica *z* en el período *t*.

La suma de los coeficientes de ponderación es la unidad:

$$\sum_{j} q_{jz}^{t} = 1$$

2.2.2. Estimaciones de precios de vivienda protegida

Las consideraciones establecidas para la estimación del precio de la vivienda libre pueden realizarse en la estimación del precio de la vivienda protegida. La única salvedad a considerar es que los valores de la vivienda protegida vienen determinados por el valor máximo legal.

3. ESTADÍSTICA DE PRECIOS DE SUELO

La Estadística de Precios de Suelo tiene como principal objetivo estimar el precio del suelo urbano y urbanizable en España a nivel provincial y regional y conocer su evolución trimestral y anual.

The weight coefficients are defined as the quotient between the number of housing *j* in the geographical area *z* in the period *t* and the total number of housing in the geographical area *z* in the period *t*:

$$q_{jz}^{t} = n_{jz}^{t} \Big/ \sum_{j} n_{jz}^{t} \quad (5)$$

where:

n_{jz}^{t} is the number of housing *j* in the geographical area *z* in the period *t*,

$\sum_{j} n_{jz}^{t}$ is the total number of housing in the geographical area *z* in the period *t*.

The addition of the weight coefficients equals one:

$$\sum_{j} q_{jz}^{t} = 1$$

2.2.2. State-subsidized housing price estimation

All the considerations made for the estimation of price of free-market housing can also be applied to the estimation of price corresponding to state-subsidized housing. The only difference is that there is a maximum legal price for this type of housing.

3. STATISTICS OF SOIL PRICES

The main aim of the Statistics of Soil Prices is to estimate the price of urban soil and building land in Spain both at a provincial and regional level, as well as to analyse its quarterly and annual evolution.

Además facilita el número, el valor y la superficie de las transacciones registradas en un determinado trimestre.

3.1. Recogida de la información

Los datos para la elaboración de esta estadística se reciben del Colegio de Registradores de la Propiedad y Mercantiles de España. El Colegio de Registradores, entidad sin fines de lucro, agrupa a todos los registradores que ejercen su actividad dentro del territorio español.

Esta Entidad coordina la recogida de la información procedente de las bases de datos de los Registros de la Propiedad y facilita datos trimestrales y anuales a la unidad promotora de la Estadística de Precios de Suelo del Ministerio de Fomento, en soporte magnético con un diseño de registro acordado, en función de los objetivos de la estadística y en plazos preestablecidos. Antes de la remisión de la información el Colegio de Registradores realiza un trabajo previo de consistencia y validación de la información.

Una vez recibida la información por el servicio promotor, se lleva a cabo un análisis exhaustivo de los datos, con el fin de detectar y depurar errores. Se utilizan diferentes técnicas de imputación dependiendo de cada tipo de error.

- La unidad de análisis es el terreno de suelo urbano o urbanizable.

- La población objeto de estudio está constituida por todos aquellos terrenos inscritos como urbanos y urbanizables que han sido objeto de transacción en un determinado trimestre. La población de estudio es variable ya que depende del número de transacciones realizadas en el período de tiempo considerado.

Moreover, it provides the number, value and area corresponding to the recorded transactions in a certain quarter.

3.1. Data collection

Data for the preparation of this statistics are provided by the *Colegio de Registradores de la Propiedad y Mercantiles de España* (Spanish Association of Land and Business Companies Registrars). It is a non-profit institution that brings together all the registrars who develop their activity in Spain.

This institution coordinates the collection of information included in the databases from the Land Registries and provides quarterly and annual data to the unit responsible of this Statistics of Soil Prices in the *Ministerio de Fomento*, in a magnetic device and with a predetermined registration design, according to the statistics aims and in the agreed deadlines. Before sending the information, the Association carries out a previous activity of data consistency and validation.

Once the information is available for the responsible unit, an exhaustive analysis of data is developed, in order to detect and mend mistakes. Different methods are used, depending on the type of mistake.

- The analysis unit is the piece of land of urban soil or building land.

- The study population is composed of all pieces of land that have been recorded as urban or building land in a certain quarter. The population varies in each quarter since it depends on the number of transactions developed in the referred period.

- El ámbito geográfico de la estadística comprende todo el territorio nacional, incluidas las Ciudades Autónomas de Ceuta y Melilla.

- La estadística se lleva a cabo con carácter trimestral y el período de referencia de los datos solicitados y tratados es el trimestre en el que se efectuado la transacción, independientemente del trimestre en que se realizó su inscripción en el Registro de la Propiedad.

Variables

Los registros que corresponden a cada uno de los terrenos registrados, tienen la misma estructura y están compuestos por las siguientes variables:

- Identificación de la operación
- Registro de la Propiedad
- Código provincial
- Código municipal
- Titular del bien
- Fecha de transmisión
- Fecha de inscripción en el Registro de la Propiedad
- Superficie
- Precio

La variable 'identificación de la operación' es una variable numérica que identifica cada transmisión.

La variable 'registro de la propiedad' es una variable numérica que identifica al Registro de la Propiedad que ha efectuado la transmisión.

Los códigos provincial y municipal, de acuerdo con la codificación del Instituto Nacional de Estadística, permiten localizar geográficamente la observación dentro del territorio nacional, es decir, su asignación a una Comunidad Autónoma, provincia y municipio.

- The geographical area covers the whole country, including Ceuta and Melilla.

- The statistics is quarterly developed and the reference period for the demanded and processed data is the quarter in which the transaction has been developed, independently of the quarter this was recorded in the Land Registry.

Variables

The entries corresponding to each of the recorded pieces of land have the same structure and are composed of the following variables:

- Identification of the operation
- Land Registry
- Province code
- Municipality code
- Owner of the asset
- Date of transfer
- Date of entry in the Land Registry
- Area
- Price

The variable 'identification of the operation' is a numerical variable which identifies each transfer.

The variable 'Land Registry' is a numerical variable which identifies the Land Registry where the transfer has been made.

The province and municipality codes, according to the Spanish Statistics Institute classification, allow the geographical location in the Spanish area, that is, its allocation to a region, province and municipality.

La variable 'titular del bien' clasifica a los solares según si la persona que lo adquiere es persona física o jurídica.

La 'fecha de transmisión' especifica el momento en el que el solar ha sido transmitido.

La 'fecha de inscripción' se refiere al momento en el que la transmisión se inscribe en el Registro de la Propiedad.

El 'precio' es el valor del solar que ha sido inscrito en el Registro de la Propiedad correspondiente.

La 'superficie' nos indica los metros cuadrados del solar.

Con el objeto de reducir la dispersión de la variable valor de tasación respecto a su media, es preciso construir un determinado número de estratos, donde se agrupen aquellas observaciones cuyos valores tengan una menor dispersión respecto a su media.

La heterogeneidad es una de las características del suelo, de ahí que la agrupación de las observaciones según el tamaño de las parcelas y de los municipios nos permita introducir una mayor homogeneidad en las observaciones correspondientes con el fin de estimar los precios del metro cuadrado en cada uno de los estratos definidos.

En el cuadro siguiente se definen los estratos poblacionales:

The variable 'owner of the asset' classifies land depending on whether the purchaser is a person or a firm.

The variable 'date of transfer' indicates the moment when the piece of land has been transferred.

The 'date of entry' refers to the moment when the transfer is recorded in the Land Registry.

The 'price' is the value of the piece of land which has been recorded in the respective Land Registry.

The 'area' specifies the square metres of the piece of land.

In order to reduce the dispersion of the variable value of appraisal with respect to its mean value, it is necessary to build a certain number of strata, which will group those units whose dispersion to their mean is lower.

Heterogeneity is one of the features of soil. Thus, the grouping of observations according to the size of the pieces of land and municipalities allows us to introduce a higher homogeneity in the corresponding observations, in order to estimate the square metre prices in each of the defined strata.

The table below shows the population strata:

Agrupación municipal / *Municipalities classification*	
Estrato / *Stratum*	Municipio según su población (nº de habitantes) *Municipality population (inhabitants)*
1	< 1.000
2	1.000 – 5.000
3	5.001 – 10.000
4	10.001 – 50.000
5	> 50.000

La estatificación de la variable superficie sería la siguiente: / Regarding the variable 'area', the strata are the following:

Agrupación de la superficie / *Area classification*	
Estrato / *Stratum*	Superficie en m² / *Area (m²)*
1	< 500 m²
2	501 – 1.000 m²
3	1.001 – 5.000 m²
4	5.001 – 10.000 m²
5	> 10.000 m²

3.2. Estimación de precios

En cada estrato de superficie *k* de la zona geográfica *z* en el período *t* se define la variable bidimensional (X,Y) donde X representa el valor del terreno en euros, e Y representa la superficie del terreno en metros cuadrados.

El precio por metro cuadrado de suelo urbano en el estrato de superficie *k* de la zona geográfica *z* en el período *t* vendrá definido por el siguiente cociente:

$$P_{kz}^{t} = \sum_{i} X_{ikz}^{t} / n_{kz}^{t} \Big/ \sum_{i} Y_{ikz}^{t} / n_{kz}^{t}$$

Simplificando, se obtiene

$$P_{kz}^{t} = \sum_{i} X_{ikz}^{t} \Big/ \sum_{i} Y_{ikz}^{t} \quad (6)$$

Donde:

P_{kz}^{t} es el precio por precio por metro cuadrado (*euros/m²*) de los terrenos urbanos en el estrato de superficie *k* de la zona geográfica *z* en el período *t*,

$\sum_{i} X_{ikz}^{t} / n_{kz}^{t}$ es el valor medio de los terrenos urbanos en el estrato de superficie *k* de la zona geográfica *z* en el período *t*,

3.2. Price estimation

In each stratum of area *k* belonging to the geographical area *z* in the period *t*, a two-dimensional variable (X, Y) is defined, where X indicates the piece of land value (in Euros) and Y indicates its area measured in square metres.

The square metre price of urban soil in the stratum of area *k* belonging to the geographical area *z* in the period *t* is expressed by the following formulae:

$$P_{kz}^{t} = \sum_{i} X_{ikz}^{t} / n_{kz}^{t} \Big/ \sum_{i} Y_{ikz}^{t} / n_{kz}^{t}$$

By simplifying the previous expression, this becomes:

$$P_{kz}^{t} = \sum_{i} X_{ikz}^{t} \Big/ \sum_{i} Y_{ikz}^{t} \quad (6)$$

where:

P_{kz}^{t} is the price per square metre (*Euros/m²*) of urban pieces of land in the stratum of area *k* belonging to the geographical area *z* and in the period *t*,

$\sum_{i} X_{ikz}^{t} / n_{kz}^{t}$ is the mean value of urban pieces of land in the stratum of area *k* belonging to the geographical area *z* and in the period *t*,

$\sum_i Y_{ikz}^t / n_{kz}^t$ es la superficie media de los terrenos urbanos en el estrato de superficie *k* de la zona geográfica *z* en el período *t*.

El precio por metro cuadrado de terreno urbano en la zona *z* en el período *t*, se obtiene como media aritmética ponderada de los precios por metro cuadrado en cada uno de los estratos *k* de los terrenos en la zona geográfica *z* y en el período *t*:

$$P_z^t = \sum_k \alpha_{kz}^t P_{kz}^t \quad (7)$$

Donde α_{kz}^t representa el número de transacciones en el estrato de superficie *k* de la zona geográfica *z* en el período *t* sobre el número total de transacciones de la zona geográfica *z* en el período *t*:

$$\alpha_{kz}^t = n_{kz}^t \Big/ \sum_k n_{kz}^t \quad (8)$$

Donde:

n_{kz}^t es el número de transacciones en el estrato de superficie *k* de la zona geográfica *z* en el período *t*,

$\sum_k n_{kz}^t$ es el número total de transacciones de la zona geográfica *z* en el período *t*.

La suma de los coeficientes de ponderación es la unidad:

$$\sum_k \alpha_{kz}^t = 1$$

4. EVOLUCIÓN DEL SECTOR VIVIENDA Y DEL SECTOR SUELO

Una vez visto cómo se elaboran los precios de vivienda y de suelo por parte de la Subdirección de Estudios Económicos y Estadísticas del Ministerio de Fomento, pasaremos

$\sum_i Y_{ikz}^t / n_{kz}^t$ is the mean area of urban pieces of land in the stratum of area *k* belonging to the geographical area *z* and in the period *t*.

The square metre price of urban soil in the geographical area *z* in the period *t* is a weighted arithmetic mean of prices in each of the strata *k* of urban pieces of land in the geographical area *z* and in the period *t*:

$$P_z^t = \sum_k \alpha_{kz}^t P_{kz}^t \quad (7)$$

where α_{kz}^t indicates the number of transactions in the stratum of area *k* belonging to the geographical area *z* and in the period *t* over the total number of transactions in the geographical area *z* in the period *t*:

$$\alpha_{kz}^t = n_{kz}^t \Big/ \sum_k n_{kz}^t \quad (8)$$

where:

n_{kz}^t is the number of transactions in the stratum of area *k* belonging to the geographical area *z* and in the period *t*,

$\sum_k n_{kz}^t$ is the number of transactions in the geographical area *z* and in the period *t*.

The addition of the weight coefficients equals one:

$$\sum_k \alpha_{kz}^t = 1$$

4. EVOLUTION IN HOUSING AND SOIL SECTORS

Once the procedure of estimating prices of housing and soil by the *Subdirección de Estudios Económicos y Estadísticas* belonging to the *Ministerio de Fomento* has been described, we will then analyse

a hacer un análisis de la evolución de dichos precios desde el año 2004.

4.1. Sector vivienda

Respecto a este sector, se diferencia la información en precios de la vivienda y en transacciones inmobiliarias.

En la Tabla 1 se expresan los precios de vivienda libre, vivienda libre nueva, vivienda libre de segunda mano y vivienda protegida desde el año 2004, así como las tasas interanuales correspondientes:

the evolution in those prices starting from 2004.

4.1. Housing sector

Regarding the housing sector, information is differentiated for prices of housing and real estate transactions.

Table 1 shows prices for free-market housing, for both new and second-hand free-market housing, as well as state-subsidized housing, starting from 2004, together with their corresponding year-on-year rates:

Tabla 1.Precios de vivienda (unidad: euros/m²)
Table 1. Housing prices (Euros/m²)

Año *Year*	Trimestre *Quarter*	Vivienda libre *Free-market housing*		Vivienda libre nueva *New free-market housing*		Vivienda libresegunda mano/*Second-hand free-market housing*		Viviendaprotegida *State-subsidized housing*	
		Precio *Price*	Tasas anuales/*Year-on-year rate*	Precio *Price*	Tasas anuales *Year-on-year rate*	Precio *Price*	Tasas anuales *Year-on-year rate*	Precio *Price*	Tasas anuales *Year-on-year rate*
2004	1º	1.456,2		1.431,8		1.465,9			
	2º	1.538,8		1.524,3		1.544,0			
	3º	1.570,8		1.544,2		1.581,4			
	4º	1.618,0		1.618,6		1.613,7			
2005	1º	1.685,4	15,7	1.653,3	15,5	1.698,6	15,9	912,6	
	2º	1.752,8	13,9	1.714,5	12,5	1.768,2	14,5	916,2	
	3º	1.781,5	13,4	1.742,6	12,8	1.800,5	13,9	931,8	
	4º	1.824,3	12,8	1.786,2	10,4	1.843,7	14,3	945,0	
2006	1º	1.887,6	12,0	1.856,7	12,3	1.900,7	11,9	977,4	7,1
	2º	1.942,3	10,8	1.912,9	11,6	1.952,4	10,4	995,6	8,7
	3º	1.956,7	9,8	1.926,1	10,5	1.968,7	9,3	1.000,2	7,3
	4º	1.990,5	9,1	1.957,5	9,6	2.002,6	8,6	1.015,7	7,5
2007	1º	2.024,2	7,2	1.993,9	7,4	2.035,0	7,1	1.020,3	4,4
	2º	2.054,5	5,8	2.028,6	6,0	2.062,1	5,6	1.035,8	4,0
	3º	2.061,2	5,3	2.036,9	5,8	2.068,9	5,1	1.053,6	5,3
	4º	2.085,5	4,8	2.069,9	5,7	2.085,9	4,2	1.071,1	5,5
2008	1º	2.101,4	3,8	2.094,7	5,1	2.102,1	3,3	1.100,0	7,8
	2º	2.095,7	2,0	2.088,1	2,9	2.098,5	1,8	1.112,5	7,4
	3º	2.068,7	0,4	2.071,6	1,7	2.062,1	-0,3	1.123,4	6,6
	4º	2.018,5	-3,2	2.022,0	-2,3	2.007,7	-3,7	1.131,6	5,6
2009	1º	1.958,1	-6,8	1.959,2	-6,5	1.956,8	-6,9	1.112,5	1,1
	2º	1.920,9	-8,3	1.923,9	-7,9	1.917,7	-8,6	1.096,9	-1,4
	3º	1.902,8	-8,0	1.911,1	-7,7	1.890,5	-8,3	1.114,3	-0,8
	4º	1.892,3	-6,3	1.899,6	-6,1	1.878,7	-6,4	1.124,3	-0,6
2010	1º	1.865,7	-4,7	1.869,9	-4,6	1.863,4	-4,8	1.133,4	1,9
	2º	1.848,9	-3,7	1.846,7	-4,0	1.854,9	-3,3	1.141,4	4,1
	3º	1.832,0	-3,7	1.846,7	-3,4	1.827,7	-3,3	1.148,4	3,1
	4º	1.825,5	-3,5	1.829,9	-3,7	1.819,5	-3,2	1.163,5	3,5
2011	1º	1.777,6	-4,7	1.793,8	-4,1	1.764,8	-5,3	1.164,9	2,8
	2º	1.752,1	-5,2	1.770,7	-4,1	1.739,4	-6,2	1.161,7	1,8
	3º	1.729,3	-5,6	1.747,5	-5,4	1.719,0	-5,9	1.158,1	0,8
	4º	1.701,8	-6,8	1.721,1	-5,9	1.691,8	-7,0	1.158,2	-0,5
2012	1º	1.649,3	-7,2	1.671,5	-6,8	1.637,5	-7,2	1.150,8	-1,2

Los resultados que se deducen de esta Tabla son:

– El precio de la vivienda libre en el primer trimestre de 2012 fue de 1.649,3 euros, lo que significa un 3,1% menos que en el cuarto trimestre de 2011, cuyo valor fue de 1.701,8 euros. Respecto al primer trimestre de 2011, dicho valor correspondía a 1.777,6 euros, lo que representa un 7,2% menos.
– El precio de la vivienda libre nueva en el primer trimestre de 2012 fue de 1.671,5 euros, un 2,9% menos que en el cuarto trimestre de 2011, que fue de 1.721,1 euros, y un 6,8% menos que en el primer trimestre de 2011, que fue de 1.793,8 euros.
– El precio de la vivienda libre de segunda mano en el primer trimestre de 2012 fue de 1.637,5 euros, un 3,2% menos que en el cuarto trimestre de 2011, que fue de 1.691,8 euros, y un 7,2% menos que en el primer trimestre de 2011, que fue de 1.764,8 euros.

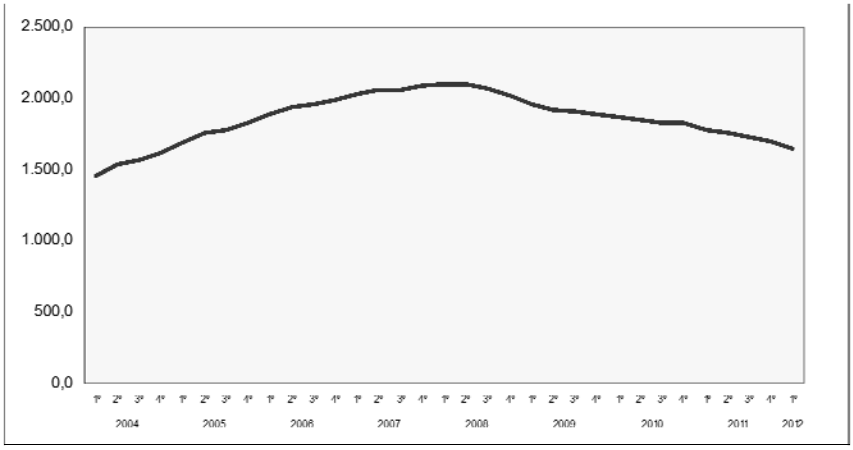
Gráficamente, se observa a continuación, la evolución de los precios de vivienda libre y de las tasas interanuales correspondientes.

From the Table above, we draw the following results:

– The price of free-market housing in the first quarter of 2012 was 1,649.3 Euros, which means a 3.1% lower than the price in the last quarter of 2011, when it was 1,701.8 Euros. Regarding the first quarter of 2011, the price was 1,778.6 Euros, which means that in the same period in 2012 it was a 7.2% lower.
– The price of new free-market housing in the first quarter of 2012 was 1,671.5 Euros, which is a 2.9% lower than in the fourth quarter of 2011, when the price was 1,721.1 Euros, and a 6.8% lower than in the first quarter of 2011, when it was 1,793.8 Euros.
– The price of second hand free-market housing in the first quarter of 2012 was 1,637.5 Euros, which is a 3.2% lower than in the fourth quarter of 2011, when the price was 1,691.8 Euros, and a 7.2% lower than in the first quarter of 2011, when it was 1,764.8 Euros.

Below we graphically show the evolution in prices of free-market housing, as well as their corresponding year-on-year rates.

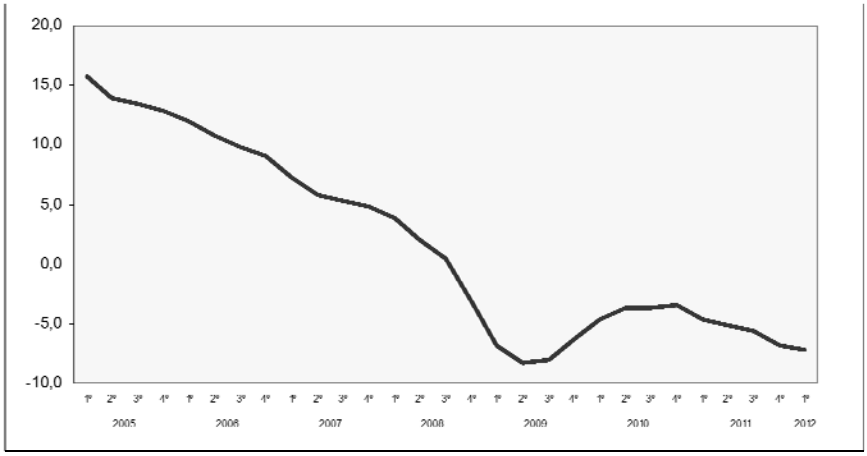
Gráfico 1. Precios de vivienda libre / *Graph 1. Free-market housing prices*



Se comprueba fácilmente que desde el año 2004 hasta principios de 2008 los precios suben, momento a partir del cual los precios empiezan a bajar. Se aprecia idéntico resultado al observar el gráfico de las tasas interanuales correspondientes que aparece a continuación.

