CUADERNOS DE TRABAJO
ESCUELA UNIVERSITARIA DE ESTADÍSTICA

LA METODOLOGÍA DEL HAZ DE RECTAS PARA LA COMPARACIÓN DE SERIES TEMPORALES

Magdalena Ferrán Aranaz
Departamento de Estadística e I.O. III
Escuela Universitaria de Estadística
Universidad Complutense de Madrid

Cuaderno de Trabajo número 04/2011

UCM

UNIVERSIDAD
COMPLUTENSE
MADRID

Los Cuadernos de Trabajo de la Escuela Universitaria de Estadística constituyen una apuesta por la publicación de los trabajos en curso y de los informes técnicos desarrollados desde la Escuela para servir de apoyo tanto a la docencia como a la investigación.

Los Cuadernos de Trabajo se pueden descargar de la página de la Biblioteca de la Escuela www.ucm.es/BUCM/est/ y en la sección de investigación de la página del centro www.ucm.es/centros/webs/eest/

CONTACTO: Biblioteca de la E. U. de Estadística
Universidad Complutense de Madrid
Av. Puerta de Hierro, S/N
28040 Madrid
Tlf. 913944035
buc_est@buc.ucm.es

ISSN: 1989-0567

LA METODOLOGÍA DEL HAZ DE RECTAS PARA LA COMPARACIÓN DE SERIES TEMPORALES

Magdalena Ferrán Aranaz

Departamento de Estadística e I.O. III
Escuela Universitaria de Estadística
Universidad Complutense de Madrid

1. INTRODUCCIÓN

En este trabajo se propone una metodología, a la que denominaremos Metodología del haz de rectas, para la comparación de series temporales que midan un mismo fenómeno o variable procedentes de diferentes ámbitos, territorios, agentes, condiciones, etc. Para ilustrar el proceso de aplicación compararemos las cincuenta series temporales relativas al número de ocupados en el sector de la construcción desde el tercer trimestre de 1976 hasta el cuarto de 2010, ambos inclusive, en cada una de las provincias españolas.

Antes de proceder a la presentación de la metodología y a la exposición de los resultados teóricos asociados, introduzcamos dos conceptos.

Definición 1:

Un conjunto $\{c_{t,k}\}$ ***de*** K ***series temporales distintas, todas ellas definidas en los mismos instantes,*** $t = 1, \ldots, T$ ***, tiene estructura de haz de*** K ***rectas si existe otra serie*** x_t ***tal que para cada*** $c_{t,k}$ ***existen cuatro coeficientes*** b_k , m_k , B_0 ***y*** B_1 ***, con al menos*** m_k ***diferente de cero, tales que:***

$$c_{t,k} = b_k + m_k \cdot x_t \qquad \forall t,k \qquad \textbf{donde} \qquad b_k = B_0 + B_1 \cdot m_k \qquad \forall k$$

Obsérvese que:

$$c_{t,k} = c_{t,k'} \quad sys \quad b_k + m_k \cdot x_t = b_{k'} + m_{k'} \cdot x_t \quad sys \quad (m_k - m_{k'}) \cdot x_t = -B_1 \cdot (m_k - m_{k'})$$

En consecuencia[1]:

[1] Dadas dos series distintas $c_{t,k}$ y $c_{t,k'}$, se verifica $m_k \neq m_{k'}$.

$$c_{t,k} = c_{t,k'} \quad sii \quad x_t = -B_1$$

En dicho caso:

$$c_{t,k} = B_0$$

y el vértice:

$$(x_t, c_{t,k}) = (-B_1, B_0)\,.$$

En el caso particular de que el coeficiente b_k sea igual a cero $\forall k$ el vértice del haz coincidirá con el origen de coordenadas:

$$(x_t, c_{t,k}) = (0,0)$$

Definición 2:

Sea $\{Y_{t,j}\}$, $\mathrm{j}=1,\ldots,\mathrm{J}$***, un conjunto de*** J ***series temporales distintas, todas ellas definidas en los mismos instantes,*** $\mathrm{t}=1,\ldots,\mathrm{T}$***. Diremos que la estructura que subyace en el conjunto*** $\{Y_{t,j}\}$ ***es la de un haz de rectas si existe otra serie*** X_t ***tal que, por un lado, para cada una de las*** J ***series es adecuado el ajuste de la ecuación de regresión***:

$$\hat{Y}_{t,j} = A_{0,j} \cdot t + A_{1,j} \cdot X_t + A_{2,j} \qquad \mathrm{j}=1,\ldots,\mathrm{J}\,,$$

y, por otro, o bien la secuencia de coeficientes $(A_{0,1},...,A_{0,J})$ ***es nula o bien su grado de asociación lineal con la secuencia*** $(A_{1,1},...,A_{1,J})$ ***es estadísticamente significativo.***