It can be easily seen that prices rise from 2004 to the beginning of 2008, when they start to fall. The same trend is observed regarding the year-on-year rates, whose evolution is shown in the graph below.

Gráfico 2. Tasas interanuales de precios de vivienda libre
Graph 2. Year-on-year rates of free-market housing prices



Los datos que se presentan en la Tabla 2 corresponden a las transacciones inmobiliarias, y de ellos se deduce:

– En el primer trimestre de 2012 el número total de transacciones inmobiliarias ha sido de 70.228, lo que significa un 34,6% menos que en el cuarto trimestre de 2011, cuyo número fue de 107.373; asimismo y, respecto al primer trimestre de 2011, cuyo valor fue de 74.455, el porcentaje correspondiente es de un 5,7% menos.
– Por su parte, el número total de transacciones inmobiliarias de vivienda libre nueva en el primer trimestre de 2012 fue de 13.721, un 59,9% menos que en el cuarto trimestre de 2011, que fue de 34.248, y un 34,8% menos que en el primer trimestre de 2011, que fue de 21.055.

Data shown in Table 2 refer to the real estate transactions. From them, the following conclusions are drawn:

– On the one hand, in the first quarter of 2012 the total number of real estate transactions was 70,228, which means it was a 34.6% lower than in the last quarter of 2011, when 107,373 transactions were developed. Likewise, the number was a 5.7% lower than in the first quarter of 2011, when the number of transactions was 74,455.
– On the other hand, the total number of real estate transactions related to new free-market housing in the first quarter of 2012 was 13,721, which is a 59.9% lower than in the last quarter of 2011, when the number was 34,248, and a 34,8% lower than in the first quarter of that year, when it was 21,055.

- En relación al número total de transacciones inmobiliarias de vivienda libre de segunda mano en el primer trimestre de 2012 fue de 50.325, un 15,5% menos que en el cuarto trimestre de 2011, que fue de 59.567, y un 10,1% más que en el primer trimestre de 2011, que fue de 45.725.

- With regard to the total number of real estate transactions related to second-hand free-market housing, in the first quarter of 2012 it was 50,325, which means a 15.5% lower than in the fourth quarter of 2011, when the total number of transactions was 59,567, and a 10.1% higher than in the first quarter of 2011, when it was 45,725.

Tabla 2. Número de transacciones inmobiliarias
Table 2. Number of real estate transactions

Año *Year*	Trimestre *Quarter*	Número total de transacciones/*Total number of transactions*		Vivienda libre nueva *New free-market housing*		Vivienda libre segunda mano/*Second-hand free-market housing*	
		Transacciones *Transactions*	Tasas anuales *Year-on-year rate*	Transacciones *Transactions*	Tasas anuales *Year-on-year rate*	Transacciones *Transactions*	Tasas anuales *Year-on-year rate*
2004	1º	190.442		58.292		120.328	
	2º	223.895		70.827		141.641	
	3º	201.089		66.171		124.720	
	4º	232.964		72.250		147.466	
2005	1º	196.438	3,1	60.950	4,6	125.717	4,5
	2º	241.398	7,8	80.216	13,3	148.922	5,1
	3º	216.333	7,6	82.842	25,2	122.405	-1,9
	4º	247.405	6,2	82.052	13,6	151.585	2,8
2006	1º	233.669	19,0	86.358	41,7	136.363	8,5
	2º	251.649	4,2	94.333	17,6	145.289	-2,4
	3º	221.610	2,4	93.081	12,4	117.109	-4,3
	4º	248.258	0,3	103.384	26,0	132.071	-12,9
2007	1º	230.755	-1,2	89.712	3,9	122.087	-10,5
	2º	227.562	-9,6	93.733	-0,6	116.508	-19,8
	3º	186.504	-15,8	86.836	-6,7	85.044	-27,4
	4º	192.050	-22,6	94.225	-8,9	80.719	-38,9
2008	1º	159.088	-31,1	80.448	-10,3	65.221	-46,6
	2º	157.008	-31,0	80.378	-14,2	61.544	-47,2
	3º	122.949	-34,1	67.294	-22,5	44.363	-47,8
	4º	125.419	-34,7	63.358	-32,8	47.274	-41,4
2009	1º	104.703	-34,2	49.806	-38,1	43.289	-33,6
	2º	119.938	-23,6	54.227	-32,5	52.385	-14,9
	3º	107.534	-12,5	46.250	-31,3	50.491	13,8
	4º	131.544	4,9	51.443	-18,8	66.242	40,1
2010	1º	107.079	2,3	38.837	-22,0	57.879	33,7
	2º	153.164	27,7	59.604	9,9	78.415	49,7
	3º	80.550	-25,1	21.212	-54,1	52.736	4,4
	4º	150.494	14,4	45.562	-11,4	90.143	36,1
2011	1º	74.455	-30,5	21.055	-45,8	45.725	-21,0
	2º	90.756	-40,7	22.546	-62,2	57.268	-27,0
	3º	76.534	-5,0	20.245	-4,6	48.484	-8,1
	4º	107.373	-28,7	34.248	-24,8	59.567	-33,9
2012	1º	70.228	-5,7	13.721	-34,8	50.325	10,1

En los Gráficos 3, 4 y 5 se expresa la tendencia del número de transacciones inmobiliarias para los tres casos considerados. Siendo, a partir de finales de 2006, la fecha en la que se aprecia un descenso del número de transacciones inmobiliarias de vivienda libre nueva y del número total, mientras que en el caso de viviendas de segunda mano se sigue la misma tendencia descendiente hasta principios de 2009, pero a partir de ese punto tiende a estabilizarse.

Graphs 3, 4 and 5 show the trend of the number of real estate transactions for the three aforementioned cases. It is from the end of 2006 when we observe a fall in the total number of real estate transactions and in the number of transactions related to new free-market housing; in the case of second-hand free-market housing, the same descending trend is observed until the beginning of 2009, but from that point onwards, the number of transactions tends to become stabilized.

Gráfico 3. Número total de transacciones inmobiliarias
Graph 3. Number of real estate transactions

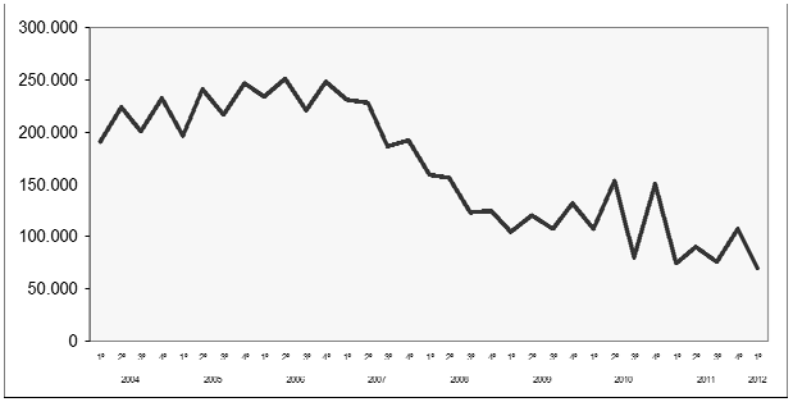


Gráfico 4. Número de transacciones de vivienda libre nueva
Graph 4. Number of new free-market housing transactions

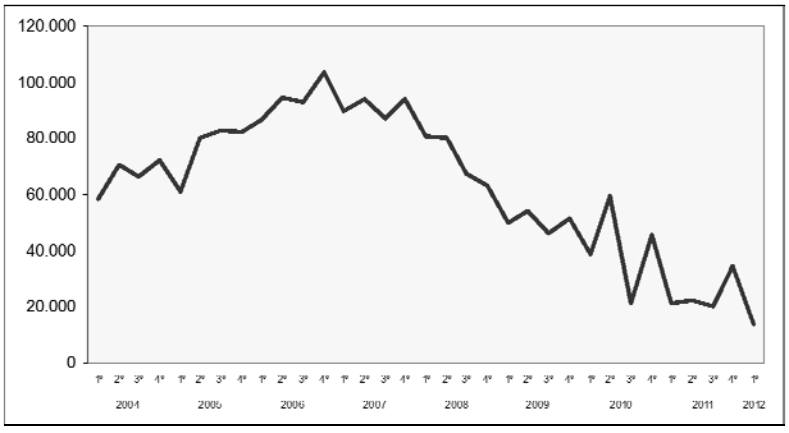
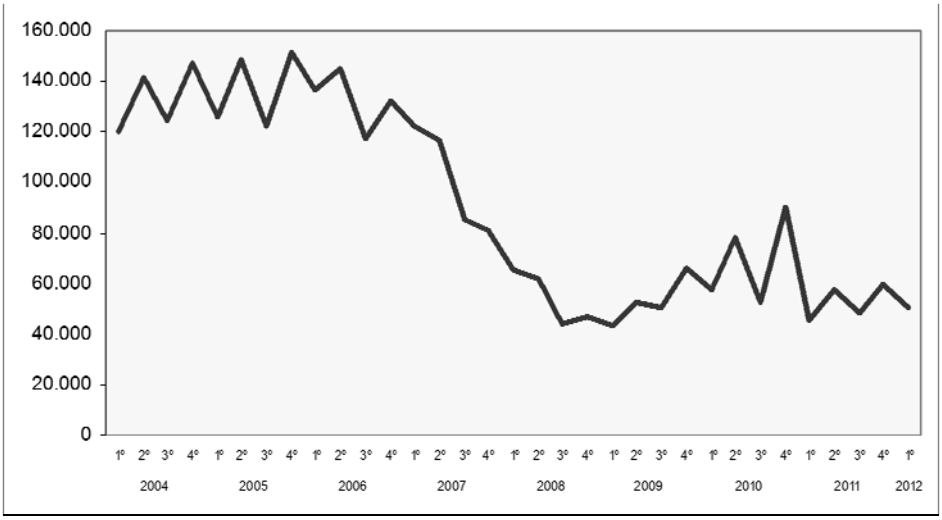


Gráfico 5. Número de transacciones de vivienda libre de segunda mano
Graph 5. Number of second-hand free-market housing transactions



4.2. Sector suelo

Siguiendo un procedimiento similar al sector vivienda, se realiza un análisis de la evolución del sector suelo a partir del año 2004, basado en las siguientes variables: precios del suelo, superficie, valor y transacciones de suelo.

- En primer lugar, se presentan en la Tabla 3 los datos relativos a los precios de suelo urbano en España desde 2004, además de las tasas trimestrales y anuales. Gráficamente estos resultados se pueden observar en los Gráficos 6 y 7.

4.2. Soil sector

Following a similar procedure to the housing sector, we carry out an analysis of the evolution in the soil sector since 2004, in which the variables are: soil prices, area, value and soil transactions.

- First, Table 3 shows the data related to urban soil prices in Spain from 2004, together with quarter-on-quarter and year-on-year rates. From a graphical viewpoint, the results can be observed in Graphs 6 and 7.

Tabla 3. Precios de suelo (unidad: euros/m²) / *Table 3. Soil prices (Euros/m²)*

Año *Year*	Trimestre *Quarter*	Precio suelo *Soil price*	Tasas trimestrales *Quarter-on-quarter rates*	Tasas anuales *Year-on-year rates*
2004	1º	206,5		
	2º	226,4	9,6	
	3º	227,1	0,3	
	4º	247,3	8,9	

2005	1º	258,6	4,6	25,2
	2º	254,5	-1,6	12,4
	3º	263,9	3,7	16,2
	4º	267,3	1,3	8,1
2006	1º	257,4	-3,7	-0,5
	2º	258,4	0,4	1,5
	3º	273,7	5,9	3,7
	4º	284,6	4,0	6,5
2007	1º	271,8	-4,5	5,6
	2º	280,6	3,2	8,6
	3º	285,0	1,6	4,1
	4º	277,0	-2,8	-2,7
2008	1º	250,9	-9,4	-7,7
	2º	258,8	3,1	-7,8
	3º	257,1	-0,7	-9,8
	4º	248,0	-3,5	-10,5
2009	1º	238,8	-3,7	-4,8
	2º	247,6	3,7	-4,3
	3º	237,7	-4,0	-7,5
	4º	232,0	-2,4	-6,5
2010	1º	204,7	-11,8	-14,3
	2º	210,7	2,9	-14,9
	3º	190,8	-9,5	-19,7
	4º	227,7	19,4	-1,9
2011	1º	212,4	-6,7	3,7
	2º	213,9	0,7	1,5
	3º	169,6	-20,7	-11,1
	4º	182,5	7,6	-19,8
2012	1º	177,6	-2,7	-16,4

Gráfico 6. Precios de suelo urbano en España
Graph 6. Urban soil prices in Spain

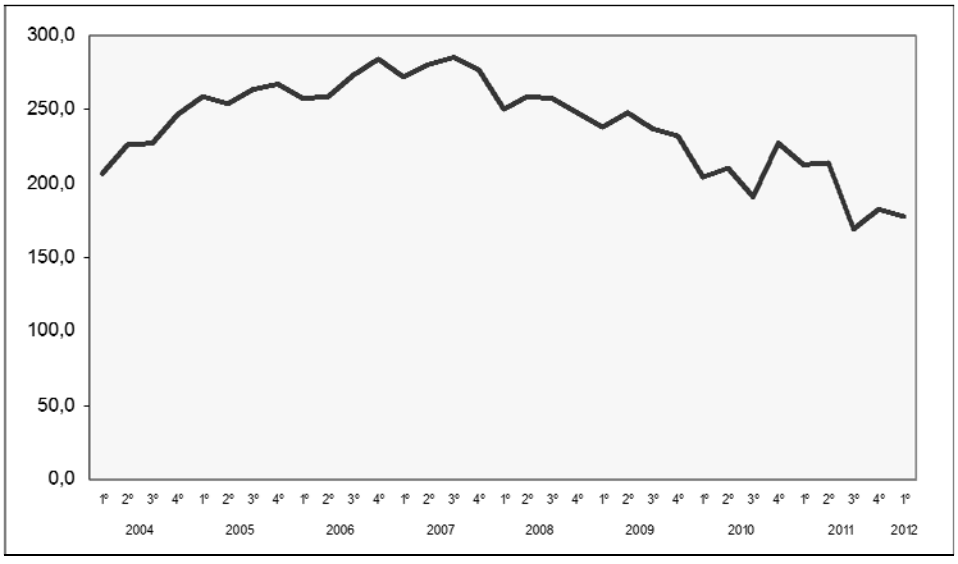
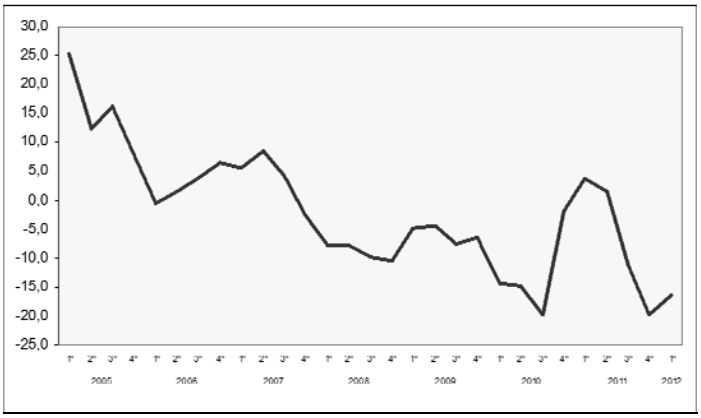


Gráfico 7. Tasas interanuales de precios de suelo
Graph 7. Year-on-year rates of urban soil prices



Se deduce que el precio medio del metro cuadrado de suelo urbano en el primer trimestre de 2012 fue de 177,6 euros, un 16,4% menos que en el mismo trimestre de 2011 y un 2,7% menos que en el cuarto trimestre de 2011.

Como puede comprobarse, es a finales de 2007 cuando se aprecia una tendencia descendente de los precios de suelo urbano en España, aspecto que también se observa en el gráfico correspondiente a las tasas interanuales.

En la Tabla 4 se presentan los datos relativos al valor total de suelo transmitido y a la superficie total de suelo transmitido desde el año 2004. De la misma forma que las variables correspondientes al sector vivienda, en los Gráficos 8 y 9 se puede observar la evolución del valor total y de la superficie total, ambas correspondientes a suelo transmitido.

Se observa que la superficie transmitida en el primer trimestre de 2012 representó 4,6 millones de metros cuadrados, un 15,2% menos que los 5,5 millones de metros cuadrados del primer trimestre de 2011 y un 22,6% menos que el cuarto trimestre de 2011.

It is deduced that the mean square metre price of urban soil in the first quarter of 2012 was 177.6 Euros, which is a 16.4% lower than in the same quarter of 2011 and a 2.7% lower than in the last quarter of that year.

As it can be seen, it is at the end of 2007 when there is a descending trend in urban soil prices in Spain, which can also be seen in Graph 7, related to the quarter-on-quarter rates.

Table 4 shows data concerning the total value of soil being transferred and its total area starting from 2004. In the same sense as the variables corresponding to the housing sector, Graphs 8 and 9 show the evolution in total value and area of transferred soil.

It is observed that the transferred area in the first quarter of 2012 involved 4.6 million of square metres, which is a 15.2% lower than 5.5 million of square metres that were transferred in the first quarter of 2011 and a 22.6% lower than in the last quarter of the same year.

El valor del suelo transmitido fue de 582,9 millones de euros, un 32,9% menos respecto al primer trimestre de 2011 (868,1 millones de euros) y un 34,4% menos que el cuarto trimestre de 2011.

Respecto a la evolución del valor total de suelo urbano transmitido y de la superficie total de suelo urbano transmitido, se aprecia, en ambos casos, un descenso de la tendencia a lo largo de los años.

The value of transferred soil was 582.9 million Euros, which is a 32.9% lower than the value for the first quarter of 2011 (868.1 million Euros) and a 34.4% lower than the last quarter of this year.

Regarding the evolution of the total transferred urban soil value and the total area of the transferred urban soil, it is observed, in both cases, a decrease in the trend with the passing of time.

Tabla 4. Superficie y valor de suelo / *Table 4. Soil area and value*

Año *Year*	Trimestre *Quarter*	Valor / *Value*			Superficie / *Areas*		
		Valor (miles de euros) *Value (thousand euros)*	Tasas trimestrales *Quarter-on-quarter rates*	Tasas anuales *Year-on-year rates*	Superficie miles de m² *Areas thousand m²*	Tasas trimestrales *Quarter-on-quarter rates*	Tasas anuales *Year-on-year rates*
2004	1º	5368707,6			28465,2		
	2º	5678322,7	5,8		30388,2	6,8	
	3º	4882934,3	-14,0		24913,6	-18,0	
	4º	7085241,3	45,1		33071,5	32,7	
2005	1º	5546449,6	-21,7	3,3	23829,7	-27,9	-16,3
	2º	6540225,7	17,9	15,2	33505,1	40,6	10,3
	3º	5926083,4	-9,4	21,4	25581,3	-23,6	2,7
	4º	4469969,7	-24,6	-36,9	20007,9	-21,8	-39,5
2006	1º	4186788,5	-6,3	-24,5	18917,8	-5,4	-20,6
	2º	3890382,7	-7,1	-40,5	23384,9	23,6	-30,2
	3º	3371871,7	-13,3	-43,1	16369,4	-30,0	-36,0
	4º	4756144,3	41,1	6,4	22083,2	34,9	10,4
2007	1º	5330805,9	12,1	27,3	19328,1	-12,5	2,2
	2º	6368256,3	19,5	63,7	20579,5	6,5	-12,0
	3º	4694289,0	-26,3	39,2	15730,3	-23,6	-3,9
	4º	4210696,5	-10,3	-11,5	13810,4	-12,2	-37,5
2008	1º	4245916,7	0,8	-20,4	14930,1	8,1	-22,8
	2º	2707490,6	-36,2	-57,5	10101,8	-32,3	-50,9
	3º	3836866,6	41,7	-18,3	12714,5	25,9	-19,2
	4º	3052507,1	-20,4	-27,5	11001,9	-13,5	-20,3
2009	1º	2571122,4	-15,8	-39,4	8716,7	-20,8	-41,6
	2º	2662399,7	3,6	-1,7	9687,5	11,1	-4,1
	3º	2036690,4	-23,5	-46,9	6107,2	-37,0	-52,0
	4º	1214462,0	-40,4	-60,2	6422,6	5,2	-41,6
2010	1º	1010482,0	-16,8	-60,7	5536,4	-13,8	-36,5
	2º	1061295,5	5,0	-60,1	6817,6	23,1	-29,6
	3º	832552,0	-21,6	-59,1	5428,5	-20,4	-11,1
	4º	1066559,0	28,1	-12,2	5895,1	8,6	-8,2
2011	1º	868103,3	-18,6	-14,1	5467,5	-7,3	-1,2
	2º	1064003,2	22,6	0,3	7150,9	30,8	4,9
	3º	683369,8	-35,8	-17,9	5121,5	-28,4	-5,7
	4º	888333,8	30	-16,7	5991,0	17	1,6
2012	1º	582863,2	-34,4	-32,9	4634,9	-22,6	-15,2

Gráfico 8. Valor de suelo / *Graph 8. Soil value*

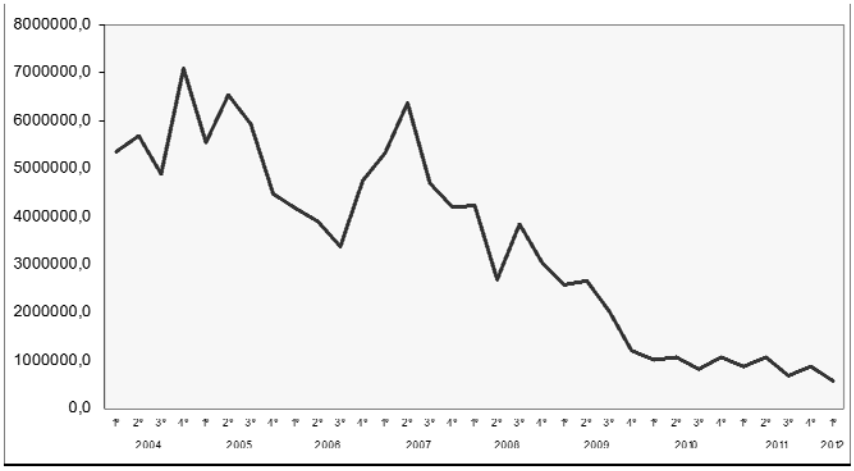
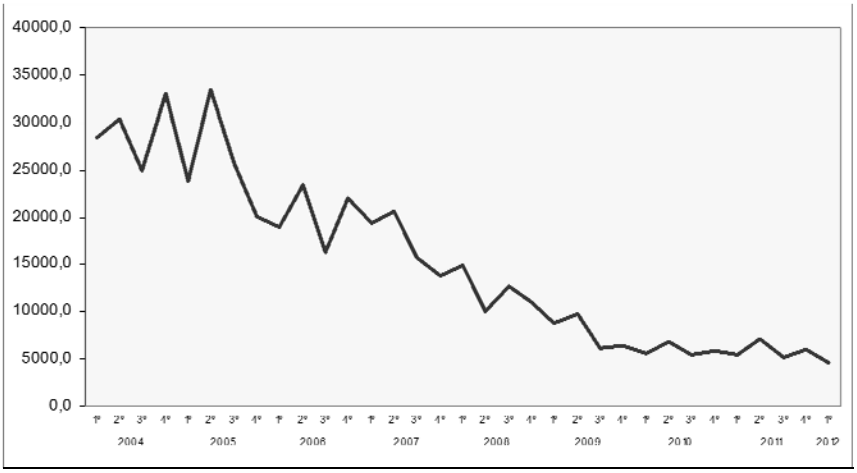


Gráfico 9. Superficie de suelo / *Graph 9. Soil area*



Los datos correspondientes a la última variable considerada para el sector suelo se presentan en la Tabla 5 y se refieren al número total de transacciones de suelo urbano realizadas en España desde 2004. La representación gráfica de dicha variable y la de las tasas interanuales se observa en los Gráficos 10 y 11.