Obsérvese que, si B_0 y B_1 son los coeficientes de la ecuación de regresión:

$$\hat{A}_{0,j} = B_0 + B_1 \cdot A_{1,j} \qquad \mathrm{j}=1,\ldots,\mathrm{J}$$

entonces, según la **Definición 1**, el conjunto de series $\{\hat{y}_{t,j}\}$, donde:

$$\hat{y}_{t,j} = \hat{A}_{0,j} + A_{1,j} \cdot x_t \qquad \mathrm{j}=1,\ldots,\mathrm{J} \qquad \text{siendo} \qquad x_t = \nabla X_t = X_t - X_{t-1}\,,$$

tiene estructura de haz de J rectas de vértice:

$$(x_t, \hat{y}_{t,j}) = (-B_1, B_0)\,.$$

Ilustremos estos dos conceptos mediante un sencillo ejemplo que, a su vez, permitirá introducir la metodología. Sean $Y_{t,1}$, $Y_{t,2}$, $Y_{t,3}$, $Y_{t,4}$ cuatro series temporales (**Fig. 1, A** y **B**) e Y_t la serie promedio. Supongamos, por un lado, que para cada una de las cuatro series, el coeficiente R_j^2 correspondiente a la ecuación de regresión:

$$\hat{Y}_{t,j} = A_{0,j} \cdot t + A_{1,j} \cdot Y_t + A_{2,j}$$

es estadísticamente significativo; por ejemplo, para $j = 1$ (**Fig. 1**, **C**):

$$\hat{Y}_{t,1} = 2{,}51 \cdot t + 1{,}61 \cdot Y_t + (-157) \qquad \text{con} \qquad R_1^2 = 0{,}983$$

Supongamos también que la asociación lineal entre las dos secuencias de coeficientes $(A_{0,1},...,A_{0,4})$ y $(A_{1,1},...,A_{1,4})$ es estadísticamente significativa. Si denominamos $\hat{A}_{0,j}$ al valor ajustado de $A_{0,j}$ mediante la ecuación de regresión lineal (**Fig. 1**, **C**):

$$\hat{A}_{0,j} = B_0 + B_1 \cdot A_{1,j} = -24{,}19 + 24{,}19 \cdot A_{1,j} \qquad \text{con} \qquad R^2 = 0{,}59$$

entonces, según la **Definición 1**, el conjunto de las cuatro series:

$$\hat{y}_{t,j} = \hat{A}_{0,j} + A_{1,j} \cdot y_t \qquad \text{j} = 1,\ldots,4 \qquad \text{con} \qquad y_t = \nabla Y_t = Y_t - Y_{t-1},$$

tiene estructura de haz de J rectas de vértice (**Fig. 1**, **D)**:

$$(y_t, \hat{y}_{t,j}) = (-B_1, B_0) = (-24{,}19, -24{,}19)$$

En otras palabras, según la **Definición 2**, la estructura que subyace en el conjunto $\{Y_{t,j}\}$ es la de un haz de rectas.

Así como el valor $\hat{y}_{t,j}$ (**Fig. 1**, **E)** se puede contemplar como una estimación del valor $y_{t,j}$ (**Fig. 1**, **F)**, también el valor:

$$\hat{Y}_{t,j} = \hat{A}_{0,j} \cdot t + A_{1,j} \cdot Y_t + \hat{A}_{2,j}$$

donde $\hat{A}_{2,j}$ es aquel valor tal que:

$$\bar{\hat{Y}}_j = \bar{Y}_j \qquad \text{con} \qquad \bar{\hat{Y}}_j = \frac{1}{10} \cdot \sum_{t=1}^{10} \hat{Y}_{t,j} \qquad \text{e} \qquad \bar{Y}_j = \frac{1}{10} \cdot \sum_{t=1}^{10} Y_{t,j}$$

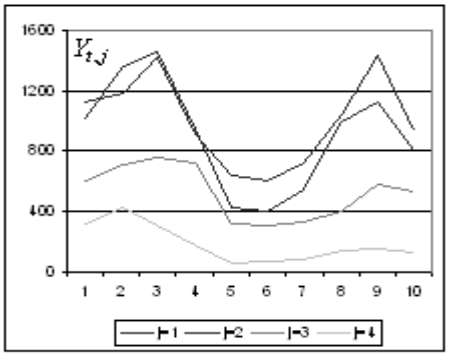
se puede contemplar como una estimación del valor $Y_{t,j}$; por ejemplo, para $j = 1$ (**Fig. 1**, **C**):

$$\hat{Y}_{t,1} = \hat{A}_{0,1} \cdot t + A_{1,1} \cdot Y_t + \hat{A}_{2,1} = 14{,}67 \cdot t + 1{,}61 \cdot Y_t + (-223{,}9)$$

A

t	$Y_{t,1}$	$Y_{t,2}$	$Y_{t,3}$	$Y_{t,4}$	Y_t
1	1018	1123	594	319	763,5
2	1360	1187	710	425	920,5
3	1454	1415	753	304	981,5
4	941	913	720	180	688,5
5	423	629	323	47	355,5
6	400	604	299	66	342,25
7	544	716	325	74	414,75
8	995	1035	404	131	641,25
9	1116	1433	581	148	819,5
10	797	935	527	128	596,75
Media	**904,8**	**999**	**523,6**	**182,2**	**652,4**

B



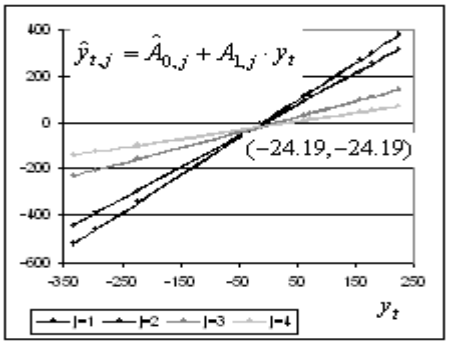
C

j	R_j^2	$A_{0,j}$	$A_{1,j}$	$A_{2,j}$	$\hat{A}_{0,j}$	$\hat{A}_{2,j}$
1	0,983	2,51	1,61	-157,0	14,67	-223,9
2	0,934	23,80	1,35	-13,9	8,51	70,2
3	0,843	-7,98	0,66	135,3	-8,16	136,3
4	0,894	-18,32	0,38	35,6	-15,02	17,4