The data corresponding to the last analysed variable for the soil sector are shown in Table 5 and refer to the total number of soil transactions developed in Spain starting from 2004. This variable is graphically presented in Graph 10, whereas Graph 11 shows the evolution in the year-on-year rates.

Tabla 5. Número de transacciones de suelo
Table 5. Number of soil transactions

Año *Year*	Trimestre *Quarter*	Transacciones del suelo *Soil transactions*	Tasas trimestrales *Quarter-on-quarter rates*	Tasas anuales *Year-on-year rates*
2004	1º	22106		
	2º	23030	4,2	
	3º	18206	-20,9	
	4º	23556	29,4	
2005	1º	19216	-18,4	-13,1
	2º	23324	21,4	1,3
	3º	18012	-22,8	-1,1
	4º	14986	-16,8	-36,4
2006	1º	14743	-1,6	-23,3
	2º	15153	2,8	-35,0
	3º	12489	-17,6	-30,7
	4º	16094	28,9	7,4
2007	1º	12819	-20,3	-13,1
	2º	12847	0,2	-15,2
	3º	10361	-19,4	-17,0
	4º	10410	0,5	-35,3
2008	1º	8892	-14,6	-30,6
	2º	7956	-10,5	-38,1
	3º	6722	-15,5	-35,1
	4º	7909	17,7	-24,0
2009	1º	5589	-29,3	-37,1
	2º	5894	5,5	-25,9
	3º	5023	-14,8	-25,3
	4º	6150	22,4	-22,2
2010	1º	5758	-6,4	3,0
	2º	5997	4,2	1,7
	3º	4279	-28,60	-14,8
	4º	5186	21,20	-15,7
2011	1º	4453	-14,10	-22,7
	2º	4377	-1,70	-27
	3º	3546	-19,00	-17,1
	4º	4420	24,60	-14,8
2012	1º	3598	-18,60	-19,2

Gráfico 10. Número de transacciones de suelo
Graph 10. Number of soil transactions

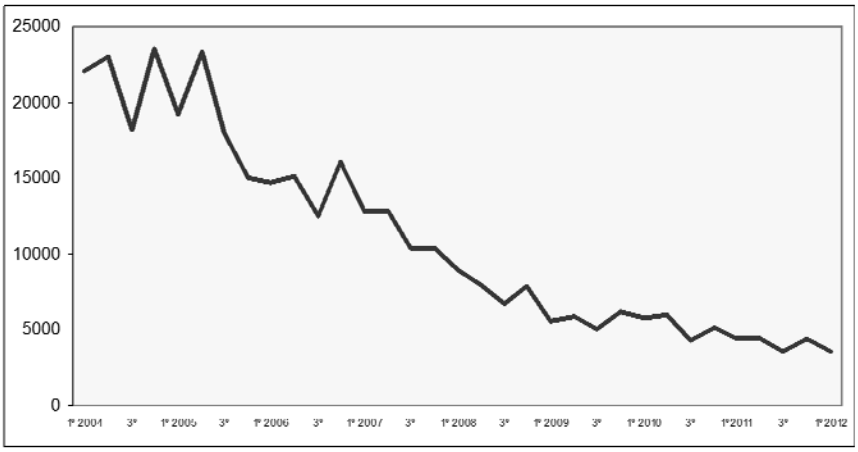
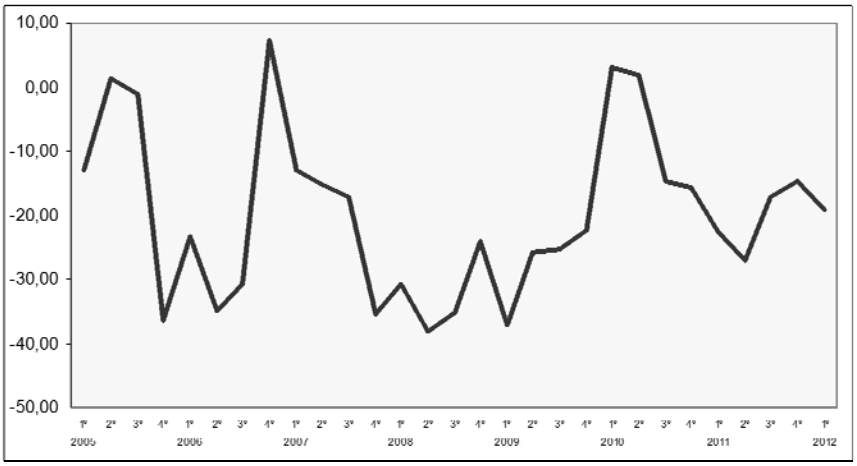


Gráfico 11. Tasas interanuales de las transacciones de suelo
Graph 11. Year-on-year rates of soil transactions



Se puede deducir que el número de transacciones realizadas en el primer trimestre de 2012 fue de 3.598, un 18,6% menos que las realizadas en el cuarto trimestre de 2011, que fue de 4.420, y un 19,2% menos que las que se realizaron en el primer trimestre de 2011, donde se transmitieron 4.453 solares.

It can be inferred that the number of transactions developed in the first quarter of 2012 was 3,598, which means an 18.6% lower than the number in the last quarter of 2011, when it was 4,420, and a 19.2% lower than the number of transactions developed in the first quarter, when 4,453 pieces of land were transferred.

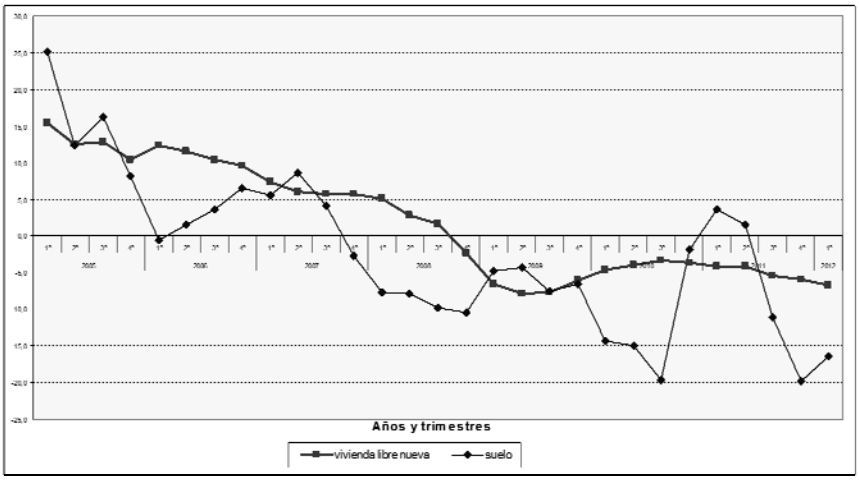
En el Gráfico 10 se aprecia claramente un descenso de la tendencia del número de transacciones.

En el siguiente Gráfico se observa la evolución de las tasas interanuales del precio del suelo y el de la vivienda de obra nueva:

In Graph 10 it is clearly observed that the trend in the number of transactions has decreased.

The graph below shows the evolution in the year-on-year rates for soil and new housing prices.

Gráfico 12. Tasas interanuales de precios de vivienda y de precios de suelo / *Graph 12. Year-on-year rates of housing and soil prices*



5. CONCLUSIONES

El objetivo del análisis que hemos llevado a cabo ha sido comprobar la evolución de la que han sido objeto tanto el sector de la vivienda como el sector del suelo, en los últimos años, a partir de la información que se obtiene de las estadísticas de precios de vivienda y de precios de suelo, elaboradas por la Subdirección de Estudios Económicos y Estadísticas del Ministerio de Fomento.

Dicha información, nos ha permitido corroborar que el ajuste, tanto del sector de la vivienda como del sector del suelo, se inicia a finales de 2007 y principios de 2008.

5. CONCLUSIONS

The aim of this study has been to analyse the evolution in both the housing and soil sectors in recent years, by using the information available in the statistics of housing and soil prices, which are prepared by the *Subdirección de Estudios Económicos y Estadísticas* belonging to the *Ministerio de Fomento*.

That information has allowed us to confirm that the adjustment in both the housing and soil sectors started at the end of 2007 and the beginning of 2008.

Desde el año 2004 hasta principios de 2008 los precios de vivienda libre son objeto de subida, mientras que a partir de esta última fecha se comprueba que los precios empiezan a bajar. En cuanto a los precios de suelo, es a finales de 2007 cuando se aprecia un descenso de la tendencia.

Asimismo, de los datos analizados en el trabajo se desprende que a partir de finales de 2006 se produce un descenso de la tendencia en el número total de transacciones inmobiliarias, lo que también se observa en lo que se refiere al número de transacciones de suelo, en las que hay un claro descenso de la tendencia a partir de finales de 2006.

From 2004 to the beginning of 2008, free-market housing prices increased, whereas from 2008 prices have started to fall. Regarding soil prices, it is the end of 2007 the one which shows a decrease in their trend.

Furthermore, from the analysed data in this paper it is deduced that from the end of 2006 onwards the trend of the total number of real estate transactions has decreased, which is also observed regarding the number of soil transactions, in which there is a clear fall in their trend starting from the end of 2006.

6. BIBLIOGRAFÍA/REFERENCES

BBVA (2010). *Situación España*. Servicio de Estudios Económicos, segundo trimestre.

Bueno Miralles, B. (2009). *El poliedro de la vivienda. Estudio de la vivienda protegida de acuerdo con el Plan Estatal 2009-2012*. Madrid: La Ley.

Carbajo Nogal, C. (2010). *La Fiscalidad de la vivienda en España: análisis de sus efectos jurídicos y económicos* (Tesis Doctoral). Universidad de Burgos.

García Delgado, J.L. y Jiménez, J.C. (1999). *Un siglo de España. La economía.* Madrid: Marcial Pons.

García Montalvo, J. y Mas, M. (2000). *La vivienda y el sector de la construcción en España*. Valencia: Caja de Ahorros del Mediterráneo.

Juárez, M. (Dir.) (1994). *Vivienda*. V Informe sociológico sobre la situación de la vivienda en España. Sociedad para todos en el año 2000 (Vol. II, Cap. 10). Fundación Focea (Fomento de Estudios Sociales y de Sociología Aplicada).

Ministerio de Fomento (2000). *Atlas estadístico de las áreas urbanas en España*. Serie Monografías.

Ministerio de Vivienda (2006). *Estudio sobre el mercado de la vivienda en España. Previsiones 2007-2009*.

Ministerio de Vivienda (2006). *Estudio sobre la situación y perspectivas futuras en el sector inmobiliario en España*.

Ministerio de Vivienda (2009). *Estudio sobre el stock de viviendas a 31 de enero de 2008*.

Ministerio de Vivienda (2010). *Estadística de precios de suelo*.

Ministerio de Vivienda (2010). *Estadística de precios de vivienda*.

Ministerio de Vivienda (2010). *Estadística de transacciones inmobiliarias*.

Ponce Solé, J. y Sibina Tomás, D. (Coord.), (2008). *El derecho de la vivienda en el siglo XXI: Sus relaciones con la ordenación del territorio y el urbanismo. Con un análisis específico de la Ley catalana 18/2007, de 28 de diciembre, del derecho a la vivienda, en su contexto español, europeo e internacional*. Madrid: Marcial Pons.

Rodríguez Alonso, R. (2005). La política de vivienda en España desde la perspectiva de otros modelos europeos. *Boletín CF+S*, 29/30. http://habitat.aq.upm.es/boletin/n29/arrod2.html.

Servicio de Estudios Económicos del BBVA. *Situación Inmobiliaria*. Diciembre 2008. http://www.bbvaresearch.com/KETD/fbin/mult/ESIES_0812_Situacioninmobiliaria_24_tcm346-183823.pdf?ts=2012014.

Tinaut Elorza, J.J. (2006). Desarrollos recientes de la política estatal de vivienda en España: el Plan 2005-2008. *Papeles de Economía Española*, 109, 273-290.

Pecvnia Monográfico 2012, pp. 165-178

CONSTRUCCIÓN DE TABLAS DE VIDA DINÁMICAS PARA UNO O DOS SEXOS / *CONSTRUCTION OF UNISEX OR SEX-DISTINCT DYNAMIC LIFE TABLES*

Ewa Dylewska[1]
ewa@dylewska.com
Metlife Amplico, Polska

Mª Purificación Galindo Villardón[2]
p.galindo@usal.es
Universidad de Salamanca

Resumen

Mientras que las tablas de vida tradicionales describen la mortalidad actual en un determinado periodo de tiempo, las tablas de vida dinámicas permiten una proyección de mortalidad futura. Además de la edad y el sexo, las tablas de vida dinámicas tienen también una tercera dimensión, que es el tiempo. Por lo tanto, permiten observar cambios en la mortalidad que resultan no solamente de cambios en la edad sino también de cambios que aparecen a lo largo del tiempo. Esto se refiere sobre todo a la tendencia de disminución de riesgo de mortalidad. Las tablas de vida dinámicas son, por lo tanto, muy útiles en la construcción de seguros de vida de larga duración y planes de pensiones. También pueden ser aplicadas en la construcción de tablas unisex (relacionado con el dictamen C-236/09 – Test -Achats).
El objetivo del estudio es comprobar la diferencia en la esperanza de vida a la edad *x* en España calculada utilizando las tablas de vida estáticas (tradicionales) y dinámicas, para uno o ambos sexos. Este fin se realiza aplicando el modelo de Lee-Carter para pronosticar la mortalidad futura.

Palabras clave: Tablas de vida dinámicas; Modelo de Lee-Carter; Proyección de mortalidad.

Abstract

Traditional life tables describe a level of mortality at one and defined period of time whereas dynamic life tables allow for projection of future mortality. Apart from age and gender, dynamic life tables also have a third dimension, which is time. In that way, it is possible to observe changes of mortality over the years. This is especially reflected in a mortality reduction trend. Dynamic life

[1] Metlife Amplico, Polska.
[2] Departamento de Estadística, Facultad de Medicina, Universidad de Salamanca, Campus de Miguel de Unamuno, c/ Alfonso X El Sabio, s/n, 37007-Salamanca.

tables are therefore very useful in pricing long-term life contracts and especially in pricing annuities. Moreover, dynamic life tables can also be used in constructing unisex life tables (in relation with decision C-236/09 – Test-Achats).
The purpose of this paper is to demonstrate differences in life expectancy at age *x* in Spain calculated by using static (traditional) and dynamic life tables, both unisex and sex-distinct. Mortality projection is done through the application of the Lee-Carter model.

Keywords: Dynamic life tables; Lee-Carter model; Mortality projection.

1. INTRODUCCIÓN: IDEA DE TABLAS DE VIDA DINÁMICAS

Las tablas de vida dinámicas representan las estimaciones del valor de la probabilidad de muerte correspondiente a la edad x en el año t $(\hat{q}_{xt})$ (Dylewska y Galindo, 2011).

Este tipo de tablas de vida se basan en unas tablas estáticas a las que se aplica una función de ajuste que representa el cambio en el riesgo de mortalidad a la edad x, que tiene lugar durante t años. Las tablas de vida dinámicas asumen que la supervivencia de una cohorte depende del tiempo físico pues el requisito de estacionariedad, en su caso, no se verifica. En otras palabras, las tablas dinámicas se construyen con el objetivo de representar una proyección de la mortalidad de una cohorte, cuyo riesgo de mortalidad depende no solamente de la edad (*x*) sino también del tiempo físico (*t*).

Existen varios modelos que permiten calcular la disminución del riesgo de mortalidad año por año y de este modo construir tablas de vida dinámicas. Entre los modelos más utilizados en la práctica actuarial se encuentran el método de Factor de Reducción (*Reduction Factor*) propuesto por Continuous Mortality Investigation Committee (1999) y el Modelo de Lee-Carter (Lee y Carter, 1992), con modificaciones de numerosos autores.

1. INTRODUCTION: THE IDEA OF DYNAMIC LIFE TABLES

Dynamic life tables represent the estimation of the probability of death at age x and time t $(\hat{q}_{xt})$ (Dylewska and Galindo, 2011). This type of life tables are based on static (traditional) life tables and on a mathematical model which describes the change in mortality risk at age x, and which takes place over t years. Dynamic life tables represent the biometric model of a population, as they assume the survival of a cohort also depends on time and assumptions regarding stationarity of the model cannot be satisfied. The main aim of constructing dynamic life tables is to represent a projection of the mortality of a cohort whose risk of death depends not only on age (*x*) but also on time (*t*).

There are various models which allow us to calculate the improvement in mortality year by year and therefore can be used to build dynamic life tables. The most popular and currently applied methods in actuarial practice are the Reduction Factor method proposed by the Continuous Mortality Investigation Committee (1999) and the Lee-Carter model (Lee and Carter, 1992) with some modifications proposed by various authors. Detailed review of methods and models currently applied was presented *inter alia* by Booth (2006).

Una detallada revisión de los métodos y modelos aplicados en la actualidad para el ajuste de la mortalidad fue presentada, por ejemplo, por Booth (2006).

Una tabla dinámica tiene dos o más dimensiones, dependiendo del número de parámetros del modelo utilizados para su creación. El estudio de Debón, Martínez y Montes (2007) presenta las tablas dinámicas creadas con la aplicación de los modelos dependientes de edad y cohorte (*age-cohort model*), edad y periodo (*age-period model*), edad, periodo y cohorte (*age-period-cohort model*) o edad y deriva (*age-drift model*). Normalmente las filas de una tabla dinámica representan la edad (*x*) y las columnas representan años consecutivos, o bien el número de años que han transcurrido desde el momento 0 (*t*). Un esquema de la tabla de mortalidad dinámica se presenta a continuación:

A dynamic life table has two or more dimensions, depending on the number of parameters used in the model. In the study of Debón, Martínez and Montes (2007) different dynamic life tables are shown: based on age and cohort (age-cohort model), age and period (age-period model), age, period and cohort (age-period-cohort model) or age and drift (age-drift model). Generally, across lines of the dynamic life table represent age (*x*) and down columns represent consecutive years or number of years (*t*) starting from a defined period 0. A sample of a dynamic life table is provided below:

Tabla 1. Esquema de tabla de mortalidad dinámica
Table 1. Sample of a dynamic life table

x | t	0	1	2	3	...	n
0	$q_{0,0}$	$q_{0,1}$	$q_{0,2}$	$q_{0,3}$	…	$q_{0,n}$
1	$q_{1,0}$	$q_{1,1}$	$q_{1,2}$	$q_{1,3}$	…	$q_{1,n}$
2	$q_{2,0}$	$q_{2,1}$	$q_{2,2}$	$q_{2,3}$	…	$q_{2,n}$
…	…	…	…	…	…	…
110	$q_{110,0}$	$q_{110,1}$	$q_{110,2}$	$q_{110,3}$	…	$q_{110,n}$

Cada diagonal de la matriz corresponde a una cohorte, es decir, a un grupo de individuos nacidos en el mismo año. La mortalidad a la misma edad de varios cohortes es distinta y normalmente disminuye a medida que aumenta el año de nacimiento. A veces ocurre que el cambio de mortalidad de unas cohortes o grupos de cohortes es muy distinto a otros. Este efecto se denomina el efecto de la cohorte (*cohort effect*).

Each diagonal of the matrix represent one cohort, which is a group of individuals born in the same year. Mortality at the same age of different cohorts is not the same and generally the lower it is, the older the cohort grows. It may occur that change of mortality of some cohorts or groups of cohorts varies a lot from one another. This effect is known as cohort effect.

La elección de una columna de la matriz permite crear una tabla de vida estática pronosticada para el horizonte de *t* años. Además, en la práctica actuarial puede resultar útil crear una tabla abreviada que recoja resultados para una sola cohorte (una diagonal de la tabla dinámica entera). Entonces, la tabla dinámica que describe la mortalidad de una cohorte a la cual pertenece un individuo de edad x_d, asumiendo una disminución de la mortalidad cada año *t*, tendrá la forma siguiente:

The selection of one column of the matrix allows us to create a static life table projected for *t* years. Moreover, in some cases, in practice, it may be useful to create a table that contains results for one only cohort (one diagonal of the whole dynamic life table). In such a case, the dynamic life table describing mortality of one cohort to which an individual in age *x* belongs, assuming mortality improvement each *t* year, will be the following:

Tabla 2. Esquema de tabla de mortalidad dinámica abreviada
Table 2. Sample of an abbreviated dynamic life table

x	$q_{x,t}$
x_d	$q_{xd,0}$
x_{d+1}	$q_{xd+1,1}$
x_{d+2}	$q_{xd+2,2}$
…	…
110	$q_{110,110-xd}$

2. EL MODELO DE LEE-CARTER

El modelo de predicción de la mortalidad llamado en la actualidad el "modelo de Lee-Carter" fue propuesto por Ronald D. Lee and Lawrence R. Carter en el año 1992 (Lee y Carter, 1992). El primer modelo fue ajustado para la población de Estados Unidos; los datos históricos consistían en ratios de mortalidad poblacional por edad para los años 1933-1987. Los autores no disponían de información detallada sobre la mortalidad a partir de 85 años de edad.

El objetivo del modelo de Lee-Carter es calcular el riesgo de mortalidad a la edad *x*, en el instante *t*, estimando previamente los parámetros de la siguiente función:

$$m_{x,t} = e^{a_x + b_x k_t + \varepsilon_{x,t}} .$$

2. LEE-CARTER MODEL

The model widely used at present for mortality improvement projection is the Lee-Carter model which was proposed in 1992 by Ronald D. Lee and Lawrence R. Carter (Lee and Carter, 1992). The first model was estimated for the population of the United States. The data consisted of historical population mortality rates focusing on age in years 1933-1987. Mortality rates for ages over 85 were not considered due to lack of reliable data.

The purpose of Lee-Carter model is to calculate the mortality risk at age *x*, in the moment *t*, estimating parameters in the following function:

$$m_{x,t} = e^{a_x + b_x k_t + \varepsilon_{x,t}} .$$

En el modelo propuesto, k_t es un índice dependiente del tiempo t que describe el nivel de mortalidad, para determinados coeficientes de a_x y b_x. La observación de los valores de k_t permite describir la tendencia de la mortalidad durante el periodo observado. Los valores del coeficiente b_x permiten determinar la fuerza de disminución de la mortalidad para un cambio de k_t:

$$k_t(\frac{dln(m_{x,t})}{dt} = b_x \frac{dk_t}{dt})$$

El coeficiente a_x representa una media de valores de logaritmos naturales de m_x para la misma edad x y varios momentos t.

Los errores del modelo $(\varepsilon_{x,t})$ siguen una distribución normal con media 0 y varianza σ^2 y reflejan los cambios de mortalidad que no están descritos por el modelo.

Con el fin de estimar los parámetros del modelo (a_x, b_x, k_t) se aplica el método de descomposición de valores propios (*single value decomposition*). Para encontrar la solución única se requieren las siguientes condiciones:

$$\sum_{\substack{i=0 \\ T}}^{\omega} b_x = 1$$
$$\sum_{t=1} k_t = 0$$

Aplicando el método de mínimos cuadrados se busca el mínimo de la función:

$$\sum_{x=0}^{\omega}\sum_{i=1}^{T}(\ln(m_{x,t}) - a_x - b_x k_t - e_t)^2 \rightarrow min.$$

Al estimar los parámetros, el modelo de Lee-Carter permite pronosticar la mortalidad futura. Con este fin se asume que los parámetros a_x y b_x son constantes y solamente k_t se modifica con el tiempo. Lee y Carter, construyendo un modelo

In the proposed model, coefficient k_t depends on time and describes level of mortality risk for defined coefficients a_x and b_x. Observation of values of k_t allows us to describe mortality evolution over the period under study. Values of the coefficient b_x allow to measure mortality improvement for each change of k_t:

$$k_t(\frac{dln(m_{x,t})}{dt} = b_x \frac{dk_t}{dt})$$

Coefficient a_x represents an average of natural logarithms of m_x for the same age x and in the different t moments.

Errors of the model $(\varepsilon_{x,t})$ have a normal distribution with average value 0 and variance σ^2 and represent mortality changes which are not described by the model.

In order to estimate parameters of the model (a_x, b_x, k_t) a single value decomposition method is applied. A unique solution of the problem requires the following conditions:

$$\sum_{\substack{i=0 \\ T}}^{\omega} b_x = 1$$
$$\sum_{t=1} k_t = 0$$

Applying the least squares method we search for the minimum in the following function:

$$\sum_{x=0}^{\omega}\sum_{i=1}^{T}(\ln(m_{x,t}) - a_x - b_x k_t - e_t)^2 \rightarrow min.$$

Having estimated parameters of the Lee-Carter model, it is possible to project future mortality. For the purpose of projection it is assumed that parameters a_x and b_x are constant over time and only k_t changes over the period in question. Lee and Carter, when estimating parameters of the model for the population

para Estados Unidos, comprobaron que, en su caso, los valores de k_t se pueden calcular mediante el modelo de camino aleatorio (*single walk with drift*). Sin embargo, para distintas series de datos históricos, pueden ser más adecuados otros modelos de series temporales (Lee y Carter, 1992).

Si la evolución del índice de mortalidad en cada momento t (k_t) sigue un camino aleatorio (*single walk with drift*), valores consecutivos del parámetro pueden describirse mediante la función:

$$k_t = k_{t-1} + c_t + e_t$$

donde c representa una constante que modifica el valor de k para cada instante t. El valor de c lo podemos estimar como:

$$c = \frac{1}{T-1}(k_T - k_1),$$

siendo T el valor del último periodo de observación.

La fórmula para la estimación de la tasa central de mortalidad futura en el momento $T + \Delta t$ sería entonces:

$$\hat{m}_{x,T+\Delta t} = e^{a_x + b_x k_{T+\Delta t}} = e^{\bar{m}_x + b_x(k_T + \Delta t \cdot c)}$$

El modelo de Lee-Carter se puede entender como un modelo de Componentes Principales (Lee y Carter, 1992; Girosi y King, 2007) que está compuesto por una sola componente ($b_x k_t$) y que es suficiente para describir la mayor parte de la varianza del sistema.

Consideramos una matriz de P x T elementos compuesta por logaritmos naturales de la tasa central de mortalidad $ln(m_{x,t})$ para varias edades $x \in P$ y en varios instantes $t \in T$. Asumimos que el espacio P dimensional puede ser reducido hasta una sola dimensión sin mayor pérdida de información. Un modelo compuesto por P componentes sería:

of the United States, found out that for their studied set of data values of k_t follow the pattern of a single walk with drift. Other populations and sets of data may, however, require the use of other methods of time series (Lee and Carter, 1992).

If k_t follows the single walk with drift pattern, then consecutives values of the parameter can be described with the formula:

$$k_t = k_{t-1} + c_t + e_t$$

where the constant c represents the change in k in each period t. The value of c can be calculated as:

$$c = \frac{1}{T-1}(k_T - k_1),$$

where T is the value of the last observation.

Future mortality at the moment $T + \Delta t$ can be described with the following equation:

$$\hat{m}_{x,T+\Delta t} = e^{a_x + b_x k_{T+\Delta t}} = e^{\bar{m}_x + b_x(k_T + \Delta t \cdot c)}$$

The basic Lee-Carter model can also be defined as a Principal Component Analysis model (Lee and Carter, 1992; Girosi and King, 2007), composed of only one component ($b_x k_t$) which is sufficient to describe the majority of the system variability.

We consider a matrix of P x T elements composed of natural logarithms of central death rate $ln(m_{x,t})$ for various ages $x \in P$ and observation periods $t \in T$. We assume that the P-dimensional space can be reduced to only one dimension without major loss of information. The model composed of P components is:

$$\ln(m_{x,t}) = a_x + b_{x1}k_{t1} + b_{x2}k_{t2} + \cdots + b_{xA}k_{tP}$$

Si el modelo anterior lo reducimos a una sola componente, se obtiene el modelo presentado por Lee y Carter. La única componente es una función lineal del tiempo.

$$\ln(m_{x,t}) = a_x + b_{x1}k_{t1} + b_{x2}k_{t2} + \cdots + b_{xA}k_{tP}$$

If the model above is reduced to only one component, we obtain the model proposed by Lee and Carter. The unique component is a lineal function of time:

$$\ln(m_{x,t}) = \underbrace{a_x + b_{x1}k_{t1} + b_{x2}k_{t2} + \cdots + b_{xP}k_{tP}}_{\text{Modelo de Lee-Carter (sin componente aleatoria)}} = \underbrace{a_x + b_{x1}k_{t1}}_{\text{Lee-Carter Model (without aleatory component)}} + \underbrace{\varepsilon_{x,t}}_{\varepsilon_{x,t}}$$

La aplicación del método de Análisis de Componentes Principales a la población de España (Dylewska y Galindo, 2011) demuestra que para una población entre 40 y 85 años y para los años 1958-2009, un modelo compuesto por una componente principal describe alrededor de un 95% de la varianza del sistema. Por lo tanto, la aplicación del modelo de Lee-Carter en su forma básica para pronosticar la mortalidad de hombres y mujeres de España se considera adecuada.

Application of Principal Component Analysis to the population of Spain (Dylewska and Galindo, 2011) shows that for a population ranging 40-85 and in years 1958-2009, the Lee-Carter model composed of only one component describes the great majority of the system variability (around 95%). Therefore, the use of the basic Lee-Carter model for projecting future mortality evolution can be considered as adequate.

Entre otros estudios donde se comprueba la adecuación del modelo de Lee-Carter a los datos destaca el estudio realizado por Girosi y King (2007) para datos históricos sobre la mortalidad en Estados Unidos y varios países europeos. Los autores analizan una matriz de datos que consiste en la mortalidad por grupos de edad, tiempo y causas de muerte. Los resultados obtenidos indican que en el caso del análisis de la mortalidad en EEUU sin distinción de causa de muerte, el modelo compuesto por una componente explica un 93% de la variabilidad. También realizan el análisis para España, obteniendo un porcentaje de varianza explicada por la primera componente igual al 89%.

Beyond other studies that verify the Lee-Carter model adequacy to describe mortality evolution, there is a study conducted by Girosi and King (2007) for historical data regarding the mortality in the United States and various European countries. The authors analyze a matrix of data regarding mortality in various age groups, years and causes of death. The results imply that in the case of the United States and for total deaths (without considering cause of death), the model composed of only one component describes 93% of the total variability. In the case of Spain, the percentage of variability described by the model with one component is 89%.

3. PARÁMETROS DEL MODELO DE LEE-CARTER

El modelo de Lee-Carter fue ajustado a las tablas de mortalidad anuales de España para los años 1985-2009 (25 años consecutivos) publicadas por HMD (*Human Mortality Database*). La razón del uso de esta base de datos en lugar de las tablas de vida oficiales (INE) es que el HMD proporciona una metodología de construcción de las tablas de vida uniforme para todos los países (Wilmoth et al., 2007).

Una buena aproximación sobre los cambios de mortalidad en la población española lo proporcionan las representaciones gráficas de los coeficientes a_x, b_x y k_t obtenidos:

- Coeficiente a_x:

La observación de los valores del coeficiente a_x permite confirmar cómo crece con la edad el valor medio del riesgo de mortalidad. Destaca un mayor riesgo de mortalidad de hombres frente al riesgo de mortalidad de mujeres y el incremento de mortalidad de jóvenes alrededor de la edad de 20 años. Este incremento es más fuerte en el caso de hombres que en el de mujeres y es la consecuencia del aumento del número de accidentes entre jóvenes.

3. PARAMETERS OF LEE-CARTER MODEL

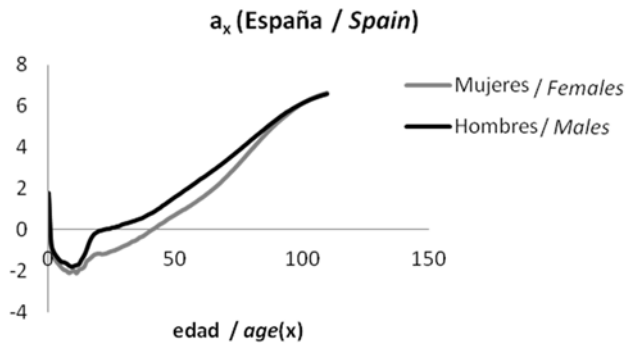
The Lee-Carter model presented in this study has been adjusted to annual mortality tables of Spain for the years 1985-2009 (25 consecutive years) published by HMD (Human Mortality Database). The reason to use this database rather than the Spanish Statistics Institute (INE) tables is that the study was also conducted for other countries and HMD applies a uniform methodology of study for all populations (Wilmoth et al., 2007).

Graphical representations of estimated coefficients a_x, b_x and k_t are also very helpful in observing changes of mortality over the period under study:

- Coefficient a_x:

The observation of values of the coefficient a_x helps to understand how much increase in age affects the average value of mortality risk. On the graph below it can be observed how much risk of male death is higher than in the case of females, as well as the increase in mortality risk in a younger group – around an age of 20 year-olds. This increase of risk is higher for males than for females and it is a consequence of accidental deaths.

Gráfico 1. ***Coeficiente a_x del modelo de LC para la población de España***
Graph 1. Coefficient a_x of the model of Lee-Carter for population of Spain



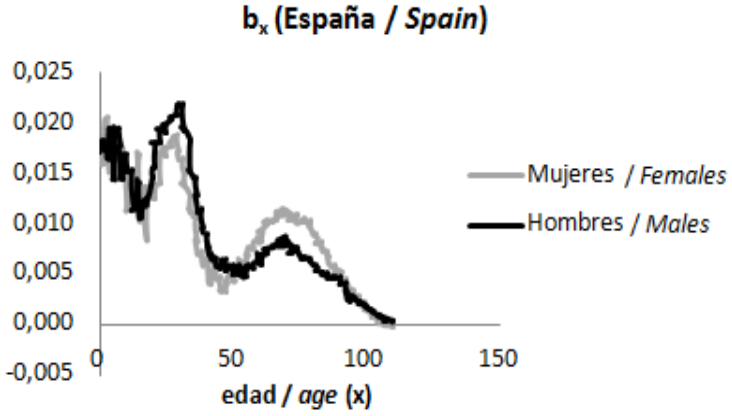
Coeficiente b_x:

El coeficiente b_x se puede entender como la parte de cambio en el valor medio (a_x) que podemos atribuir al cambio de edad. Este coeficiente modifica la tendencia de a_x dependiendo de si el cambio de mortalidad es más rápido o más lento que en el caso de la tendencia (a_x) (Bowers et al., 1997). La evolución de b_x también da una idea de lo rápidamente que decrecen los ratios en respuesta a cambios de k_t (Dylewska y Galindo, 2011). Destaca un incremento de la mortalidad entre 20 y 30 años de edad.

Coefficient b_x:

Coefficient b_x can be defined as the part of change in the average value (a_x) that can be assigned to a change in age. Coefficient b_x adjusts values of a_x depending if the change in mortality is faster or slower than the change in the average mortality (a_x) (Bowers et al., 1997). The observation of the evolution of b_x allows us to define how fast the coefficient decreases in relation to the changes in k_t (Dylewska and Galindo, 2011). The graph below confirms the previous observation regarding the increase in mortality between 20 and 30 year-olds.

Gráfico 2. Coeficiente b_x del modelo de LC para la población de España
Graph 2. Coefficient b_x of the model of Lee-Carter for population of Spain



Coeficiente k_t:

Los valores de k_t representan la tendencia de la mortalidad a lo largo del periodo de observación. Observando los valores de los coeficientes para hombres y mujeres, podemos decir que la mortalidad tiene tendencia decreciente. Las líneas para los dos sexos son casi iguales.

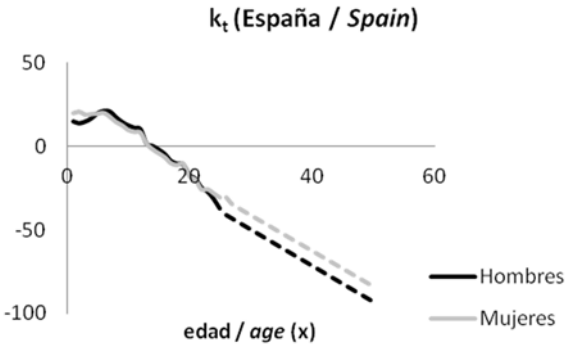
Coefficient k_t:

Values of the coefficient k_t represent the overall evolution of mortality over the period under study. The observation of values of this coefficient for males and females makes us come to the conclusion that the tendency is decreasing and that the lines for both sexes are very similar.

La observación de los valores de k_t tiene también otro propósito que es confirmar si los futuros valores de k_t se pueden estimar como un camino aleatorio (*random walk with drift*). Si no es así, habría que buscar otro modelo para la estimación de futuros valores de k_t. En este caso las representaciones de la función de k_t para el momento t tienen una forma decreciente que podría ajustarse a una función lineal. Por lo tanto, se concluye que los futuros valores de k_t pueden ser estimados con la función $k_t = k_{t-1} + c_t + e_t$.

Moreover, the observation of values of k_t allows us to confirm that future values of k_t can be estimated as a random walk with drift. Otherwise, other model for estimation of future values of k_t should be adjusted. In this case, the function k_t in time t is decreasing and can be described by a lineal function. Therefore, we can conclude that future values of k_t can be defined by the function $k_t = k_{t-1} + c_t + e_t$.

Gráfico 3. ***Coeficiente k_t del modelo de LC para la población de España***
Graph 3. ***Coefficient k_t of Lee-Carter model for population of Spain***



4. ESPERANZA DE VIDA A LA EDAD *X* SEGÚN LAS TABLAS ESTÁTICAS Y DINÁMICAS

Los valores de la esperanza de vida a la edad 0 pronosticados para la población de España para el año 2029 según las tablas de vida estáticas son 81,5 años de vida en el caso de los hombres y 87,2 años de vida en el caso de las mujeres. Esto supone un aumento de 3 años respecto a la esperanza de vida de los hombres en España en el año 2009 y 2,6 años respecto a la esperanza de vida de mujeres en el mismo año. La esperanza de vida a la edad 0 pronosticada según las tablas de vida dinámicas para un individuo nacido en el año 2009 es de 87,2 años en el caso de un hombre y de 92,3 años para una mujer.

4. REMAINING LIFE EXPECTANCY AT AGE *X* BASED ON STATIC AND DYNAMIC LIFE TABLES

Life expectancy at age 0 projected for the population of Spain in the year 2029 and calculated on a basis of static life tables is 81.5 years in the case of males and 87.2 years in the case of females. This implies an increase of 3 years in the expected life-time of males in Spain in year 2009 and a 2.6-year increase in the expected lifetime of females in the same year. Remaining life expectancy at age 0 projected with the use of dynamic life tables for an individual born in 2009 is 87.2 years in the case of males and 92.3 years in the case of females.

La razón de la diferencia entre la esperanza de vida calculada según las tablas estáticas y dinámicas resulta del distinto método de predicción. Mientras que las tablas estáticas pronosticadas proporcionan solamente un movimiento del riesgo de mortalidad a la edad x en el año s hasta el riesgo de mortalidad a la misma edad en el año $s+t$, las tablas de vida dinámicas permiten el cálculo del riesgo de mortalidad ajustado año por año según transcurre el tiempo y aumenta la edad del individuo. Por lo tanto, las tablas de vida dinámicas permiten el cálculo del valor de la esperanza de vida más adecuado que las tablas de vida estáticas (bien pronosticadas o actuales).

The reason of the difference in life expectancy calculated on the basis of static and dynamic life tables is the different projection method. Static projected life tables assume that the mortality risk in year s at each age x moves into different mortality risk for the same age in year $s+t$. In contrast, dynamic life tables allow us to calculate the mortality risk adjusted year by year, as time passes and population get older. Therefore, dynamic life tables allow for more precise calculation of life expectancy in comparison with static life tables (both projected and actual).

Tabla 3. Esperanza de vida futura en España de un individuo de edad 0 en los años 1989, 2009, 2029 / *Table 3. Remaining life expectancy in Spain of an individual of age 0 in years 1989, 2009, 2029*

	e_0 España / *Spain*					
	Mujeres / *Females*			Hombres / *Males*		
	1989	2009	2029	1989	2009	2029
Tablas de vida estáticas *Static life tables*	78,5	84,6	87,2	73,4	78,5	81,5
Tablas de vida dinámicas *Dynamic life tables*	-	92,3	94,0	-	87,2	89,4

Tabla 4. Esperanza de vida futura en España de un individuo de edad 67 en los años 1989, 2009, 2029 / *Table 4. Remaining life expectancy in Spain of an individual of age 67 in years 1989, 2009, 2029*

	e_{67} España / *Spain*					
	Mujeres / *Females*			Hombres / *Males*		
	1989	2009	2029	1989	2009	2029
Tablas de vida estáticas *Static life tables*	17,5	20,2	22,1	14,1	16,6	18,4
Tablas de vida dinámicas *Dynamic life tables*	-	21,5	23,4	-	17,6	19,4

Un aspecto adicional es la necesidad de creación de las tablas de vida unisex como consecuencia de la implementación de la Directiva de la Unión Europea 2004/113/WE. No cabe duda que tanto el nivel de mortalidad como la tendencia de disminución de la mortalidad son muy distintos para hombres y mujeres. La aplicación de las tablas de vida dinámicas en la construcción de tablas unisex permite proteger un producto de seguros ante el riesgo de longevidad y de cambio de estructura de sexo resultante del distinto riesgo de mortalidad para hombres y mujeres.

Las tablas de vida dinámicas unisex las podemos construir utilizando dos modelos de Lee-Carter, estimados por separado para la población de hombres y mujeres, obteniendo los valores de m_x y q_x pronosticados. Sumando los parámetros d_x y l_x para los dos sexos podemos asumir previamente una proporción de cada sexo, que refleja la estructura en el momento de entrada al grupo de asegurados (el momento de adquisición). Una gran ventaja de este método de creación de las tablas de vida unisex consiste en el hecho de que las tablas reflejan no solamente el cambio de mortalidad resultante del incremento de edad y tiempo, sino también cambios de estructura de sexo en un grupo de asegurados. Esto se refiere sobre todo al incremento de la proporción de mujeres en el grupo debido al menor riesgo de mortalidad que presentan.

5. CONCLUSIONES

La construcción del modelo de Lee-Carter ha permitido realizar un pronóstico de la mortalidad futura en España para los próximos 20 años. Dado el alto porcentaje de variabilidad explicada por el modelo en su forma básica (con una sola componente), el modelo tiene una alta probabilidad de pronosticar bien el nivel de morta-

Another aspect to bear in mind is the need of constructing unisex mortality tables as a consequence of the implementation of the European Union Directive 2004/113/WE. Without doubt both risk of mortality and mortality improvement of males and females are very different, which makes the issue of unisex product pricing even more complex. The application of dynamic life tables in constructing unisex life tables allows for better fit of the model and mitigates the longevity risk and risk of change of sex structure resulting from different mortality risk of males and females.

Unisex mortality tables can be constructed by using two Lee-Carter models estimated separately for populations of males and females, obtaining projected values of m_x and q_x. By adding up parameters d_x and l_x for both sexes, we can make an assumption about the proportion of each sex which reflects the gender structure at the contract issue date. The main advantage of this approach is that unisex life tables constructed with this method reflect changes of mortality in portfolio which result not only from an increase in age but also the mortality improvement and the change in insured's gender structure in portfolio. This especially relates to the increase in the proportion of females over years due to their lower mortality risk.