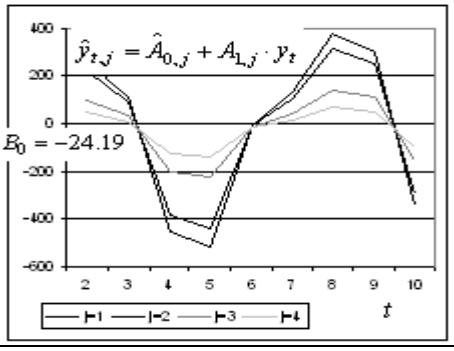
$$\hat{A}_{0,j} = B_0 + B_1 \cdot A_{1,j} = -24{,}19 + 24{,}19_1 \cdot A_{1,j}$$

$$R^2 = 0{,}59$$

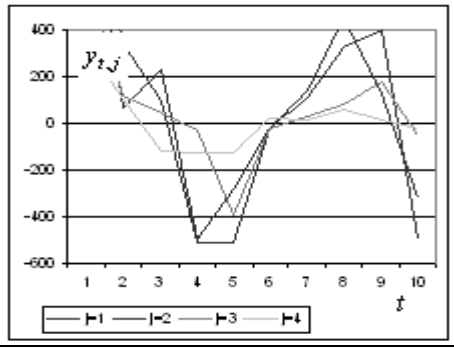
D



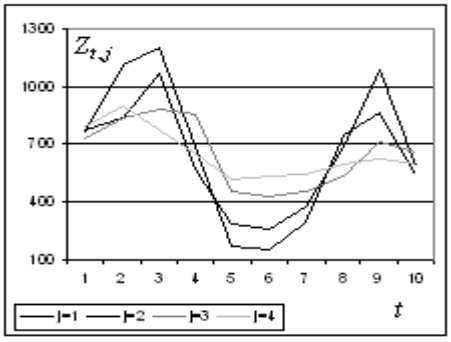
E



F



G



H

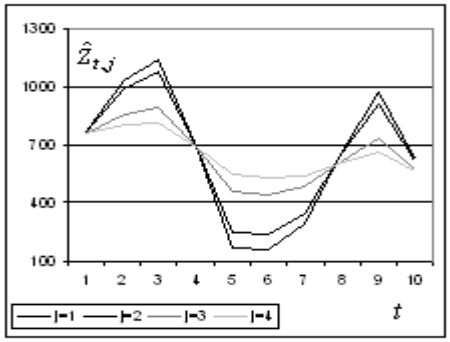


Fig. 1 Ejemplo introductorio.

Fuente: Elaboración propia

Si homogeneizamos la escala de las cuatro series (por ejemplo, al nivel de la serie promedio, **Fig. 1**, **G**):

$$Z_{t,j} = Y_{t,j} - \overline{Y}_j + \overline{Y} \qquad j = 1,...,4 \qquad \text{con} \qquad \overline{Y} = \frac{1}{10} \cdot \sum_{t=1}^{10} Y_t$$

entonces también:

$$\hat{Z}_{t,j} = \hat{Y}_{t,j} - \overline{Y}_j + \overline{Y}$$

se puede contemplar como estimación de $Z_{t,j}$ (**Fig. 1**, **H**).

Obsérvese que, respecto de una tendencia constante ($B_0 = -24{,}19$), las series del conjunto $\{\hat{y}_{t,j}\}$ presentan fluctuaciones más a menos pronunciadas dependiendo del coeficiente $A_{1,j}$ (**Fig. 1**, **E**). Si el conjunto $\{\hat{y}_{t,j}\}$ reproduce el patrón de comportamiento del conjunto $\{y_{t,j}\}$ (**Fig. 1**, **F**), cabe suponer que existe una variable x_t responsable de las fluctuaciones ya que, por su naturaleza, la serie promedio no puede serlo. En otras palabras, si respecto de la serie promedio la estructura que subyace en el conjunto $\{Y_{t,j}\}$ es la de un haz de rectas, cabe suponer que es debido a que existe una serie X_t que la genera, aunque en la práctica no es necesario conocerla.

La metodología que se describe en este trabajo, a la que denominaremos "Metodología del haz de rectas", se fundamenta en la hipótesis de que la estructura que subyace en el conjunto de series objeto de análisis es la de un haz de rectas: dado un conjunto $\{Y_{t,j}\}$ (**Fig. 2,** *sup. izqda.*) que mide el mismo fenómeno o variable en distintos ámbitos, territorios, etc., como paso previo al proceso de construcción de las series resumen comprobaremos si la metodología es aplicable. La comparación se realizará sobre el correspondiente conjunto de series homogeneizadas, $\{Z_{t,j}\}$ (**Fig. 2,** *sup. dcha.*). Para ello, del conjunto $\{\hat{Z}_{t,j}\}$ (**Fig. 2,** *centro izqda.)* se "extraerá un subconjunto de series resumen"; concretamente, considerando que cada serie $\hat{Z}_{t,j}$ procede de un punto $(A_{1,j}, \hat{A}_{0,j})$ situado en un segmento de la recta $y = B_0 + B_1 \cdot x$ (**Fig. 2,** *centro dcha.*), las series resumen se construirán a partir de K puntos representativos de este mismo segmento (**Fig. 2,** *inf. izqda.)*:

$$(m_k, b_k), \quad \text{donde} \quad b_k = B_0 + B_1 \cdot m_k, \quad \mathrm{k} = 1, \ldots, \mathrm{K}:$$

En definitiva, la expresión de las series resumen (**Fig. 2,** *inf. dcha.)* vendrá dada por:

$$C_{t,k} = b_k \cdot t + m_k \cdot Y_t + \mu_k \qquad \mathrm{k} = 1, \ldots, \mathrm{K}$$

siendo μ_k aquel valor que sitúe la k-ésima serie resumen a la misma escala que las J series homogeneizadas.