5. CONCLUSIONS

As a result of the estimation of the Lee-Carter model for the population of Spain, a prognosis of the future mortality for next 20 years has been made. Taking into account that the proportion of mortality explained by the first component of this model is very high, the chances for good projection using the basic formula of the model are very high. Values of the parameters of the Lee-Carter model, despite

lidad futura. Sin embargo, los valores de los parámetros del modelo de Lee-Carter calculados en este estudio, aunque permiten una estimación a largo plazo, deberían ser verificados periódicamente ya que los valores del parámetro b_x no suelen ser constantes a lo largo de periodo de observación y vienen afectados por el horizonte temporal elegido. Además, el parámetro a_x, siendo una media aritmética, depende fuertemente del periodo elegido para la estimación del modelo.

El cálculo de la esperanza de vida a la edad 0 (e_0) para tablas estáticas y dinámicas, difiere considerablemente cuando se trabaja con cada tipo de tabla, hecho esperable ya que las tablas dinámicas consideran el continuo decrecimiento del riesgo de mortalidad. La esperanza de vida de un individuo-hombre que nacerá en España en el año 2029 según las tablas estáticas (tradicionales) es de 81,5 años mientras que la esperanza de vida calculada con las tablas de vida dinámicas alcanza el valor de 89,4 años. La previsión de vida de las españolas es de 87,2 años para tablas estáticas y 94,0 años para las tablas dinámicas.

allowing for a long term prognosis, should be periodically verified as values of the coefficient b_x are likely to change over the period under study. Moreover, the parameter a_x is the arithmetic average of the mortality risk and it is strongly affected by the selected period in question.

Remaining life expectancy at age 0 (e_0) calculated with the use of static and dynamic life tables differs a lot, which results from continuous mortality improvement assumed for dynamic life tables. The life expectancy for a male who will be born in Spain in year 2029 according to static (traditional) projected life tables is 81.5 years while life expectancy calculated using dynamic mortality tables is 89.4 years. The relevant figures for females are 87.2 years of life according to static projected life tables and 94 years for dynamic life tables.

BIBLIOGRAFÍA / REFERENCES

Booth, H. (2006). Demographic forecasting: 1980 to 2005 in review. *International Journal of Forecasting*, *22*(3), 547– 581.

Bowers, N.L., Gerber, H.U., Hickman, J.C., Jones, D.A. y Nesbitt, C.J. (1997). *Actuarial mathematics* (2nd ed.). Schaumburg (Illinois): Society of Actuaries.

Continuous Mortality Investigation (CMI) Committee (2007). Stochastic projection methodologies: Lee-Carter model features, example results and implications. CMI Working Paper No. 25.

Continuous Mortality Investigation (CMI) Committee (1999). Standard tables of mortality based on the 1991-1994 experiences. The distribution of policies per life assured. CMI Report No. 17.

Debón Aucejo, A., Martínez Ruiz, F. y Montes Suay, F. (2006). Dynamic life tables. Age-period-cohort models. Paper presented on the *10th International Congress on Insurance: Mathematics and Economics (IME)*. Leuven (Bélgica), 18th-20th July.

Debón Aucejo, A., Martínez Ruiz, F. y Montes Suay, F. (2007). Modelo Lee-Carter extendido. *XV Jornadas de ASEPUMA y III Encuentro Internacional*. Palma de Mallorca, 20-21 de septiembre.

Dylewska, E. y Galindo Villardón, M.P. (2011). *Aplicación del modelo de Lee-Carter para la construcción de tablas de mortalidad dinámicas para Polonia y España*. Universidad de Salamanca: Trabajo Fin de Máster no publicado.

Girosi, F. y King, G. (2007). *Understanding the Lee-Carter Mortality forecasting method*. Copy at http://gking.harvard.edu/files/lc.pdf.

Lee, R.D. y Carter, L.R. (1992). Modelling and forecasting U.S. mortality. *Journal of the American Statistical Association*, *87*(419), 659-671.

López Cachero, M. y López de la Manzanara Barbero, J. (1996). *Estadística para actuarios*. Madrid: MAPFRE.

Rossa A. (2009). Dynamiczne tablice trwania życia oparte na metodologii Lee-Cartera i ich zastosowanie do obliczania wysokości świadczeń emerytalnych. *Acta Universitatis Lodziensis. Folia Oeconomica, 231*, 367-384.

Wilmoth, J.R., Andreev, K., Jdanov, D. y Glei, D.A. (2007). *Methods protocol for the human mortality database*. Version 5. [http://www.mortality.org/Public/Docs/MethodsProtocol.pdf].

Pecvnia Monográfico 2012, pp. 179-210

ASPECTOS TÉCNICOS DE LAS ESTADÍSTICAS OFICIALES

José Ignacio Alonso Cimadevilla[1]

Instituto Nacional de Estadística de León

Resumen

Los múltiples cambios sociales que han tenido lugar en las últimas décadas, han repercutido de forma especial en la elaboración de las estadísticas en nuestro país. En este trabajo se describen las estadísticas que mayor repercusión tienen a nivel económico y que son elaboradas por el organismo oficial de las estadísticas oficiales, el Instituto Nacional de Estadística.
Concretamente, se desarrolla la metodología utilizada en la elaboración de la Encuesta de Población Activa (EPA), cuyos resultados permiten conocer la actividad económica en relación al componente humano. Por otro lado, y respecto al Índice de Precios al Consumo (IPC), se describen los aspectos técnicos del índice base 2006 y base 2011, entendiendo que éste último constituye una actualización del anterior. El trabajo se completa con los Censos Demográficos 2011, cuya metodología ha experimentado una gran variación. Aunque se describe el proceso de elaboración del Censo de Edificios y del Censo de Viviendas, se concreta de forma más detallada el Censo de Población, considerado como el mayor proyecto estadístico de un país.

Palabras clave: Encuesta de población activa; Índice de precios al consumo; Censo de población.

INTRODUCCIÓN

El Instituto Nacional de Estadística (INE) es un organismo autónomo de carácter administrativo, con personalidad jurídica y patrimonio propio, adscrito al Ministerio de Economía y Competitividad a través de la Secretaría de Estado de Economía y Apoyo a la Empresa. Se rige, básicamente, por la Ley 12/1989, de 9 de mayo, de la Función Estadística Pública, que regula la actividad estadística para fines estatales, la cual es competencia exclusiva del Estado, y por el Estatuto aprobado por Real Decreto

[1] Delegado Provincial del Instituto Nacional de Estadística. León.

508/2001 de 11 de mayo, y modificado por el Real Decreto 947/2003, de 18 de julio, por el Real Decreto 759/2005, de 24 de junio y por el Real Decreto 950/2009, de 5 de junio.

La Ley asigna al INE un papel destacado en la actividad estadística pública encomendándole expresamente la realización de las operaciones estadísticas de gran envergadura (censos demográficos y económicos, cuentas nacionales, estadísticas demográficas y sociales, indicadores económicos y sociales, coordinación y mantenimiento de los directorios de empresas, formación del Censo Electoral...). También se ocupa de las siguientes funciones: la formulación del Proyecto del Plan Estadístico Nacional con la colaboración de los Departamentos Ministeriales y del Banco de España; la propuesta de normas comunes sobre conceptos, unidades estadísticas, clasificaciones y códigos; y las relaciones en materia estadística con los Organismos Internacionales especializados y, en particular, con la Oficina de Estadística de la Unión Europea (EUROSTAT).

Su creación se debe a la Ley de 31 de diciembre de 1945 (BOE del 3 de enero de 1946), asignándole como misiones principales la elaboración y perfeccionamiento de las estadísticas demográficas, económicas y sociales ya existentes, la creación de otras nuevas y la coordinación con los servicios estadísticos de las áreas provinciales y municipales. Además de regular la coordinación entre otros servicios estadísticos como el Servicio Sindical de Estadística, la Ley crea el Consejo Superior de Estadística y se organiza en Servicios Centrales, Delegaciones provinciales y Delegaciones en los Ministerios.

El comienzo de la estadística oficial en España viene marcado por el Decreto firmado el 3 de noviembre de 1856 por el general Narváez, presidente del Consejo de Ministros de Isabel II, por el que se creaba una Comisión, compuesta por personas de reconocida capacidad, para la formación de la Estadística General del Reino. Unos meses más tarde, el 21 de abril de 1857, la Comisión pasa a denominarse Junta de Estadística. Su primer trabajo es el Censo de Población, con fecha de referencia del 21 de mayo del mismo año. Es la Ley de Instrucción Pública, de 9 de septiembre de 1857, la que establece que la Estadística será una disciplina académica.

Ya en el siglo XX, el Decreto de 1 de octubre de 1901 establece la formación de las estadísticas oficiales y la publicación de las mismas. Se crea la Dirección General y se crean departamentos en los Ministerios para completar su labor. En 1924, el Consejo del Servicio Estadístico, creado en 1921, es reformado, cuatro años antes de que pase a depender del Ministerio de Trabajo y Previsión. En 1931, la adscripción se hace al Ministerio de la Presidencia. Durante la Guerra Civil (1936-1939) comienza a funcionar el Servicio Sindical de Estadística en coordinación con los Servicios de Estadística del Estado, dentro de la zona controlada por los militares sublevados.

Dentro del marco de la Comunidad Económica Europea, el Sistema Estadístico Europeo (SEE) está formado por Eurostat (la oficina de estadística de la UE), las oficinas de estadística de todos los estados miembros y otros organismos que elaboran estadísticas europeas. El SEE garantiza que las estadísticas europeas elaboradas en todos los Estados miembros de la Unión Europea sean fiables, siguiendo unos criterios y definiciones comunes y

tratando los datos de la manera adecuada para que sean siempre comparables entre los distintos países de la UE.

En la actualidad el SEE se regula fundamentalmente por la Ley Estadística Europea, aprobada en 2009 mediante el Reglamento (CE) 223/2009 del Parlamento Europeo y del Consejo. Las estadísticas de la Unión Europea se preparan, elaboran y difunden tanto por el Sistema Estadístico Europeo (SEE) como por el Sistema Europeo de Bancos Centrales (SEBC). El SEE cuenta con un Programa Estadístico Europeo que recoge la planificación estadística para un periodo de cinco años. Este Programa es aprobado por el Parlamento Europeo y por el Consejo y para asegurar que se tengan en cuenta las necesidades de los usuarios en la elaboración del Programa Estadístico Europeo, se creó el Comité Consultivo Europeo de Estadística. En él están representados los usuarios, informantes, instituciones académicas y sociales y la administración comunitaria.

No obstante, la planificación de la actividad se hace conjuntamente entre los INE y Eurostat, la producción de estadísticas nacionales armonizadas corresponde a las autoridades de los Estados miembros, mientras que Eurostat recopila los datos que aportan los Estados, los analiza y en base a ellos ofrece cifras comparables y armonizadas, de forma que se puedan definir, acometer y analizar las políticas comunitarias. Además, Eurostat se encarga de asegurar la coordinación necesaria para garantizar el funcionamiento de este complejo sistema (lenguas diferentes, formas de organización administrativa muy diversa, nomenclaturas específicas...) y para asegurar la coherencia y calidad de los datos.

En este trabajo se van a desarrollar los principales aspectos técnicos de tres de las investigaciones con mayor repercusión de las muchas que se realizan en el INE. En primer lugar, la Encuesta de Población Activa (EPA) como ejemplo de muestreo probabilístico; a continuación, el Índice de Precios de Consumo (IPC) como ejemplo de muestreo no probabilístico y, finalmente, el Censo de Población, operación que, aunque tradicionalmente ha sido dirigida a toda la población, en esta última ocasión, se ha efectuado por muestreo.

1. ENCUESTA DE POBLACIÓN ACTIVA (EPA)

La EPA es una investigación por muestreo de periodicidad trimestral, dirigida a la población que reside en viviendas familiares del territorio nacional y cuya finalidad es averiguar las características de dicha población en relación con el mercado de trabajo.

Se publicó por primera vez en 1964 y hasta finales de 1968 se obtuvieron resultados con referencia trimestral; de 1969 a 1974 la referencia temporal pasó a ser semestral y a partir de 1975 volvió a ser trimestral. En 1987 se modificó el cuestionario de la encuesta para adaptarse a las últimas recomendaciones internacionales de entonces (Conferencia Internacional de Estadísticos del Trabajo) y la exigencia de adaptar la EPA a la Encuesta de Fuerza de Trabajo de la Comunidad Económica Europea con motivo de la incorporación de España en las Comunidades Europeas en 1986. En 1999 se convierte en una 'encuesta continua' dado que las entrevistas se realizan a lo largo de las 13 semanas de cada trimestre y no de 12 de las 13 como se venía haciendo hasta ese momento.

En 2002 se introduce una nueva definición operativa de paro produciéndose una ruptura en las series de parados y activos, cuyo impacto se valoró elaborando doble estimación de ambas definiciones a lo largo del año 2001; en 2005 se produjo el último cambio metodológico sustancial hasta el momento, introduciendo un nuevo cuestionario y el control centralizado del sistema de recogida mediante encuesta telefónica asistida por ordenador. También en este año se calcularon series retrospectivas para el periodo 1996-2004 con la nueva base de población instaurada ese año, con el fin de mantener la homogeneidad de las estimaciones. El periodo anterior quedó sin variaciones. Las cifras actuales de la encuesta se encuadran en la metodología instaurada en 2005.

Su **finalidad** principal es conocer la actividad económica en lo relativo a su componente humano. Está orientada a dar datos de las principales categorías poblacionales en relación con el mercado de trabajo (ocupados, parados, activos, inactivos) y a obtener clasificaciones de estas categorías según diversas características. También posibilita confeccionar series temporales homogéneas de resultados. Por último, al ser las definiciones y criterios utilizados coherentes con los establecidos por los organismos internacionales que se ocupan de temas laborales, permite la comparación con datos de otros países.

Las **ventajas** que presenta son:

- se puede realizar de forma continua con la periodicidad que se desee.
- permite profundizar en los aspectos que interesen en relación con la fuerza laboral, al ser una investigación enfocada directamente a estos temas.
- entrevistadores especializados se encargan de la cumplimentación de los cuestionarios.
- los resultados se obtienen con rapidez, al ser una encuesta por muestreo.
- las definiciones y tratamiento de la información son uniformes a lo largo de las sucesivas realizaciones de la encuesta, lo que origina series homogéneas de resultados.
- se pueden obtener resultados para el total nacional y para subconjuntos territoriales (fundamentalmente para las comunidades autónomas y las provincias).

El principal inconveniente se deriva de su propia condición de encuesta por muestreo y es el no poder dar información de algunas características con la mayor desagregación posible. Así, por ejemplo, el número de activos de cada una de las sesenta divisiones de la Clasificación Nacional de Actividades Económicas en cada provincia resulta poco fiable, ya que a una mayor desagregación en la información corresponde un mayor error de muestreo.

Las **unidades** son de dos tipos: por una parte, de muestreo, donde las primarias son las secciones censales (áreas geográficas perfectamente delimitadas) y las secundarias son las viviendas; y, por otra, como unidades de análisis se toman las viviendas y las personas. Cubre todo el territorio nacional desde la inclusión de Ceuta y Melilla en el segundo trimestre de 1988 y va dirigida a la población que reside en viviendas familiares principales, es decir, las utilizadas todo el año o la

mayor parte de él como vivienda habitual o permanente.

Se distinguen dos **periodos de referencia**: el de los resultados, que es el trimestre natural y, el periodo de referencia de la información, que es, en general, la semana natural inmediatamente anterior a la de la entrevista según el calendario, aunque hay preguntas con periodos de referencia especiales.

1.1. Principales definiciones en la EPA

Todas las definiciones están basadas en las recomendaciones aprobadas por la Organización Internacional del Trabajo (OIT) en la Decimotercera y Decimosexta Conferencia Internacional de Estadísticos del Trabajo (Ginebra, 1982 y 1998, respectivamente). Por otra parte, todas las características definidas están referidas al concepto nacional y no al interior de acuerdo con las definiciones del Sistema Europeo de Cuentas Nacionales y Regionales (SEC-95). Esto se debe a que no es posible recoger información de la población que trabaja en España y reside en el extranjero, ya que la Encuesta va dirigida a la población que habita en las viviendas familiares del territorio nacional.

- **Población económicamente activa**: Es el conjunto de personas de unas edades determinadas que, en un periodo de referencia dado, suministran mano de obra para la producción de bienes y servicios económicos o que están disponibles y hacen gestiones para incorporarse a dicha producción. Por tanto, la población económicamente activa comprende todas las personas de 16 o más años que durante la semana de referencia (la anterior a aquélla en que corresponde realizar la entrevista según el calendario) satisfacen las condiciones necesarias para su inclusión entre las personas ocupadas o paradas, según se define más adelante.
- **Población ocupada**: Es la formada por todas aquellas personas de 16 o más años que durante la semana de referencia han tenido un trabajo por cuenta ajena o han ejercido una actividad por cuenta propia.
- **Población parada o desempleada**: Se consideran paradas a todas las personas de 16 o más años que reúnen simultáneamente las condiciones de estar sin trabajo, buscarlo y estar disponible para trabajar.
- **Población económicamente inactiva**: Abarca a todas las personas de 16 o más años, no clasificadas como ocupadas ni paradas ni población contada aparte durante la semana de referencia.
- **Población contada aparte**: Los varones que cumplían el servicio militar obligatorio (o servicio social sustitutorio) se consideraban población contada aparte, esto es, no se les incluían entre los activos ni entre los inactivos, independientemente de que en la semana de referencia hubieran trabajado o no. El servicio militar obligatorio desapareció en diciembre de 2001.

Tabla 1. Principales definiciones en la EPA

<table>
<tr><td colspan="5">Menores de 16 años</td></tr>
<tr><td rowspan="5">Personas de 16 y más años</td><td rowspan="4">Activos</td><td rowspan="3">Ocupados</td><td>Asalariados</td><td>del sector público
del sector privado</td></tr>
<tr><td>Trabajadores por cuenta propia</td><td>Empleadores
Empresarios sin asalariados y trabajadores independientes
Miembros de cooperativas
Ayudas familiares</td></tr>
<tr><td colspan="2">Otros</td></tr>
<tr><td>Parados</td><td colspan="2">que buscan primer empleo
que han trabajado antes</td></tr>
<tr><td>Inactivos</td><td colspan="3">Estudiantes
Jubilados o pensionistas
Labores del hogar
Incapacitados para trabajar
Otra situación (rentistas, ...)
No sabe</td></tr>
<tr><td>Población Contada Aparte (PCA)*</td><td colspan="3">que trabaja
que no trabaja</td></tr>
<tr><td colspan="5">* El año 2002 desapareció el servicio militar obligatorio, por lo que a partir de esa fecha no existe esta categoría poblacional.</td></tr>
</table>

Las principales tasas que se consideran en la EPA son:

- **Tasa global de actividad**, cociente entre el número total de activos y la población total. Se calcula para ambos sexos y para cada uno de ellos por separado.
- La **tasa específica de actividad** para un intervalo de edad determinado, es el cociente entre el número de activos de esas edades y la población correspondiente al intervalo.
- La **tasa global de empleo** es el cociente entre el número total de ocupados y la población total. Se calcula para ambos sexos y para cada uno de ellos por separado.
- La **tasa específica de empleo** para un intervalo de edad determinado es el cociente entre el número de ocupados de esas edades y la población correspondiente al intervalo.
- La **tasa de paro** es el cociente entre el número de parados y el de activos. Se calcula para ambos sexos y para cada uno de ellos por separado.
- La **tasa específica de paro** para un intervalo de edad determinado, que es el cociente entre los parados de edades comprendidas entre los extremos del intervalo y los activos de dicho intervalo.
- **Tasa de asalarización**, cociente entre el número de asalariados y el número total de ocupados.
- **Tasa de temporalidad**, cociente entre el número de asalariados con contrato temporal y el número total de asalariados.
- **Tasa de trabajo a tiempo parcial**, cociente entre el número de ocupados a

tiempo parcial y el número total de ocupados.

1.2. El diseño de la muestra

Tomando como referencia el marco de la encuesta, se utiliza un muestreo bietápico con estratificación de las unidades de primera etapa. Las unidades de primera etapa están constituidas por las secciones censales, cuya muestra permanece fija indefinidamente salvo pocas excepciones. Siempre que una sección sale de la muestra es sustituida por otra seleccionada aleatoriamente.

Las unidades de segunda etapa están constituidas por las viviendas familiares principales (ocupadas permanentemente) y los alojamientos fijos (chabolas, cuevas, etc.). No se consideran encuestables las viviendas secundarias (ocupadas sólo una parte del año), ni las disponibles para alquiler o venta, ya que no forman parte del ámbito poblacional definido anteriormente. Dentro de las unidades de segunda etapa no se realiza submuestreo alguno, recogiéndose información de todas las personas que tengan su residencia habitual en las mismas.

Las unidades de primera etapa se estratifican atendiendo a un doble criterio:

- Criterio geográfico (de estratificación). Las secciones se agrupan en estratos dentro de cada provincia, de acuerdo con la importancia demográfica del municipio al que pertenecen.
- Criterio socioeconómico (de subestratificación). Las secciones censales se agrupan en subestratos dentro de cada uno de los estratos, según las características socioeconómicas de las mismas.

Para llegar a la formación de los estratos se consideran los siguientes tipos de municipios:

- Municipios autorrepresentados: Son aquellos que dada su categoría dentro de la provincia deben tener siempre secciones en la muestra. Se consideran así la capital de la provincia y aquéllos que tienen un número de habitantes tal, que en la afijación proporcional dentro de la provincia le corresponden al menos 12 secciones en la muestra, así como los que teniendo una situación demográfica destacada dentro de la provincia no hay otros similares con que agruparlos, aunque proporcionalmente le correspondan menos de 12 secciones en la muestra.
- Municipios correpresentados: Son aquellos que dentro de la misma provincia forman parte de un grupo de municipios demográficamente similares y que son representados en común.

De acuerdo con esta clasificación se consideran nueve estratos, de tal manera que en los tres primeros se encuentran las tres categorías de municipios autorrepresentados y en el resto los correpresentados, abarcando desde el estrato 4 los municipios entre cincuenta y cien mil habitantes hasta el estrato 9 los menores de dos mil habitantes. Cada diez años, con la información procedente de los Censos de Población, se actualiza la definición de los estratos en cada provincia.

Tamaño de la muestra: Para determinar el tamaño de la muestra, se aplica un procedimiento de mínima varianza para coste fijo. Se parte de un presupuesto (Q) y a partir de él se procede a determinar el número de secciones (n) y el número de viviendas (m) que minimizan la varianza de las estimaciones. Para ello se utiliza una

función de coste de tipo lineal y la expresión del coeficiente de variación para una proporción en el muestreo de conglomerados con submuestreo.

- La función de coste considerada es: $Q = n\,Q_S + n\,m\,Q_V$, con $Q_S = Q_F + d\,Q_D$, donde:

Q = Presupuesto total; Q_S = Coste por unidad primaria (sección); Q_V = Coste por unidad última (vivienda)
n = Número de secciones; m = Número de viviendas por sección
Q_F = Coste fijo por sección; Q_D = Coste diario del trabajo de campo
d = Número de días necesarios para el trabajo de campo

- El coeficiente de variación para una proporción P viene dado por la siguiente expresión:

$$CV^2(\hat{P}) = \frac{V(\hat{P})}{\hat{P}^2} = \frac{1-\hat{P}}{\hat{P}} \frac{1+\delta(m-1)}{nm} = \frac{1-\hat{P}}{\hat{P}} F(\delta, m, n)$$

siendo $F(\delta, m, n) = \dfrac{1+\delta(m-1)}{nm}$

siendo δ el coeficiente de correlación intraclásica, que para el caso de la población activa se ha calculado y vale 0,05. El mínimo de la expresión $CV^2(\hat{P})$ respecto de las variables m y n se obtiene calculando el mínimo de la expresión $\mathrm{F}(\delta, m, n)$ que es independiente de $\hat{P}$. Para distintos valores de m compatibles con el trabajo de campo, m = 4, 6, 8, 10, 11, 14, 17, 18, 19..... 91, 100 y los correspondientes valores de n, dados por $n = \dfrac{Q}{Q_s + mQ_v}$ se obtienen distintos valores para $\mathrm{F}(\delta, m, n)$.

El valor mínimo de $\mathrm{F}(\delta, m, n)$ respecto de m y n corresponde a m=20 y n=3000 y en base a este resultado, la muestra se fija en un total de 3.000 secciones, investigándose una media de 20 viviendas por sección. Posteriormente, la muestra ha tenido diferentes ampliaciones con el objeto de dar cumplimiento a las exigencias de la Unión Europea y mejorar la representación de áreas más desagregadas. A partir del primer trimestre de 2005 se establece un tamaño muestral de 3.588 secciones y 18 viviendas por sección, excepto en las provincias de Madrid, Barcelona, Sevilla, Valencia y Zaragoza, en las cuales el número de entrevistas por sección es de 22. En el tercer trimestre de 2009 se firma un convenio con la Comunidad Autónoma de Galicia por el que se incrementa la muestra en esta comunidad hasta un total de 468 secciones y se asignan estratos separados a los municipios de Santiago de Compostela y Ferrol. El tamaño de muestra final para el total nacional se establece en 3.822 secciones.

Para distribuir las secciones de la muestra entre las provincias, dentro de la provincia entre estratos y dentro de éstos entre subestratos (afijación de la muestra), se tienen en cuenta los siguientes aspectos:

- Los resultados nacionales deben tener la mayor fiabilidad posible. A este respecto hay que recordar que, en general, cuanto más lejos se esté de la afijación óptima por provincias y estratos, mayor es la pérdida de precisión en la estimación nacional para un tamaño fijo de la muestra.
- Disponer en cada provincia de un tamaño mínimo de muestra que permita dar estimaciones de la misma.
- En cada provincia el número de secciones debe ser múltiplo de trece. Con ello se facilita la distribución de la muestra entre las semanas del trimestre.

Para compatibilizar las tres condiciones antes expuestas se ha adoptado una

afijación de compromiso entre la uniforme y la proporcional, a base de agrupar provincias de importancia demográfica similar y asignarles de 3 a 12 entrevistadores, es decir, de 39 a 156 secciones muestrales (con las excepciones de Ceuta y Melilla, que debido a su reducido tamaño poblacional tienen un solo agente y por tanto 13 secciones en la muestra cada una de ellas).

Dentro de cada provincia la afijación entre estratos es proporcional al tamaño de cada uno de ellos, si bien se han potenciado los estratos donde se encuentran los municipios de mayor tamaño, ya que se espera que la mayor parte de las características que se estudian estén correlacionadas con los niveles económico-sociales y culturales de los habitantes y es precisamente en estos estratos donde, en general, la dispersión debe ser mayor y donde el costo por entrevista es menor. Dentro de los estratos, la afijación entre subestratos es estrictamente proporcional al tamaño (medido en número de viviendas familiares).

Selección de la muestra: se realiza de forma que dentro de cada estrato cualquier vivienda familiar tenga la misma probabilidad de ser seleccionada, es decir, se tengan muestras autoponderadas dentro de cada estrato. Este tipo de muestras proporciona pesos de diseño iguales por estrato en los estimadores. Para ello, las unidades de primera etapa (secciones censales) se seleccionan con probabilidad proporcional al número de viviendas familiares principales, según los datos del último Censo o del Padrón Continuo. Dentro de cada sección seleccionada en primera etapa, se selecciona un número fijo de viviendas familiares con igual probabilidad, mediante la aplicación de un muestreo sistemático con arranque aleatorio. Para esta encuesta se ha determinado seleccionar 18 viviendas por sección.

- La probabilidad de selección de la vivienda i, perteneciente a la sección j del estrato h, donde se han afijado K_h secciones, será:

$$P(V_{ijh}) = P(S_{jh}).P\left(\frac{V_{ijh}}{S_{jh}}\right) = K_h.\frac{V_{jh}}{V_h}.\frac{18}{V_{jh}} = K_h.\frac{18}{V_h}$$

$P\ (S_{jh})$ = Probabilidad de selección de la sección j del estrato h

$P\left(\frac{V_{ijh}}{S_{jh}}\right)$ = Probabilidad de selección de la vivienda i condicionada a la selección de la sección j.

V_{jh} = Total de viviendas de la sección j;

V_h = Total de viviendas del estrato h.

Esta probabilidad no depende ni de la vivienda (i) ni de la sección (j), por lo tanto, la muestra es autoponderada.

La muestra se distribuye uniformemente en el tiempo, lo que equivale a que en cada provincia el número de secciones por semana es constante. Además, se ha procurado que la distribución de secciones muestrales por provincia, estrato y semana sea homogéneo, al igual que por provincia, turno de rotación y semana. Cada periodo de la encuesta es de un trimestre siendo cada una de las secciones de la muestra visitada en una de las 13 semanas del mismo. La totalidad de la muestra está dividida en tres submuestras independientes representativas, cada una de ellas, de toda la población.

Como se ha dicho en el párrafo anterior, cada periodo de la encuesta es de un trimestre, repitiéndose ésta sucesivamente. Las secciones censales permanecen fijas en la muestra indefinidamente (salvo las excepciones), sin embargo las viviendas

familiares son renovadas parcialmente cada trimestre de encuesta, a fin de evitar el cansancio de las familias. Esta renovación se efectúa en una sexta parte de las secciones.

A estos efectos, la muestra total se halla dividida en seis submuestras que denominamos *Turnos de rotación*. Cada trimestre se renuevan las viviendas que pertenecen a las secciones de un determinado turno de rotación. Por tanto cada vivienda permanece en la muestra durante seis trimestres consecutivos, al cabo de los cuales sale de la misma para ser reemplazada por otra de la misma sección.

La distribución del número de secciones por estrato y semana es similar en cada turno de rotación. De esta forma se trata de evitar posibles sesgos de medida en las estimaciones, debidos al diferente comportamiento de las familias colaboradoras en función del tiempo que lleven en la encuesta.

Cada turno de rotación puede ser considerado, por tanto, como una submuestra representativa. Este hecho facilita la obtención de estimaciones de variables estructurales mediante la unión de dichas submuestras.

Hasta el año 2001, se utilizaron estimadores de razón tomando como variable auxiliar las Proyecciones Demográficas de población elaboradas por el INE, siendo la expresión del estimador de una determinada característica Y en un trimestre de encuesta la siguiente:

$\hat{Y} = \sum_h \frac{P_h}{p_h} \sum_{i=1}^{n_h} y_{hi}$, extendiéndose el sumatorio h a los estratos de una provincia, una comunidad autónoma o al total nacional, y donde:

P_h: es la proyección de la población, que reside en viviendas familiares, en el estrato h, referida a la mitad del trimestre.

p_h: es el número de personas que habitan en las viviendas de la muestra, en el estrato h, en el momento de la entrevista.

n_h: es el número de viviendas en las secciones de la muestra en el estrato h.

y_{hi}: es el valor de la característica investígada en la vivienda i-ésima, del estrato h.

Las continuas variaciones de población, bien en sus características, bien en su distribución espacial exigen realizar actualizaciones en el marco que necesariamente repercuten en la estructura muestral. En el marco de la EPA se consideran cuatro tipos de actualizaciones: las relativas a las secciones muestrales, consecuencia de las modificaciones producidas por diversas incidencias como particiones, fusiones o variaciones de límites en las secciones seleccionadas; las relativas a las viviendas; las correspondientes a las probabilidades de selección y las relativas a todas las secciones y viviendas de la población.

La forma más directa de actualizar las probabilidades de selección del seccionado es la selección de una nueva muestra a partir del marco disponible más actualizado. Pero un cambio tan radical en una encuesta continua, como es la EPA, genera tres tipos de problemas:

- Pérdida de información imprescindible para la selección y visita de las viviendas que resulten elegidas en la segunda etapa. Esta información, que es necesario rehacer, tiene aspectos tangibles como los directorios de viviendas o la planimetría de la zona, y otros intangibles pero no

menos importantes, como el conocimiento por parte de la población de la sección de la figura del entrevistador, hecho éste que facilita el acceso a las familias y disminuye notablemente la falta de respuesta.
- Pérdida de precisión en las estimaciones de variaciones trimestrales interanuales, al disminuir considerablemente la muestra común entre ambos periodos.
- Posible presencia de discontinuidades en la serie temporal de la encuesta, debidas a la causa citada en el apartado anterior.

A causa de ello se decidió arbitrar un procedimiento que, sin distorsionar las probabilidades de selección que realmente corresponden a cada sección, mantenga la muestra de secciones con las mínimas variaciones, considerando dos tipos de actualizaciones de las probabilidades de selección en función de la información disponible para las mismas.

Errores: los errores que afectan a toda encuesta pueden agruparse en dos grandes grupos:

- **Errores de muestreo**, que se originan por la obtención de resultados sobre las características de una población, a partir de la información recogida en una muestra de la misma.
- **Errores ajenos al muestreo**, que son comunes a toda investigación estadística, tanto si la información es recogida por muestreo como si se realiza un Censo. Estos errores se presentan en cualquier fase del proceso estadístico, tanto antes como durante o tras la recogida de datos: deficiencias del marco e insuficiencias en las definiciones y cuestionarios, por defectos en la labor de los entrevistadores e incorrecta declaración por parte de los informantes o en la depuración, codificación, grabación o tabulación de los resultados.

➢ Los errores de muestreo se calculan trimestralmente de las estimaciones de algunas de las principales características investigadas. Para la obtención de los errores de muestreo se utiliza el método de las semimuestras reiteradas, consistente en obtener sucesivas semimuestras de la muestra inicial. A partir de cada semimuestra se calcula la estimación de la característica de la que queremos obtener el error de muestreo.

Una vez calculadas todas las estimaciones con cada una de las semimuestras, así como la estimación con la muestra completa, el estimador de la varianza viene dado por:

$$\hat{V}(\hat{Y}) = \frac{1}{r}\sum_{i=1}^{r}\left(\hat{y}_i - \hat{Y}\right)^2$$, donde:

r: es el número de semimuestras obtenidas, esto es, el número de reiteraciones.

$\hat{y}_i$: es la estimación obtenida con la i-ésima reiteración. Para cada reiteración se repite el proceso de estimación general, es decir, se aplica la técnica de reponderación (software CALMAR).

$\hat{Y}$: es la estimación basada en la muestra completa.

En el caso de la EPA el número de reiteraciones que se utiliza es de 40. Para formarlas se agrupan todas las secciones de cada estrato por pares, procurando que las dos secciones de cada par pertenezcan al mismo turno de rotación de la EPA. A continuación se asigna aleatoriamente la primera sección de cada par a 20 reiteraciones y la otra sección a las otras 20. De esta forma cada reiteración queda constituida por un número de secciones equivalente al 50 por ciento de la muestra (semimuestra) y cada sección

aparece en la mitad de las reiteraciones. En las tablas se publica el error de muestreo relativo en porcentaje (coeficiente de variación).

- El estudio de los errores ajenos al muestreo presenta numerosas dificultades debido a la gran variedad de causas que los originan, así como a las hipótesis en que se basan los modelos teóricos que, en general, no se cumplen en la realidad, lo que lleva a obtener resultados aproximados.

En la EPA el análisis de los errores ajenos al muestreo se basa en el modelo matemático elaborado por la Oficina de Censos de los Estados Unidos, debido a Hansen, Hurwitz y Bershad (1961), y que, operativamente, consiste en repetir las entrevistas de la encuesta en una submuestra de la muestra de viviendas originalmente seleccionada. Posteriormente se cotejan los datos obtenidos en ambas ocasiones, con objeto de investigar las inconsistencias y cuantificar los errores mediante la aplicación de diversos índices de calidad.

Esta encuesta de evaluación persigue un doble objetivo: por una parte, controlar el trabajo de recogida de la información en todas las comunidades autónomas y, por otra, evaluar la calidad de los resultados.

La comparación de los resultados obtenidos en la encuesta de evaluación (entrevista repetida, ER) con los obtenidos en la entrevista original (EO) permite evaluar dos grandes tipos de errores ajenos al muestreo:

- Errores de cobertura, producidos por la omisión o por la inclusión errónea de unidades (viviendas y personas) en la encuesta original.
- Errores de contenido, que afectan a las características investigadas en las personas encuestables.

El trabajo de campo se lleva a cabo por agentes especializados, los cuales realizan la entrevista repetida a lo sumo tres semanas después de la original, refiriéndose los datos de ambas entrevistas al mismo periodo de tiempo.

El hecho de que más del 70 por ciento de las negativas por primera vez se producen en la primera entrevista a las familias, unido a la existencia de dificultades técnicas para la realización de la encuesta de evaluación (ER) mediante entrevistas telefónicas, han determinado que se investiguen en ER únicamente secciones que en EO se encuentran en primera entrevista. El método de recogida utilizado en estas secciones, tanto en EO como en ER, es de visita domiciliaria. Como consecuencia se dispone de menos muestra en la encuesta de evaluación, respecto a años anteriores, por lo que las cuatro muestras trimestrales se van a agrupar para ofrecer los resultados en cómputo anual, a fin de que éstos sean más representativos.

Para la selección trimestral de la muestra de la encuesta de evaluación se han creado cuatro zonas, agrupando en cada una varias comunidades autónomas, de forma que cada una de éstas esté incluida en una y sólo una de las zonas. Cada semana se investigan las secciones (de primera entrevista) de la muestra en una de las zonas, siendo aleatoria la asignación de las semanas a las zonas, de modo que cada una de éstas se investigue al menos en tres de las semanas del trimestre. De este modo se investigan aproximadamente entre 130 y 150 secciones cada trimestre.

En las secciones seleccionadas se repite la entrevista en la mitad de las viviendas, utilizándose en ER un cuestionario ligeramente reducido respecto al de EO, es decir, con algunas preguntas menos. Con este procedimiento se investiga un número de viviendas de entre 1.300 y 1.500, lo que representa aproximadamente un 2 por ciento de la muestra de la EPA.

Además de la encuesta de evaluación, y con objeto de detectar errores cometidos en el proceso de actualización de las secciones de la muestra, cada trimestre se selecciona una muestra de cincuenta secciones (una de cada provincia, salvo Ceuta y Melilla) para evaluar la calidad de las actualizaciones.

Aparte de la entrevista repetida se realiza un estudio específico de aquellas unidades seleccionadas que son encuestables pero que se negaron a facilitar los datos solicitados. Para estas unidades que se niegan a colaborar en la encuesta se cumplimenta un cuestionario de negativas, en el que se recogen una serie de características básicas, como son el sexo, la edad y la relación con la persona principal de la persona que rehúsa ser entrevistada, así como la edad, el sexo, la nacionalidad, los estudios terminados, la relación con la actividad, la rama de actividad y la ocupación de la persona principal.

2. EL ÍNDICE DE PRECIOS DE CONSUMO (BASE 2006)

La operación del cambio de Sistema del Índice de Precios de Consumo (IPC) consiste, fundamentalmente, en revisar y actualizar cada uno de sus componentes y determinar las mejores opciones para conseguir un indicador representativo y preciso que se adapte a las tendencias de la economía.

Hasta la entrada en vigor de la base 2001, el IPC basaba su cálculo en lo que se denomina sistema de base fija, cuya principal característica es que tanto la composición de la cesta de la compra como sus ponderaciones se mantienen inalterables a lo largo del tiempo que dura la base. Los cambios de base se llevaban a cabo cada ocho o nueve años, debido a que ésa era la periodicidad de la Encuesta Básica de Presupuestos Familiares (EBPF), la fuente utilizada para la elaboración de las ponderaciones y de la cesta de la compra.

A partir de 1997, las dos encuestas de presupuestos familiares que convivían (una continua, con periodicidad trimestral, y una básica, que se realizaba cada ocho o nueve años) fueron sustituidas por una sola, con periodicidad trimestral, que proporcionaba una información más cercana a la encuesta básica, en cuanto al nivel de desagregación. Esta encuesta, denominada Encuesta Continua de Presupuestos Familiares (ECPF), proporcionó la información necesaria para la actualización de las ponderaciones así como la renovación de la composición de la cesta de la compra en el cambio de base del IPC 2001. Y, además, posibilitó la actualización permanente de dichas ponderaciones y la revisión de la cesta de la compra, lo que supuso una mejora en los cambios de Sistema del IPC.

El IPC base 2001 se actualizó con la revisión permanente de su sistema metodológico, mejorándolo contactando con los distintos foros académicos y organismos productores nacionales e internacionales. Pero, también, es más dinámico que sus predecesores en la medida en que anualmente revisa las

ponderaciones para ciertos niveles de desagregación funcional e incluye en el plazo más breve posible cualquier cambio detectado en los componentes del mercado, ya sea la aparición de nuevos productos, cambios en la estructura de consumo o en la muestra de municipios o establecimientos. Además, establece los cambios de base cada cinco años, realizando una revisión completa de la metodología y la muestra y la actualización de ponderaciones a todos los niveles de desagregación.

Como consecuencia de este nuevo esquema de funcionamiento, en enero de 2007, entra en vigor el Sistema de Índices de Precios de Consumo, con base de referencia en el año 2006. Este Sistema sustituye al IPC que, con base 2001, estuvo vigente hasta diciembre de 2006. El IPC, base 2006, mantiene las principales características del IPC, base 2001, y, al igual que éste, revisa anualmente las ponderaciones para cierto nivel de desagregación funcional. Para realizar esta actualización utiliza la información proporcionada por la nueva Encuesta de Presupuesto Familiares (EPF) que, desde el año 2006, sustituye a la ECPF-97 y cuya principal característica es su periodicidad anual. Asimismo la información proporcionada por esta nueva encuesta también se utilizará en los cambios de base posteriores al año 2008.

El Índice de Precios de Consumo, que se publica mensualmente, tiene como objetivo medir la evolución del nivel de precios de los bienes y servicios de consumo adquiridos por los hogares residentes en España. En el Sistema Base 2006 se utiliza la siguiente definición de gasto de consumo de la EPF: "el gasto de consumo es el flujo monetario que destina el hogar y cada uno de sus miembros al pago de determinados bienes y servicios, con destino al propio hogar o para ser transferidos gratuitamente a otros hogares o instituciones".

Las aplicaciones del IPC son numerosas y de gran importancia en los ámbitos económico, jurídico y social. Entre ellas cabe destacar su utilización como medida de la inflación. También se aplica en la revisión de los contratos de arrendamiento de inmuebles, como referencia en la negociación salarial, en la fijación de las pensiones, en la actualización de las primas de seguros y otros tipos de contrato, y como deflactor en la Contabilidad Nacional.

El **campo de consumo** es el conjunto de los bienes y servicios que los hogares del estrato de referencia (población que reside en viviendas familiares en España) destinan al consumo; por lo tanto no se consideran los gastos en bienes de inversión, los autoconsumos y autosuministros, ni los alquileres imputados, ni los gastos subvencionados por las administraciones públicas. Tampoco forman parte del campo de consumo algunos impuestos no considerados consumo desde el punto de vista de la EPF ni otros gastos, como los destinados a loterías y juegos de azar.

La **cesta de la compra** es el conjunto de los bienes y servicios seleccionados en el IPC cuya evolución de precios representa la de todos aquellos que componen la parcela COICOP a la que pertenecen. La selección de los artículos que componen la cesta de la compra se realiza a partir del IPC, base 2001, y los datos de la ECPF 2004-2005. El criterio para determinar qué parcelas deben estar incluidas sigue siendo el mismo que para la base 2001: se han tenido en cuenta todas aquellas parcelas que superan el 0,3 por mil del gasto total. Una vez determinadas las parcelas de gasto que están representadas en el

índice, se revisan los artículos que componen la cesta de la compra de la base 2001, aumentando, disminuyendo o manteniendo los artículos de cada parcela, en función de la ponderación de ésta y de la variabilidad de los precios de dichos artículos. Así, el número total de artículos que componen la cesta de la compra del IPC base 2006 es 491. Para cada uno de los artículos se elabora su descripción o especificación con el fin de facilitar su identificación por parte del encuestador y permitir la correcta recogida de los precios. Estas especificaciones tienen en cuenta las particularidades propias de cada región.

El IPC base 2006 se adapta completamente a la clasificación internacional de consumo COICOP. Los artículos de la cesta de la compra se agregan en subclases, éstas en clases, posteriormente en subgrupos, y por último, los subgrupos en grupos. La estructura funcional del IPC consta de 12 grupos, 37 subgrupos, 79 clases y 126 subclases. Además, se mantienen las 57 rúbricas y los 28 grupos especiales existentes en el IPC base 2001.

Los artículos están distribuidos en los doce grandes grupos siguientes:

1	Alimentos y bebidas no alcohólicas	176
2	Bebidas alcohólicas y Tabaco	12
3	Vestido y calzado	67
4	Vivienda	18
5	Menaje	60
6	Medicina	13
7	Transporte	31
8	Comunicaciones	3
9	Ocio y cultura	43
10	Enseñanza	7
11	Hoteles, cafés y restaurants	23
12	Otros bienes y servicios	38
Total		491

Como en la mayoría de los países de la Unión Europea (UE), el diseño de la muestra de los precios que intervienen en el cálculo del IPC es intencional, y por tanto se trata de un diseño no probabilístico, dadas las características de la población objeto de estudio.