En lo que sigue se ofrece una serie de resultados teóricos relacionados con la construcción del conjunto de series resumen.

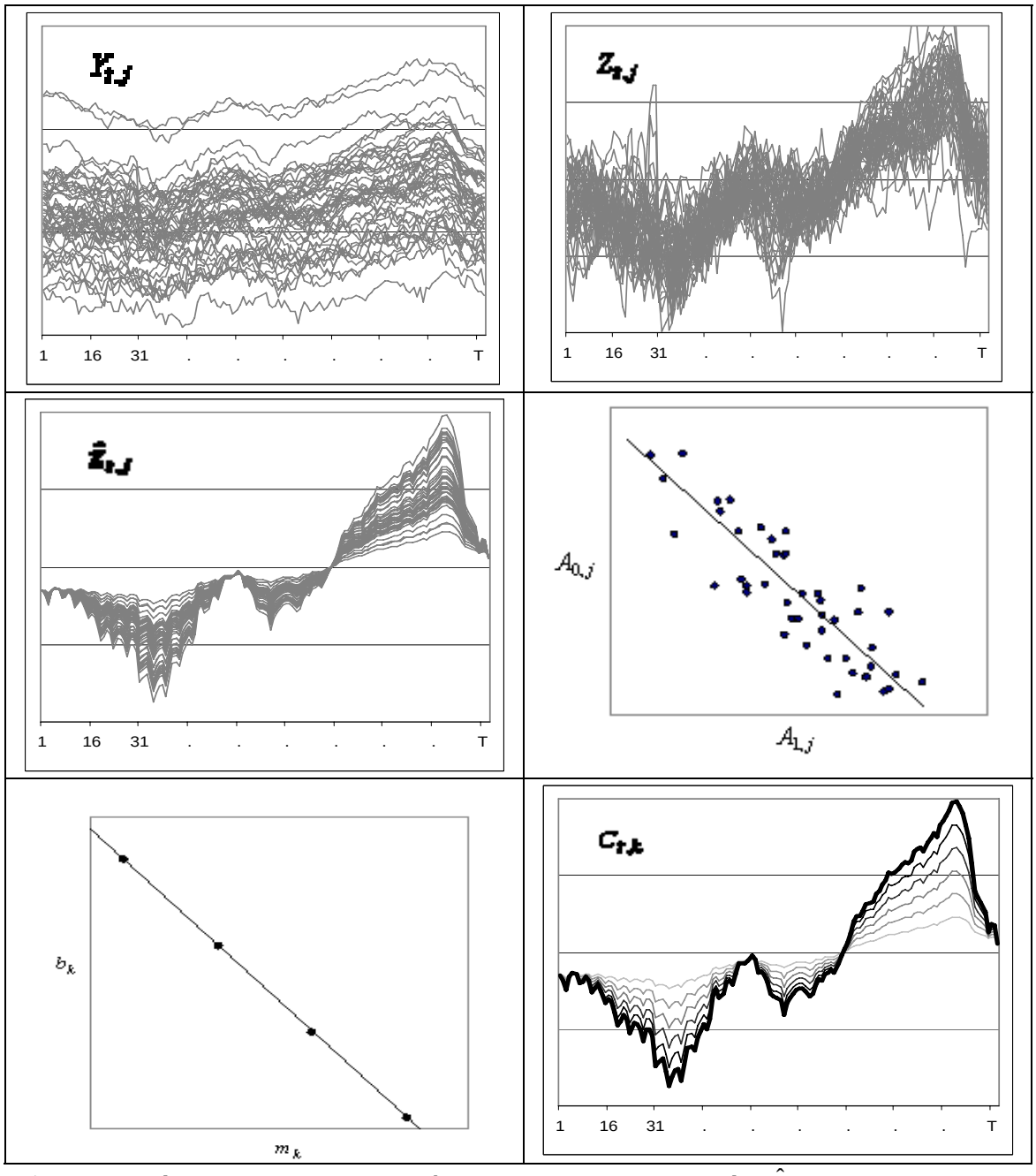


Fig. 2 *Sup. Izqda.*: $Y_{t,j}$, $j = 1,\ldots,J$; *Sup. Dcha.*: $Z_{t,j}$, $j = 1,\ldots,J$; *Ctro. Izqda.*: $\hat{Z}_{t,j}$, $j = 1,\ldots,J$; *Ctro. Dcha*: $A_{0,j}$ versus $A_{1,j}$, $j = 1,\ldots,J$; *Inf. Izqda*: b_k versus m_k, $j = 1,\ldots,J$; *Inf. Dcha*: $C_{t,k}$, $k = 1,\ldots,K$.

Fuente: Elaboración propia

2. FUNDAMENTOS TEÓRICOS DE LA METODOLOGÍA DEL HAZ DE RECTAS

Proposición 1:

Sea $\{C_{t,k}\}$, $\mathrm{k}=1,\ldots,\mathrm{K}$, ***un conjunto de*** K ***series temporales distintas, todas ellas definidas en los mismos instantes temporales,*** $\mathrm{t}=1,\ldots,\mathrm{T}$, ***y con la misma media:***

$$\overline{C}_k = \frac{1}{T}\cdot\sum_{t=1}^{T} C_{t,k} = \alpha \qquad \forall k \qquad \textbf{[1]}$$

Sea X_t ***otra serie temporal y supongamos que para cada*** $C_{t,k}$ ***existen cinco coeficientes*** b_k, m_k, μ_k, B_0 ***y*** B_1 ***con al menos*** m_k ***distinto de cero, tales que:***

$$C_{t,k} = b_k \cdot t + m_k \cdot X_t + \mu_k \qquad \forall t,k \qquad \textbf{[2]}$$

siendo:

$$b_k = B_0 + B_1 \cdot m_k \qquad \forall k \qquad \textbf{[3]}$$

Entonces **(Fig. 2,** ***inf. dcha.):***

- ***A) Si*** $C_{t,q}$, $C_{t,r}$ ***y*** $C_{t,s}$ ***son tres series temporales cualesquiera del conjunto*** $\{C_{t,k}\}$ ***tales que*** $m_q < m_r < m_s$ ***entonces*** $d(C_{t,q},C_{t,r}) < d(C_{t,q},C_{t,s})$, ***donde*** d ***es la distancia euclidea.***
- ***B) Para cualquier par de series temporales del conjunto*** $\{C_{t,k}\}$ ***existe al menos un punto en su trayectoria[2] en el que se cortan. Además, los puntos de corte de cualquier par de trayectorias son los puntos de corte de todas ellas.***
- **C)** ***Si las trayectorias se cortan en más de un punto entonces la diferencia entre dos puntos de corte cualesquiera es independiente de la media de las series temporales.***

Dem. de A):

Si denominamos: $\bar{t} = \frac{1}{T}\cdot\sum_{t=1}^{T} t = \frac{T+1}{2}$ y $\overline{X} = \frac{1}{T}\cdot\sum_{t=1}^{T} X_t$

Las expresiones [2] y [3] implican que:

$$\frac{1}{T}\sum_{t=1}^{T} C_{t,k} = \frac{1}{T}\sum_{t=1}^{T}(b_k \cdot t + m_k \cdot X_t + \mu_k) = b_k \cdot \bar{t} + m_k \cdot \overline{X} + \mu_k =$$
$$= (B_0 + B_1 \cdot m_k)\cdot \bar{t} + m_k \cdot \overline{X} + \mu_k = B_0 \cdot \bar{t} + (B_1 \cdot \bar{t} + \overline{X})\cdot m_k + \mu_k$$

[2] Utilizaremos el término trayectoria para referirnos a la línea continua que conecta la secuencia de puntos en la representación gráfica de la serie.

entonces, por [1]:

$$\alpha = B_0 \cdot \bar{t} + (B_1 \cdot \bar{t} + \overline{X}) \cdot m_k + \mu_k \qquad \forall k$$

Luego:

$$\mu_k = D_0 + D_1 \cdot m_k \qquad \forall k \qquad \textbf{[4]}$$

donde:

$$D_0 = \alpha - B_0 \cdot \bar{t} \qquad \text{y} \qquad D_1 = -B_1 \cdot \bar{t} - \overline{X}$$

En consecuencia, [2] también puede expresarse como:

$$C_{t,k} = D_0 + B_0 \cdot t + (B_1 \cdot t + X_t + D_1) \cdot m_k \qquad \textbf{[5]}$$

Entonces, si $m_k < m_{k'}$:

$$d(C_{t,k}, C_{t,k'}) = \left(\sum_{t=1}^{T} (D_0 + B_0 \cdot t + (B_1 \cdot t + X_t + D_1) \cdot m_{k'} - (D_0 + B_0 \cdot t + (B_1 \cdot t + X_t + D_1) \cdot m_k))^2 \right)^{1/2} =$$

$$= \left(\sum_{t=1}^{T} (B_1 \cdot t + X_t + D_1)^2 \cdot (m_{k'} - m_k)^2 \right)^{1/2} = (m_{k'} - m_k) \cdot \left(\sum_{t=1}^{T} (B_1 \cdot t + X_t + D_1)^2 \right)^{1/2} = (m_{k'} - m_k) \cdot \varepsilon$$

donde:

$$\varepsilon = \left(\sum_{t=1}^{T} (B_1 \cdot t + X_t + D_1)^2 \right)^{1/2}$$

es un valor constante positivo independiente de t y k. Luego, si $m_q < m_r < m_s$ entonces:

$$d(C_{t,q}, C_{t,r}) = (m_r - m_q) \cdot \varepsilon < (m_s - m_q) \cdot \varepsilon = d(C_{t,q}, C_{t,s})$$

En otras palabras, bajo las hipótesis [1], [2] y [3], las series temporales del conjunto $\{ C_{t,k} \}$ pueden ser ordenadas en términos de su similitud: si suponemos que la secuencia de coeficientes m_k está ordenada[3]:

$$m_1 < m_2 < ... < m_K$$

entonces la secuencia $C_{t,1}, C_{t,2}, ..., C_{t,K}$ es tal que:

$$d(C_{t,k}, C_{t,k+1}) < d(C_{t,k}, C_{t,k'}) \qquad k = 1, ..., K-2 \qquad k' = k+2, ..., K$$

$$d(C_{t,k}, C_{t,k-1}) < d(C_{t,k}, C_{t,k'}) \qquad k = 3, ..., K \qquad k' = 1, ..., k-2$$

Dem. de B):

Si dos trayectorias no tuvieran ningún punto de corte una de ellas tomaría siempre valores mayores que la otra:

[3] Ver nota a pie de página 1.