2.1. Selección de la muestra

Para obtener indicadores significativos en todos los niveles de desagregación funcional y geográfica para los que se publica el IPC, se estructura el proceso de selección de la muestra en tres grandes apartados, cada uno de los cuales tiene como objetivo la selección de los diferentes componentes de la misma, a saber: municipios, zonas comerciales y establecimientos y artículos.

La selección de los **municipios** que forman parte del nuevo Sistema de IPC se realiza atendiendo a criterios demográficos y a la representatividad geográfica. Los datos oficiales de población que se han utilizado para realizar la selección de municipios son los obtenidos de la revisión del Padrón

Municipal de Habitantes a 1 de enero de 2003. Se parte de los criterios demográficos que se utilizaron en el IPC base 2001, y se han introducido algunos adicionales, con el fin de obtener indicadores representativos para cada nivel de desagregación funcional y geográfica.

En la base 2001, el criterio de cobertura geográfica se basaba, principalmente, en la población del conjunto de municipios seleccionados. De esta forma, los municipios seleccionados debían cubrir el 30% de la población de la provincia y el 50% de la población de la comunidad autónoma. Con este criterio se iban seleccionando los municipios por tamaño hasta cumplir el requisito, sin tener en cuenta la distribución geográfica de los mismos dentro de la provincia. Además, la muestra de municipios donde se realizaba la recogida de precios de artículos de alimentación era siempre mayor que para el resto de artículos.

Para la base 2006 se ha mantenido este criterio de partida, pero se ha completado de la siguiente forma:

- representatividad geográfica: es importante que los municipios de la muestra estén repartidos por toda la provincia, evitando la concentración en determinados focos de población.
- representatividad poblacional: se hace especial hincapié en la representatividad de los municipios pequeños; hasta ahora, dado que el criterio de selección era exclusivamente poblacional, quedaba excluida de la muestra parte de la población residente en municipios de menor tamaño.
- representatividad de la cesta: todos los municipios deben contener artículos de todos los grupos; para ello se ha elaborado una cesta reducida, a partir de la cesta total, en la que se han incluido artículos de consumo básicos. Con ello, se ha aumentado considerablemente la representatividad del IPC.

Así, la muestra de municipios obtenida con los criterios antes citados consta de 177 (las 52 capitales de provincia y 125 municipios no capitales), frente a los 141 municipios de la Base 2001.

En 97 de estos 177 municipios se recogen precios de toda la cesta de artículos, en 44 se recogen precios de toda la cesta de alimentación y parte del resto de la cesta, y en los 36 restantes se recogen precios de una parte reducida de la cesta (compuesta por el 48% de los artículos). Estos últimos son los nuevos municipios que han entrado a formar parte del IPC base 2006.

Por lo tanto, en la base 2006 se ha mejorado la representatividad y diversificación poblacional, ya que de los 36 municipios nuevos, 31 tienen menos de 50.000 habitantes; se ha mejorado la cobertura de bienes y servicios, ya que en todos los municipios se recogen precios de todos los grupos de consumo.

Es importante destacar que en la práctica se cubren porcentajes de población superiores a los indicados, ya que algunos establecimientos incluidos en la muestra, como hipermercados, centros comerciales, talleres de reparaciones o tiendas de muebles, se encuentran en las afueras de los municipios o en municipios limítrofes, con lo que la población real representada en el índice es mayor que la teórica.

Para la selección del **número de establecimientos**, se utiliza como punto de partida la muestra del IPC base 2001, y se estudia la red existente de establecimientos disponibles en cada provincia, prestando especial atención a los distintos tipos y características de dichos establecimientos.

Como criterio general, el número de establecimientos que mensualmente informan de los precios de un artículo se calcula en función de la ponderación del artículo en el índice y de la variabilidad de sus precios: cuanta más ponderación y/o variabilidad de precios, mayor número de establecimientos se deben seleccionar. Además, para el cálculo del número de establecimientos, se establece un número mínimo para cada artículo en cada provincia, dependiendo del tipo de artículo y del tipo de recogida del mismo.

En relación a la selección de los tipos de establecimiento, se considera la distribución de los porcentajes de ventas por tipo de establecimiento (hipermercados, supermercados, mercados y tiendas especializadas), dependiendo de cada artículo. Para ello se cuenta con información de diversas fuentes, entre ellas la Encuesta de Comercio (INE) y el Ministerio de Agricultura, Pesca y Alimentación.

Se dedica especial atención a los centros comerciales, hipermercados y supermercados, dada su importancia en cuanto al volumen de ventas, si bien este aspecto ya se recogía en el IPC base 2001. En muchos casos la situación de dichos centros, así como la presencia de mercados, condiciona la creación de las "zonas comerciales" que de forma explícita se definen, en cada municipio de la muestra, para los artículos perecederos de alimentación (carnes, pescados, frutas y hortalizas frescas), e implícitamente para el resto de artículos.

Para los artículos perecederos de alimentación se definen tres tipos de áreas comerciales, atendiendo al tamaño del municipio y al número de establecimientos susceptibles de ser seleccionados para los distintos tipos de artículos considerados. Estos artículos se clasifican en dos grandes grupos atendiendo a la variabilidad de los precios que presentan y al peso que tienen en la cesta de la compra. Esta clasificación determina el número de establecimientos en los que se recogen precios según el tipo de zona comercial y el tipo de artículo de que se trate.

Para el resto de artículos, aunque no se ha efectuado la delimitación estricta de las zonas comerciales, la selección de establecimientos se hace cumpliendo el objetivo de representatividad: la muestra de establecimientos debe representar, con la evolución de los precios de los artículos que en ellos se venden, a todos los establecimientos de la localidad.

Partiendo de las premisas anteriores, la selección de los establecimientos informantes se realiza tomando en consideración las siguientes normas básicas:

- En la muestra deben estar representadas todas las zonas comerciales y los distintos tipos de establecimientos que existen.
- Los establecimientos deben ser los más frecuentes y los de mayor afluencia de público en la localidad, y/o los de mayor volumen de venta.
- Los establecimientos deben ser representativos del tipo de artículo del cual se recoge información.
- En cada establecimiento no se pueden recoger, en el mismo día, más de un precio del mismo artículo.
- Un establecimiento no debe concentrar un número importante de observaciones de precios de diferentes artículos. Se intenta evitar con esto que la política de precios de un solo establecimiento pueda condicionar la evolución del índice.
- No forman parte de la muestra los establecimientos de acceso restringido a un sector de la población como cooperativas, economatos o establecimientos

similares. Tampoco se considera la venta ambulante ni la venta a domicilio.
- Los establecimientos seleccionados han de ofrecer suficientes garantías de continuidad en la venta de artículos de los cuales se recogen precios, ya que esta muestra permanecerá fija a lo largo del tiempo, salvo que se produzca el cierre, el cambio de actividad, la pérdida de representatividad en lo que al consumo se refiere o se deje de comercializar el artículo del cual se recogían precios. En estos casos, el establecimiento será sustituido por otro que cumpla los requisitos necesarios para pertenecer a la muestra.

En el IPC base 2006 se han seleccionado aproximadamente 33.000 establecimientos con estas características, repartidos por todo el territorio nacional.

Para seleccionar los **artículos** representativos de las parcelas de gasto de la ECPF se consultan diferentes organismos, asociaciones de empresarios, fabricantes, comerciantes y establecimientos, los cuales facilitan información de aquellos artículos que mejor representan las distintas parcelas, de acuerdo con los siguientes criterios de selección:

- La evolución de los precios de los artículos seleccionados debe ser similar a la del resto de los artículos de la parcela a la que representan.
- Los artículos deben ser los consumidos habitualmente por la población.
- Deben tener precios que sean fácilmente observables.
- Deben ofrecer garantías razonables de permanencia en el mercado.

Así, en el IPC base 2006, la cesta de la compra está compuesta por 491 artículos, frente a los 484 con los que se elaboraba el IPC base 2001. Se han eliminado artículos cuyo consumo o representatividad había disminuido, como el tejido para confección o la reparación de ciertos electrodomésticos, y se han incorporado nuevos artículos como los relacionados con los productos dietéticos e infantiles y nuevos servicios relacionados con la salud, entre otros.

2.2. Cálculo

La fórmula empleada para calcular los índices del IPC, base 2006, es la fórmula de Laspeyres encadenado, que se empezó a utilizar en el IPC, base 2001. El índice general correspondiente al mes m del año t se expresa matemáticamente del siguiente modo, y, al ser encadenado establece comparaciones entre el periodo corriente (t) y el periodo base (0), pero considerando las situaciones intermedias (k).

$$I_0^t = \prod_{k=1}^{t} \frac{\sum_i p_i^k q_i^{k-1}}{\sum_i p_i^{k-1} q_i^{k-1}}$$

- El principal inconveniente de los índices encadenados es la falta de aditividad. Esto hace que no sea posible obtener el índice de cualquier agregado como media ponderada de los índices de los agregados que lo componen. Así, por ejemplo, el índice general no se puede calcular como media ponderada de los índices de los doce grupos.
- En el IPC español se calcula un índice elemental para cada artículo de la cesta de la compra en cada una de las provincias, por lo que el agregado elemental es el artículo-provincia.
  - El índice del agregado elemental i se obtiene como cociente del precio medio de dicho agregado elemental en el periodo actual y el

precio medio en el periodo de referencia de los precios, es decir, diciembre del año anterior.

$I^{mt}_{dic(t-1),i} = \frac{\overline{p}^{mt}_{i}}{\overline{p}^{dic(t-1)}_{i}} x100$, donde:

$I^{mt}_{dic(t-1),i}$ es el índice, referido a diciembre del año (*t-1*), del agregado elemental *i*, en el mes *m* del año *t*.

$\overline{p}^{mt}_{i}$ es el precio medio del agregado elemental *i*, en el mes *m* del año *t*.

$\overline{p}^{dic(t-1)}_{i}$ es el precio medio del agregado elemental i, en diciembre del año (*t-1*).

- El precio medio del agregado *i*, en el periodo (*m,t*) es la media geométrica simple de los precios recogidos en dicho periodo y otorga la misma importancia a las variaciones de todos los precios, independientemente del nivel de los mismos.
- Las ponderaciones que intervienen en el cálculo de los índices agregados provienen de la EPF.

  En el cálculo de las mismas, para los artículos que componen la cesta de la compra del IPC, base 2006, ha sido necesaria la desagregación de estas parcelas para obtener información más detallada. Los datos empleados en el cálculo de las ponderaciones, utilizadas durante el año 2007, son los correspondientes a los años 2004 y 2005. Para obtener el gasto total se calculó una media aritmética ponderada de las estructuras anuales, donde el segundo año tiene asignado un peso mayor que el primero por estar más próximo al momento de actualización.

Además, para corregir el desfase que se producía entre este periodo de referencia de las ponderaciones y el periodo de referencia de los precios (diciembre del año 2006) se actualizaron las ponderaciones mediante la utilización de información sobre evolución de precios y de cantidades, procedente del IPC y de otras fuentes. De esta forma, el periodo de referencia de las ponderaciones, utilizadas durante el año 2007, es diciembre de 2006.

Las ponderaciones de cada artículo representan la relación entre el gasto realizado en las parcelas representadas por dicho artículo y el gasto total realizado en todas las parcelas cubiertas por el índice:

Wi = gasto realizado en las parcelas representadas por el artículo *i* / gasto total.

Estas ponderaciones son diferentes en cada una de las agregaciones geográficas (provincias, comunidades autónomas y conjunto nacional) y a partir de ellas se obtienen las ponderaciones de las distintas agregaciones funcionales. Así, la ponderación de un agregado funcional se obtiene como suma de las ponderaciones de los artículos que componen dicha agregación. Las actualizaciones anuales de ponderaciones, que se llevan a cabo en el IPC base 2006, se realizan con la última información anual disponible de la nueva ECP.

En líneas generales, los índices agregados se pueden obtener como:

- Agregaciones funcionales dentro de una provincia: el índice, referido a diciembre del año anterior, de cualquier agregación funcional en una provincia, se obtiene como agregación de los índices elemen-

tales de los artículos pertenecientes a dicha agregación con las ponderaciones vigentes en el año del índice. Una vez calculados los índices agregados como se ha detallado anteriormente, es preciso encadenarlos. Estos índices son los que finalmente se difunden y dan continuidad a las series publicadas en base 2006.

- Agregaciones geográficas de una agregación funcional: De la misma forma que en el caso anterior, se calculan índices de una agregación geográfica superior a la provincia para una agrupación funcional.

Se calculan las siguientes tasas de variación:

- La **tasa de variación mensual** de un índice en un periodo: es el cociente entre el índice del mes corriente y el índice del mes anterior.
- La **tasa de variación acumulada** (o en lo que va de año): es el cociente entre el índice del mes corriente y el índice de diciembre del año anterior.
- La **tasa de variación anual:** cociente entre los índices publicados del mes corriente y del mismo mes del año anterior, ambos en base 2006.

La repercusión de la variación mensual de un artículo o conjunto de artículos en el índice general se define como la parte de la variación mensual del índice general que corresponde a dicho artículo o conjunto de artículos. Por tanto, la suma de las repercusiones mensuales de todos los artículos de la cesta de la compra es igual a la variación mensual del índice general. Esto significa que la repercusión que la variación mensual de precios de un artículo o conjunto de artículos tiene en la variación mensual del índice general, es la variación que éste habría experimentado si todos los precios del resto de artículos hubieran permanecido estables ese mes. La fórmula de la repercusión mensual de un artículo (o agregado funcional) determinado i, en el mes m del año t, es la siguiente:

$$R_i^{mt/(m-1)t} = \frac{I_{dic(t-1),i}^{mt} - I_{dic(t-1),i}^{(m-1)t}}{I_{dic(t-1),G}^{(m-1)t}} x W_{dic(t-1),i} x100$$, donde:

$I_{dic(t-1),i}^{mt}$ es el índice, referido a diciembre del año (t-1), del artículo i, en el mes m del año t.

$I_{dic(t-1),G}^{(m-1)t}$ es el índice general, referido a diciembre del año (t-1), en el mes (m-1) del año t.

La repercusión de la variación en lo que va de año (o variación acumulada) de un artículo o conjunto de artículos en el índice general, representa la variación acumulada que experimentaría el índice general si el resto de artículos no hubieran sufrido variación de precios alguna en lo que va de año. O lo que es lo mismo, es la parte de la variación acumulada debida a dicho artículo o conjunto de artículos. La fórmula de la repercusión acumulada de un artículo (o agregado funcional) determinado i en el mes m del año t es la siguiente:

$$R_i^{mt/dic(t-1)} = \frac{I_{dic(t-1),i}^{mt} - I_{dic(t-1),i}^{dic(t-1)}}{I_{dic(t-1),G}^{dic(t-1)}} x W_{dic(t-1),i} x100$$, donde:

$I^{mt}_{dic(t-1),i}$ es el índice, referido a diciembre del año (t-1), del artículo i, en el mes m del año t.

$W_{dic(t-1),i}$ es la ponderación, referida a diciembre del año (t-1), del artículo i, en tanto por uno.

2.3. Cesta de la compra

Las operaciones incluidas en el proceso de cálculo del IPC, desde la recogida de precios hasta el cálculo de los índices, son diferentes en función de las particularidades propias de cada artículo de la cesta de la compra. Así, la periodicidad de la recogida de los precios varía según la frecuencia con la que se modifican los precios de los artículos. La forma de recogerlos también es diferente dependiendo de la homogeneidad geográfica de los precios y de la disposición de los mismos. Por último, según las características de cada artículo, el método de cálculo de los índices es diferente.

En la siguiente tabla se muestran los diferentes tipos de artículo según los criterios utilizados en su clasificación:

Tabla 2. Criterios de clasificación de los artículos de la cesta de la compra

Criterios	Tipos de artículo			
Periodicidad en la recogida de precios	grupos 1 y 2	mensual	perecedero	estacional
				no estacional
			no perecedero	
	grupos del 3 al 12	mensual		
		trimestral	trimestral	
			trimestral de rebajas	
Lugar de recogida y grabación de precios	provincias			
	servicios centrales			
Método de cálculo de índices elementales	estacionales			
	de recogida centralizada			
	de recogida provincial con precio elaborado			
	alquiler de vivienda			
	sin tratamiento especial			

- Según la periodicidad y frecuencia en la recogida de precios se establece una primera forma de clasificar los artículos. Así, se consideran dos tipos de artículos, los mensuales y los trimestrales.

  - Artículos de recogida mensual: Los precios de los artículos mensuales se observan todos los meses en todos los establecimientos de la muestra, mediante visita personal, en la mayoría de los casos. En general, cada establecimiento se visita una vez al mes, excepto aquellos en donde se recogen precios de artículos perecederos que, debido a la mayor frecuencia

con que varían sus precios, el entrevistador visita dos o tres veces al mes, dependiendo del municipio. En esta categoría se encuentran los alimentos frescos no elaborados, que sufren fluctuaciones periódicas en sus precios y presentan constantes cambios de calidad. Dentro de éstos, se encuentran los artículos estacionales (frutas y hortalizas frescas), algunos de los cuales sólo se comercializan unos determinados meses del año; para estos artículos sólo se lleva a cabo la recogida de precios en los meses en que están disponibles. Dentro de los artículos de recogida mensual destacan los artículos de temporada, que son aquellos cuyo consumo tiene lugar únicamente algunos meses del año, ya que el resto del mismo no se comercializan. En el IPC se consideran de temporada los artículos de vestido y calzado, que tienen 2 temporadas definidas (primavera-verano y otoño-invierno). La recogida de precios de estos artículos se realiza una vez al mes durante la temporada en que se comercializan y el tratamiento de los precios los meses que el artículo desaparece es la repetición del último precio recogido.

- Artículos de recogida trimestral: Los artículos trimestrales son aquellos cuyos precios tienen un comportamiento bastante estable, es decir, no suelen experimentar muchas variaciones de precios a lo largo del tiempo (electrodomésticos, muebles, servicios de reparaciones, etc.). La recogida trimestral permite ampliar el número de precios recogidos con el mismo coste. El tratamiento de estos precios consiste en dividir la muestra de establecimientos seleccionados en tres submuestras, de modo que cada mes sólo se visitan los establecimientos de una de ellas y se repite el último precio recogido en los establecimientos de las otras dos submuestras. Con ello se consigue que todos los meses haya establecimientos que informen sobre los precios de estos artículos. Además, en caso de que varíen más de la mitad de los precios recogidos en el mes, el mes siguiente se solicitará información en todos los establecimientos. Con la inclusión de los precios rebajados, a partir de enero de 2002, se estableció una nueva categoría de artículos dentro de los trimestrales: los trimestrales de rebajas. Son aquellos que, a pesar de cumplir con el requisito de la estabilidad de precios propia de los trimestrales, muestran variaciones significativas en periodos típicos de rebajas (electrodomésticos, muebles, ropa de cama, etc.). Por ello, la recogida de precios en los meses de rebajas se realiza visitando todos los establecimientos de la muestra y no sólo los de la submuestra correspondiente a ese mes.

➢ Según el lugar donde se recogen y graban los precios, se puede distinguir entre artículos de recogida provincial y artículos de recogida centralizada. Los precios de los primeros se recogen en cada provincia, por medio de visita personal, a través del teléfono o fax, o de los boletines oficiales de las comunidades autónomas o las provincias, y se graban en las delegaciones provinciales.

Por otro lado, el seguimiento de los precios de los artículos de recogida centralizada se realiza desde los Servicios Centrales del INE. Forman parte de este

tipo de artículos aquellos bienes y servicios que tienen una o varias de las siguientes características:

- sus precios son los mismos en una amplia zona geográfica,
- sus precios están sujetos a tarifas publicadas en el BOE,
- existen pocas empresas que comercializan el artículo,
- se dispone de un directorio perfectamente definido de informantes,
- son artículos con cambios de calidad habituales (como los artículos tecnológicos), lo que conlleva dificultad para realizar ajustes de calidad; al hacerse su recogida de forma centralizada se homogeneiza el tratamiento de estos ajustes.

- En lo que respecta al método de cálculo, existen ciertos grupos de artículos cuyos índices elementales se obtienen de forma diferente a la fórmula general descrita anteriormente. Como ya se ha especificado, el índice elemental de cualquier artículo de la cesta de la compra se obtiene como media simple de los precios recogidos, sin tener en cuenta ningún tipo de ponderación. La excepción a esta norma general la constituye la fórmula de cálculo de los artículos de recogida centralizada y los artículos con precio elaborado; en ambos casos, el índice elemental se calcula teniendo en cuenta un conjunto de variedades o modalidades representativas del artículo, ponderadas adecuadamente por el gasto realizado en cada una de ellas.

En el IPC base 2006, la ponderación de cada modalidad se obtiene a partir del gasto realizado por los consumidores. De esta forma, se mantiene la coherencia con la estructura general de ponderaciones de la cesta de la compra. Otro conjunto de artículos que, por sus características, recibe un tratamiento especial son los estacionales. Debido a las oscilaciones periódicas de sus precios y cantidades, los índices de las frutas, verduras y hortalizas frescas se calculan con un método diferente, basado en medias móviles, que tiene en cuenta los calendarios de producción y comercialización de las mismas. Por último, y debido a las características especiales del mercado, el alquiler de vivienda también recibe un tratamiento diferenciado de la fórmula general del IPC.

- La recogida de precios de los artículos se realiza tanto en provincias como en Servicios Centrales, mediante visita personal de los agentes del INE a los establecimientos en las fechas correspondientes, con la excepción de algunos de ellos para los que, por sus características especiales, la recogida de la información se realiza por teléfono, fax, correo electrónico, catálogo o Internet. La recogida se realiza mediante un cuestionario generado automáticamente para cada establecimiento, en el que el entrevistador anota los precios e incidencias relativos a los artículos que aparecen en el mismo. Cada establecimiento es visitado por un solo entrevistador, excepto los hipermercados y las grandes superficies. Los precios recogidos son precios efectivos de venta al público con pago al contado.

La recogida de precios de un mismo artículo en los distintos establecimientos informantes, se ha distribuido a lo largo de ese periodo para recoger el mayor número posible de fluctuaciones de precios. Todos los meses se visitan los establecimientos seleccionados aproxima-

damente el mismo día; con ello se pretende que la variación reflejada por el índice corresponda a una variación mensual. Como los artículos perecederos están sujetos a fluctuaciones importantes de precios, sus precios se recogen tres veces a lo largo del mes en cada uno de los establecimientos seleccionados en todas las capitales de provincia, manteniendo una distancia de al menos siete días entre las tres visitas al establecimiento. En el resto de municipios, se recogen los precios de estos artículos dos veces en cada uno de los establecimientos que pertenecen a la muestra. Para el resto de artículos cada establecimiento se visita una sola vez al mes, a excepción de los artículos trimestrales, para los cuales la recogida de precios en cada establecimiento se realiza una vez cada tres meses.