$$C_{t,k} > C_{t,k'} \quad \forall t$$

y en dicho caso:

$$\overline{C}_k = \frac{1}{T} \cdot \sum_{t=1}^{T} C_{t,k} > \frac{1}{T} \cdot \sum_{t=1}^{T} C_{t,k'} = \overline{C}_{k'}$$

en contradicción con la hipótesis [1]. En consecuencia, existe al menos un punto en el que las trayectorias de las dos series temporales se cortan. Además, si t es un instante de corte[4], entonces:

$$C_{t,k} - C_{t,k'} = 0$$

y, por [5]:

$$(B_1 \cdot t + X_t + D_1) \cdot (m_k - m_{k'}) = 0$$

Dado que[5]:

$$m_k \neq m_{k'} \qquad \forall k' \neq k$$

entonces en el instante t:

$$B_1 \cdot t + X_t + D_1 = 0$$

y, en consecuencia:

$$(B_1 \cdot t + X_t + D_1) \cdot (m_k - m_{k'}) = 0 \qquad \forall k' \neq k$$

Luego, por [5], en el instante t:

$$C_{t,k} = C_{t,k'} \qquad \forall k' \neq k$$

Es decir, un punto de corte de dos trayectorias cualesquiera es un punto de corte de todas ellas.

Dem. de C):

Como hemos visto en la demostración de **B)**, si t es un instante en el que todas las trayectorias se cortan, entonces:

$$B_1 \cdot t + X_t + D_1 = 0$$

y, por [5]:

$$C_{t,k} = D_0 + B_0 \cdot t \qquad \forall k$$

Luego, si t' es otro instante de corte, la diferencia**:**

$$C_{t,k} - C_{t',k} = B_0 \cdot (t - t')$$

es independiente de la media de las series temporales.

[4] El punto de intersección puede darse entre dos observaciones consecutivas; en dicho caso el valor de t estaría en el segmento temporal delimitado por los dos instantes correspondientes. A pesar de ello, con el fin de no hacer más compleja la nomenclatura, utilizaremos la misma denominación.

[5] Ver nota a pie de página 1.

Observación 1:

Si denominamos:

$$c_{t,k}^{s} = C_{t,k} - C_{t-s,k} \quad \textbf{y} \quad x_t^s = X_t - X_{t-s}\textbf{,}$$

las hipótesis [2] y [3] implican que, respecto de x_t^s***, el conjunto de*** K ***series temporales {*** $c_{t,k}^{s}$ ***},*** $\mathrm{k} = 1, \ldots, \mathrm{K}$ ***, tiene estructura de haz de*** K ***rectas de vértice:***

$$(x_t^s, c_{t,k}^{s}) = (-s \cdot B_1 , s \cdot B_0)\,.$$

En el caso particular de que el coeficiente b_k ***sea igual a cero*** $\forall k$ ***el vértice coincidirá con el origen.***

Dem:

Por definición de $c_{t,k}^{s}$ y de x_t^s, y por [2]:

$$c_{t,k}^{s} = s \cdot b_k + m_k \cdot x_t^s \qquad \textbf{[6]}$$

donde:

$$s \cdot b_k = s \cdot B_0 + s \cdot B_1 \cdot m_k \qquad \forall k$$

En consecuencia, respecto de x_t^s, el conjunto $\{ c_{t,k}^{s} \}$ tiene estructura de haz de K rectas de vértice:

$$(x_t^s, c_{t,k}^{s}) = (-s \cdot B_1 , s \cdot B_0)$$

Observación 2:

Si C_t ***es la serie promedio:***

$$C_t = \frac{1}{K} \cdot \sum_{k=1}^{K} C_{t,k}$$

las hipótesis [2] y [3] implican que la serie:

$$c_t^{s} = C_t - C_{t-s}$$

es tal que existen dos coeficientes b ***y*** m ***tales que:***

$$c_t^{s} = s \cdot b + m \cdot x_t^s \qquad \textbf{donde} \qquad b = B_0 + B_1 \cdot m$$

Dem.:

Por definición de C_t y por [2]:

$$C_t = \frac{1}{K}\sum_{k=1}^{K}(b_k \cdot t + m_k \cdot X_t + \mu_k) = \frac{1}{K}\sum_{k=1}^{K} b_k \cdot t + \frac{1}{K}\sum_{k=1}^{K} m_k \cdot X_t + \frac{1}{K}\sum_{k=1}^{K} \mu_k$$

Si denominamos:

$$b = \frac{1}{K}\sum_{k=1}^{K} b_k \ , \qquad m = \frac{1}{K}\sum_{k=1}^{K} m_k \qquad \text{and} \qquad \mu = \frac{1}{K}\sum_{k=1}^{K} \mu_k$$

entonces:

$$C_t = b \cdot t + m \cdot X_t + \mu \qquad \textbf{[7]}$$

Luego:

$$c_t^s = s \cdot b + m \cdot x_t^s \qquad \textbf{[8]}$$

donde, por [3]:

$$b = B_0 + B_1 \cdot m \qquad \textbf{[9]}$$

Observación 3:

Las hipótesis [2] y [3] implican que, respecto de c_t^s***, el conjunto de*** K ***series temporales {*** $c_{t,k}^s$ ***},*** $\mathrm{k} = 1, \ldots, \mathrm{K}$ ***, tiene estructura de haz de*** K ***rectas de vértice:***

$$(c_t^s, c_{t,k}^s) = (s \cdot B_0, s \cdot B_0)$$

Dem.:

Por [6] y [8]:

$$c_{t,k}^s = s \cdot b_k^c + m_k^c \cdot c_t^s \qquad \textbf{[10]}$$

donde:

$$m_k^c = \frac{m_k}{m} \qquad \text{y} \qquad b_k^c = b_k - b \cdot m_k^c \qquad \textbf{[11]}$$

Entonces, por [3] y [9]:

$$s \cdot b_k^c = s \cdot B_0 - s \cdot B_0 \cdot m_k^c \qquad \textbf{[12]}$$

Luego, respecto de c_t^s, el conjunto $\{ c_{t,k}^s \}$ tiene estructura de haz de K rectas. Además:

$$c_{t,k}^s = c_{t,k'}^s \quad \textit{sii} \quad s \cdot b_k^c + m_k^c \cdot c_t^s = s \cdot b_{k'}^c + m_{k'}^c \cdot c_t^s \quad \textit{sii} \quad c_t^s = s \cdot B_0$$

En tal caso, por [10] y [12]:

$$c_{t,k}^s = s \cdot B_0$$

Observación 4:

Bajo las hipótesis [1], [2] y [3] los puntos de corte de las trayectorias de las series temporales del conjunto $\{C_{t,k}\}$ ***son también los puntos de corte de la trayectoria de la serie promedio con cualquiera de ellas.***

Dem:

Si t es el instante correspondiente al corte de las trayectorias de todas las series $C_{t,k}$, entonces:

$$C_{t,k} = C_{t,k'} \qquad \forall k' \neq k$$

Luego, en el instante t:

$$C_t = \frac{1}{K} \cdot \sum_{k=1}^{K} C_{t,k} = \frac{K}{K} \cdot C_{t,k} = C_{t,k} \qquad \forall k$$

Observación 5:

Bajo las hipótesis [1], [2] y [3]:

$$C_{t,k} = b_k^c \cdot t + m_k^c \cdot C_t + \mu_k^c \qquad \forall t,k \qquad \textbf{[13]}$$

donde b_k^c ***y*** m_k^c ***son los coeficientes previamente mencionados y:***

$$\mu_k^c = \mu_k - \mu \cdot m_k^c$$

Dem:

Por [7]:

$$X_t = -\frac{b}{m} \cdot t + \frac{1}{m} \cdot C_t - \frac{\mu}{m}$$

Luego, por [2]:

$$C_{t,k} = b_k \cdot t + m_k \cdot \left(-\frac{b}{m} \cdot t + \frac{1}{m} \cdot C_t - \frac{\mu}{m} \right) + \mu_k = \left(b_k - b \cdot \frac{m_k}{m} \right) \cdot t + \frac{m_k}{m} \cdot C_t + \left(\mu_k - \mu \cdot \frac{m_k}{m} \right)$$

y por [11]:

$$C_{t,k} = b_k^c \cdot t + m_k^c \cdot C_t + \mu_k^c$$

Observación 6:

Bajo las hipótesis [1], [2] y [3], si $C_{t,q}$, $C_{t,r}$ ***y*** $C_{t,s}$ ***son tres series temporales cualesquiera del conjunto*** $\{C_{t,k}\}$ ***tales que*** $m_q^c < m_r^c < m_s^c$ ***entonces:***

$$d(C_{t,q}, C_{t,r}) < d(C_{t,q}, C_{t,s})$$

donde d ***es la distancia euclidea.***

Dem:

Si $m_q^c < m_r^c < m_s^c$ entonces**,** por [11]:

$$m_q < m_r < m_s \quad (m > 0) \qquad \text{o} \qquad m_s < m_r < m_q \quad (m < 0)$$

Luego, por **A)**:

$$d(C_{t,q}, C_{t,r}) < d(C_{t,q}, C_{t,s})$$

En otras palabras, si suponemos que la secuencia de coeficientes $m_1^c, m_2^c, ..., m_K^c$ está ordenada de menor a mayor valor:

$$m_1^c < m_2^c < ... < m_K^c$$

entonces la secuencia $C_{t,1}, C_{t,2}, ..., C_{t,K}$ es tal que:

$$d(C_{t,k}, C_{t,k+1}) < d(C_{t,k}, C_{t,k'}) \qquad k = 1, ..., K-2 \qquad k' = k+2, ..., K$$

$$d(C_{t,k}, C_{t,k-1}) < d(C_{t,k}, C_{t,k'}) \qquad k = 3, ..., K \qquad k' = 1, ..., k-2$$

Observación 7:

Las hipótesis [2] y [3] implican que en aquellos instantes en que dos de las trayectorias $c_{t,k}^s$ ***se cortan lo hacen a la altura del valor*** $s \cdot B_0$ ***; además, en dichos instantes y a dicha altura, se cortan las restantes trayectorias así como la de*** c_t^s **.**

Dem:

Es consecuencia directa de la Observación 3.

Observación 8:

Por las **Observaciones 1** y **3**, el hecho de que el conjunto $\{ c_{t,k}^s \}$ tenga estructura de haz de K rectas respecto de una serie temporal x_t^s significa que también la tiene respecto de la serie promedio c_t^s. Cabe suponer entonces que si, respecto de la serie promedio, la estructura que subyace en un conjunto $\{ Y_{t,j} \}$ es la de un haz de rectas es porque existe una serie X_t que la genera. En la práctica, para construir el conjunto de series resumen, no será necesario conocer esta serie generadora, en su lugar se considerará la serie promedio Y_t .