Se recogen los precios que han sufrido reducciones por motivos tales como ofertas y promociones, así como aquellos cuyos descuentos son debidos a los periodos oficiales de rebajas. Esto afecta a la mayoría de las parcelas que componen el IPC, aunque los descuentos por rebajas se producen de forma más acentuada en las parcelas de Vestido y calzado y Menaje, donde éstas son más habituales.

Los criterios, seguidos en la base 2006, para la recogida de precios con descuento son los siguientes:

- que el descuento se realice sobre artículos que se espera estén disponibles de nuevo a sus precios habituales, es decir, no se trata de descuentos por liquidaciones o saldos;
- que el descuento se realice sobre artículos que puedan ser adquiridos por todos los consumidores, no sólo por una parte de ellos (por ejemplo, no se tendrán en cuenta descuentos realizados por tener tarjetas de fidelidad del establecimiento o por cumplir determinadas condiciones);
- y que estos descuentos sean efectivos en el momento de la compra (por ejemplo, no se consideran los reembolsos posteriores a la compra).

Se recogen, por tanto, descuentos debidos a:

- Las rebajas de temporada (periodos de rebajas oficiales regidos por la Ley de Ordenación de Comercio Minorista).
- Ofertas de cualquier tipo (siempre que no se trate de liquidaciones o saldos).

➢ Un aspecto muy relevante en cualquier IPC es el ajuste que se debe realizar sobre los precios cuando hay un cambio en la variedad del artículo o en el establecimiento, ya que el IPC tiene como objetivo recoger la evolución de los precios de los mismos productos a lo largo del tiempo, sin que ésta se vea influida por dichos factores. Estos ajustes se conocen como ajustes por cambio de calidad. Los cambios de calidad son un problema al que se enfrentan todos los países, y que en los últimos años se ha visto acentuado por el rápido progreso técnico que han experimentado algunos artículos. Por ello, es uno de los temas de los que, con mayor prioridad, se ocupa EUROSTAT, en el ámbito de la armonización de los IPC de los países de la UE.

En la elaboración del IPC español, en las distintas bases, han sido varios los procedimientos que se han utilizado para la estimación de los cambios de calidad. La elección de estos métodos ha venido determinada por la disponibilidad de información en cada momento y por el tipo de artículo de que se trate.

Un ajuste por cambio de calidad es necesario cuando un artículo (producto, variedad o modalidad), cuyo precio forma parte del cálculo del IPC, se sustituye por otro, y en ese momento es necesario determinar qué parte de la diferencia de precios entre el artículo sustituto y el sustituido se debe a una calidad diferente entre los mismos.

Las sustituciones de los artículos pueden deberse a varios motivos: cuando deja de ser representativo y surge otro más representativo en el mercado; cuando desaparece del mercado y cuando el establecimiento donde se recoge el precio del artículo deja de ser representativo, cierra o cambia de actividad económica.

En definitiva, los métodos de ajuste de calidad utilizados de forma más habitual en el IPC, base 2006, son los siguientes:

- Ajuste total de calidad. Parte del supuesto de que la diferencia entre el precio del artículo sustituido y del artículo sustituto está totalmente motivada por la diferencia de calidad entre ambos, o que los artículos son tan diferentes que no se pueden comparar. Se considera, entonces, que la diferencia de precios entre ambos artículos es debida únicamente a la distinta calidad de los mismos, con lo que el índice no reflejará variación de precios. Con este ajuste se supone que de haber seguido a la venta el artículo sustituido, su precio no habría variado.
- Ajuste por calidad idéntica. Se parte de la idea de que el artículo sustituto tiene la misma calidad que el artículo sustituido, es decir, que la diferencia de precios existente entre ambos se debe a una variación real de precios. Con este ajuste se supone que de haber seguido a la venta el artículo sustituido, su precio habría sido el mismo que el del artículo sustituto.
- Otros ajustes. Se incluyen en este apartado todos los ajustes para los cuales se estima el valor de la diferencia de calidad entre un artículo y su sustituto.

Hasta el IPC base 2006, en los cambios de Sistema de Índices de Precios de Consumo se producía una ruptura en la continuidad de las series. La actualización de ponderaciones, la composición de la nueva cesta de la compra y especialmente, los cambios metodológicos, hacían que la serie nueva difiriera de la antigua. Estas diferencias, desde el punto de vista teórico eran insalvables. No obstante, la necesidad de disponer de series continuadas por parte de los usuarios hizo necesario el cálculo de unos coeficientes de enlace que unían las series publicadas en base antigua con las series en base nueva. Sin embargo, para el IPC, base 2006, por tratarse de un índice encadenado, no ha sido necesario calcular ningún coeficiente de enlace, ya que el método de cálculo del encadenamiento permite realizar cambios en ponderaciones, muestra y metodología cada mes de diciembre y encadenar los índices obtenidos con los nuevos cálculos, con la serie que se venía publicando calculada con muestra, ponderaciones y metodología antigua.

Así, en el IPC, base 2006, sólo se ha cambiado el periodo de referencia de los índices o periodo base, que ha pasado de ser el año 2001 a ser el año 2006. Para ello se ha calculado un coeficiente de re-escala, que ha convertido los índices publicados en base 2001, desde enero de 2002 hasta diciembre de 2006, en índices en base 2006.

Este coeficiente es aquel que hace que la media aritmética simple de índices publicados del año 2006, en base 2001, sea igual a 100:

$$\frac{1}{12}\sum_{m=1}^{12} I_{01}^{m06} x\, C_{re-escala} = 100 \qquad \Rightarrow \qquad C_{re-escala} = \frac{100}{\frac{1}{12}\sum_{m=1}^{12} I_{01}^{m06}}$$

Multiplicando la serie publicada en base 2001 por este coeficiente de re-escala, se obtiene una serie de índices en base 2006, que conserva las tasas de variación publicadas, y con la que se han encadenado los nuevos índices en base 2006, calculados a partir de enero de 2007.

2.4. El Índice de Precios de Consumo (Base 2011)

El IPC base 2011 se publica en febrero del año 2011 con el objetivo de mejorar la representatividad de este indicador mediante cambios en la composición de la cesta de la compra y la actualización de la estructura de ponderaciones. Introduce un nuevo tratamiento para los artículos estacionales (frutas frescas y verduras y hortalizas frescas), permitiendo realizar una medición más precisa de la evolución de precios en el corto plazo para este tipo de productos.

2.4.1. Cambios en la cesta de la compra

Los cambios más relevantes en la cesta de la compra están relacionados con los bienes y servicios relativos a los soportes para el registro de imagen y sonido y con el material para el tratamiento de la información. Así, se incorporan los discos duros portátiles y se excluye el CD grabable y el alquiler de película. En lo referente a materiales para el tratamiento de la información la nueva base incluye los *notebooks* y las *tablets*.

También cabe destacar la incorporación en la cesta de la compra de nuevos servicios de estética, como la fotodepilación y la depilación láser, y paramédicos, como el logopeda.

Como consecuencia de estos ajustes, la cesta de la compra del IPC base 2011 pasa a tener **489 artículos**, frente a los 491 de la base anterior.

2.4.2. Actualización de las ponderaciones

La continua adaptación del IPC a los cambios en el comportamiento de los consumidores incluye también la revisión permanente de su estructura de ponderaciones. Cada año se actualiza el peso o importancia de los grandes agregados que componen este indicador, lo que mantiene la actualidad del mismo.

Además de la revisión anual de las ponderaciones para los grandes agregados, cada cinco años se actualiza la estructura completa para todos los niveles de desagregación. Así pues, el IPC base 2011 incluye una nueva estructura de ponderaciones que representa de forma más precisa las pautas de consumo de los hogares. Para su elaboración se ha considerado la EPF como fuente principal de información, y otras fuentes como la evolución del consumo privado de la Contabilidad Nacional, la evolución de precios del IPC y otras fuentes de diferentes sectores.

En la siguiente tabla se incluye el peso de cada uno de los 12 grandes grupos y su comparación con los pesos vigentes hasta el año 2011.

Tabla 3. Ponderación de los grupos de artículos de la cesta de la compra

	Ponderaciones de grupos (%)		
Grupo	2011	2012	%
01. Alimentos y bebidas no alcohólicas	18,16	18,26	0,6
02. Bebidas alcohólicas y tabaco	2,87	2,89	0,7
03. Vestido y calzado	8,59	8,34	-2,9
04. Vivienda	11,70	00	2,6
05. Menaje	6,84	6,67	-2,5
06. Medicina	3,21	3,14	-2,1
07. Transporte	14,74	15,16	2,9
08. Comunicaciones	3,98	3,85	-3,3
09. Ocio y Cultura	7,64	7,54	-1,3
10. Enseñanza	1,38	1,42	2,8
11. Hoteles, cafés y restaurantes	11,52	11,46	-0,5
12. Otros bienes y servicios	9,37	9,26	-1,2
TOTAL	100	100	

3. LOS CENSOS DEMOGRÁFICOS 2011

Los Censos Demográficos son el proyecto estadístico de mayor envergadura que periódicamente debe acometer la oficina de estadística de cualquier país. Bajo la denominación *Censos Demográficos* se engloban en realidad tres censos diferentes: el Censo de Población, el Censo de Viviendas y el Censo de Edificios. De los tres, el Censo de Población es, sin duda, el de mayor repercusión y el de más amplia tradición.

El primer censo moderno de población, entendiendo como tal el que utiliza a la persona como unidad de análisis, se realizó en España en 1768 por el Conde de Aranda bajo el reinado de Carlos III. También son de destacar por su interés el Censo efectuado en 1787 por Floridablanca y el realizado diez años más tarde por Godoy en tiempos de Carlos IV. No obstante, la serie de censos de la organización estadística oficial se inicia en 1857 con el primero de la Comisión General de Estadísticas del Reino, al que siguió, en un lapso inusualmente corto, el de 1860. Después vinieron los de 1877, 1887 y 1897. A partir de 1900 ha habido Censo de Población cada diez años sin excepción alguna.

El Censo de Población de 2011 ha sido el decimoséptimo de los Censos oficiales realizados en España. Su realización se enmarca dentro del Programa Mundial 2010 que abarca el periodo 2005-2014 promovido por Naciones Unidas y en cuyo marco se han finalizado, en el momento de redactar este proyecto, los censos de 121 países. Continuando con las actuaciones internacionales que impulsan la realización de los censos, hay que destacar que por primera vez se ha desarrollado una reglamentación comunitaria. El reglamento 763/2008 del Parlamento Europeo y del Consejo (junto con otros que lo desarrollan), además de implantar la obligatoriedad de realizar el Censo durante el año 2011, asegura la comparabilidad de los resultados a nivel de la Unión Europea.

El marco metodológico general en el que se desarrolla el proyecto de censo para España está fijado por las recomendaciones de la Conferencia de Estadísticos Europeos para la ronda censal de 2010 y con un mayor nivel de concreción por el reglamento antes citado y los tres reglamentos de la Comisión que desarrollan el anterior (sobre definiciones de variables y clasificaciones, sobre hipercubos de datos y sobre calidad de la operación).

La Reglamentación desarrollada por la Unión Europea contempla un amplio rango de opciones posibles para recopilar la información de las variables censales. Ese rango va desde los censos clásicos basados en una recogida exhaustiva de los datos, hasta un censo basado en información tomada exclusivamente de registros administrativos. Entre ambos extremos cita un número de situaciones intermedias generadas por el mayor o menor peso de la recogida de datos en campo y de los registros administrativos. Entre ellas figura expresamente el modelo de un censo basado en registros administrativos completado con una encuesta por muestreo. Este ha sido el modelo censal para España en 2011. De hecho, España, con el Padrón Municipal como registro de población consolidado, se sitúa entre los países con mejores condiciones para realizar un censo de estas características.

La introducción de elementos como la georreferenciación de los edificios, el aprovechamiento de la abundante información administrativa y la recogida de datos multicanal (Internet, entre ellos) son algunos de los ejes sobre los que se construye el primer *censo basado en registros y encuesta por muestreo* en España.

Para su elaboración, se llevó a cabo una recopilación de información procedente de diversas fuentes estadísticas y administrativas que permitieron desarrollar esta estrategia formando un directorio territorial inicial y acumulando datos relativos a personas para su posterior uso. Asimismo, se realizó un recorrido del territorio para completar y contrastar la información territorial disponible, enumerando las viviendas y recogiendo las variables de edificios, disponiendo al final de este recorrido de un directorio completo de la operación.

El Censo de 2011 se planteó como una operación basada en la combinación de los siguientes elementos:

- Un "**fichero precensal**" realizado a partir de un aprovechamiento máximo de los registros administrativos disponibles, tomando al Padrón como elemento básico de su estructura.
- Un trabajo de campo que incluye dos grandes operaciones:
  - Un **Censo de Edificios** exhaustivo que permite la georreferenciación de todos

los edificios y conocer sus características.
- Una **gran encuesta por muestreo**, dirigida a un porcentaje relativamente alto de la población para conocer el resto de características de las personas y las viviendas.

Algunos de los aspectos sustanciales en la estrategia del Censo 2011 han sido:

- La combinación del fichero precensal con la información obtenida de la encuesta proporciona toda la información censal. En particular, la cifra de población se obtiene mediante el recuento de los registros que contenga el fichero precensal, ponderados –cuando sea necesario– con unos factores de recuento obtenidos a partir de la encuesta.
- En tanto que no hay operación de campo exhaustiva, no se prevé que el Censo se utilice para introducir rectificaciones en las inscripciones padronales. El Censo tiene, por tanto, fines exclusivamente estadísticos. Sin embargo sí que se basa en el Padrón, como ya se ha apuntado, y sus resultados se utilizan para contrastar los datos padronales.
- La fracción muestral global está en torno al 12,3% de la población (11,9% de viviendas). Por otro lado, la distribución territorial no es uniforme, al pretender proporcionar un conjunto de tablas e indicadores para los municipios de menor tamaño y, al mismo tiempo, descender del nivel municipal en la difusión de los datos para los municipios que superen un determinado tamaño.
- Se plantea un censo completo de edificios, con especial hincapié en la enumeración de todos los inmuebles de los edificios destinados a vivienda. Análogamente a lo planeado para la población, se realiza una fase de "fichero precensal" de territorio, consistente en un cruce previo entre los datos del Censo 2001, Padrón y Catastro fundamentalmente, completado con datos de otras fuentes como los procedentes de Oficinas de Estadística de las Comunidades Autónomas. Esta fase se complementa con un recorrido para completar el cruce anterior.
- Los datos muestrales se elevan al Fichero Precensal ponderado, calibrados de forma que se reprodujeran las distribuciones marginales del mismo a nivel municipal.
- Las nuevas tecnologías juegan un papel relevante en la operación censal, utilizándose dispositivos portátiles (*tablets*) para el Censo de Edificios y para las entrevistas a la población. Además se ofrecen diversos canales de cumplimentación de la información a los ciudadanos: cuestionarios en papel para ser devueltos por correo e Internet.

3.1. Fichero precensal (FPC)

El fichero precensal se basa en un aprovechamiento máximo de los registros administrativos disponibles, tomando al Padrón como elemento de partida de su estructura al que se va asociando información de otros registros administrativos y de operaciones estadísticas. Esta operación tiene objetivos múltiples:

- Disponer de información adicional a la padronal para decidir si un registro concreto debe ser incluido en el recuento censal o no.
- Aportar información directa de variables censales.
- Servir de marco inicial para realizar una primera selección de la muestra de personas y viviendas que formarán parte de la encuesta.
- Ser el directorio de partida del recorrido del Censo de Edificios.
- Aportar información adicional para las fases de tratamiento de los datos.

El FPC abarca en sentido amplio dos aspectos: territorio y personas. El Fichero territorial intenta reflejar la situación del territorio en el momento más cercano posible a la fecha de referencia censal. Por otra parte, el contenido del fichero de personas se fundamenta sobre la base padronal.

El **Censo de Edificios** es una operación estadística coincidente en el tiempo con la fase postal de la encuesta de población. Dicha operación está diseñada para recopilar información exhaustiva de todos los edificios del territorio nacional en los que hay situada alguna vivienda, con enumeración de todos los inmuebles situados en ellos. Sus objetivos son:

- Enumerar y georreferenciar todos los edificios que tengan algún inmueble que sea una vivienda.
- Determinar las características de los edificios mediante un cuestionario de edificio.
- Enumerar todos los inmuebles contenidos en cada edificio.
- Seleccionar los inmuebles, a medida que se daban altas en el recorrido, que formarán parte de la muestra de la encuesta de población y viviendas.

Al efectuarse con carácter exhaustivo permite disponer de un directorio georreferenciado completo de edificios con alguna vivienda y de todos sus inmuebles. La georreferenciación permite la identificación del edificio ante cambios en la dirección postal y, en consecuencia, su identificación inequívoca.

Como ya se ha apuntado, el **Censo de Población y Viviendas** 2011 se basa en tres pilares: el fichero precensal, un Censo de Edificios exhaustivo y una encuesta por muestreo, para conocer las características de las personas y las viviendas, con un tamaño de muestra adecuado para cumplir con la normativa de cobertura establecida por Eurostat. Sus objetivos son:

- Estimar el total poblacional correspondiente a determinados colectivos, de forma que se corrija la cifra padronal correspondiente a los mismos.
- Estimar las características de la población y de las viviendas a distintos niveles de desagregación geográfica.

De acuerdo con lo anterior, y considerando que el censo de población es el único medio que se dispone para obtener información desagregada a nivel de sección censal, unidad primaria de muestreo utilizada en las encuestas de hogares, se selecciona una muestra en todas las secciones censales.

3.2. Selección de la muestra

El marco para la selección de la muestra ha sido el FPC obtenido del cruce de los ficheros de Padrón, Catastro y otros ficheros de tipo administrativo. Después de los distintos cruces realizados, las viviendas figuran clasificadas en el FPC como localizables o no localizables. Las primeras son aquellas que a través de la dirección postal es posible que se puedan localizar en el trabajo de campo. Las segundas son aquellas que no tienen una dirección completa y consecuentemente no pueden ser localizadas en el trabajo de campo. Durante la realización del trabajo de campo se dieron altas y bajas en las viviendas del FPC.

A efectos de selección de la muestra, el total de viviendas de cada municipio se agrupa en dos marcos: Marco A formado por el conjunto de viviendas localizables del FPC tal y como se ha definido anteriormente, y Marco B formado por el

conjunto de inmuebles que son dados de alta durante el recorrido exhaustivo que se realiza en campo.

Para alcanzar los objetivos del censo se selecciona una muestra de inmuebles diferente según se trate del marco A o B. La muestra se selecciona en todas las secciones censales.

- MARCO A: La muestra procedente de este marco selecciona de entre los inmuebles cuyo uso fuera el de vivienda, antes de comenzar los trabajos de campo. De este grupo se contacta por correo ordinario con las que según el FPC sean principales, dando la opción de cumplimentar el cuestionario censal bien por Internet o bien devolverlo cumplimentado por correo ordinario. Pasado un periodo de tiempo, del conjunto residual de viviendas que no habían respondido se seleccionó una submuestra para recoger el cuestionario censal mediante entrevista personal asistida por ordenador (CAPI). Las viviendas no principales se investigaron durante la realización del recorrido exhaustivo y aquellas que no pudieron ser resueltas durante el mismo pasaron a la muestra de recogida por CAPI. En consecuencia, se consideran dos grupos diferentes de viviendas, uno conteniendo a las principales y el otro a las no principales.
- MARCO B: La muestra procedente de este marco se selecciona durante el recorrido exhaustivo en campo, para lo cual se utiliza el procedimiento de Bernouilli de selección aleatoria. Durante dicho recorrido, a esta parte de la muestra se le deja un cuestionario en papel y el hogar pudo colaborar de la misma forma a la indicada para las viviendas del marco A.

Teniendo en cuenta los objetivos de obtener una cierta información a nivel municipal y el presupuesto disponible, el tamaño de la muestra es de, aproximadamente, tres millones de viviendas, representando una fracción de muestreo global del 11,9% por ciento. En porcentaje de población representa una fracción de muestreo del 12,3%. La distribución de esta muestra depende del tamaño del municipio, fijando unos criterios de precisión para las estimaciones. De esta forma las fracciones de muestreo varían desde las aplicadas a los municipios de menor tamaño que eran investigados exhaustivamente hasta las de los municipios de mayor tamaño a los que corresponden las menores fracciones de muestreo.

La muestra se selecciona en cada municipio, dentro de cada grupo de viviendas, con igual probabilidad. La procedente del marco A, se selecciona mediante muestreo sistemático con arranque aleatorio. La muestra del marco B se selecciona del conjunto de inmuebles que son alta en el recorrido exhaustivo utilizando el procedimiento de Bernouilli que asigna igual probabilidad a las unidades de muestreo.

Los estimadores de las características de viviendas y personas, en un determinado municipio, son estimadores de expansión con corrección de falta de respuesta a los que se aplican técnicas de calibrado, según los casos, y el factor de expansión inicial se obtiene como la inversa de la probabilidad de selección.

BIBLIOGRAFÍA

Carrasco Carpio, C. y García Serrano, C. (2012). Inmigración y mercado de trabajo. *Informe 2011*. Ministerio de Trabajo y Seguridad Social.

Censos de Población y Viviendas 2011. *Estadística y Sociedad* (2011), nº 48.

Delgado, M. (2011). El próximo censo de población de 2011. *Estadística y Sociedad*, 48.

Escuder Bueno, J. y Escuder Vallés, R. (2009).Síntesis histórica y metodológica de los Índices de Precios al Consumo españoles. *Estadística Española*, *51*(171), 257-280.

Fuentes Castro, D. (2007). Análisis del poder adquisitivo de los asalariados desde la entrada en circulación del euro. *Boletín Económico del ICE* (Información Comercial Española), nº 2926, 39-64.

García, M.A. (2005). Cambios en la encuesta de población activa en 2005. *Estadística y Sociedad*, 11.

García Villar, J. y Gómez del Moral, M. (2010). La estadística oficial como bien público en una sociedad democrática. *Estadística y Sociedad*, *43*, 5-7.

Hansen, M.H., Hurwitz, W.N. y Bershad, M.A. (1961). Measurement errors in censuses and surveys. *Bulletin of the International Statistical Institute*, *38*, 359-374.

www.ine.es/censos2011/censos2011.htm

www.ine.es/daco/daco43/notaepa.htm

www.ine.es/daco/daco43/meto_res_ipc.htm