3. LA METODOLOGÍA DEL HAZ DE RECTAS

3.1 CONDICIONES DE APLICACIÓN

Sea $\{O_{t,j},\ t=1,...,T=138\}$, $J=1,...,50$, el conjunto de las series temporales relativas al número de ocupados en el sector de la construcción en cada una de las cincuenta provincias españolas desde el tercer trimestre de 1976 hasta el cuarto de 2010, ambos inclusive (**Fig. 3**, *sup. izqda.*), sea $\{Y_{t,j}=\ln O_{t,j}\}$ el conjunto de series objeto de análisis[6] y sea Y_t la correspondiente serie promedio:

$$Y_t=\frac{1}{50}\cdot\sum_{j=1}^{50}Y_{t,j}$$

El objetivo es construir un conjunto de K series temporales $\{C_{t,k}\}$ que resuman el comportamiento de las cincuenta series objeto de análisis. La aplicación de la metodología propuesta estará justificada si la estructura que subyace en el conjunto $\{Y_{t,j}\}$ es la de un haz de rectas[7]. Para comprobarlo, calcularemos los coeficientes $A_{0,j}$, $A_{1,j}$ y $A_{2,j}$ mediante el ajuste de la ecuación de regresión lineal:

$$\hat{Y}_{t,j}=A_{0,j}\cdot t+A_{1,j}\cdot Y_t+A_{2,j}\qquad j=1,...,J$$

en cada una de las cincuenta provincias.

La **Tabla 1** ofrece el valor del coeficiente R_j^2 y los valores de $A_{0,j}$ y $A_{1,j}$ para las cincuenta provincias. En todos los casos tanto R_j^2 como $A_{1,j}$ son significativamente distintos de cero.

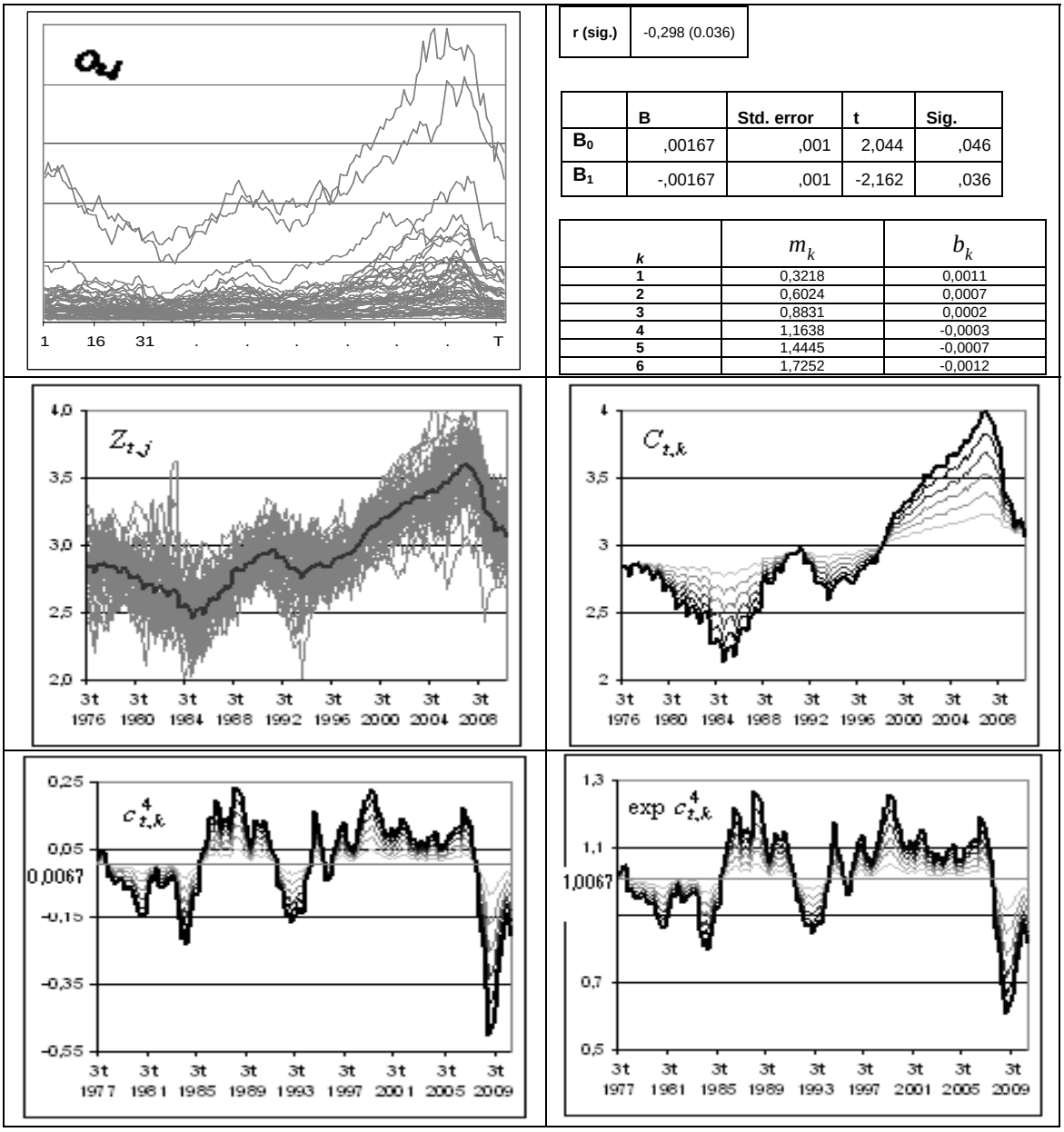
Por otro lado, el coeficiente de correlación de Pearson entre las secuencias $(A_{0,1},...,A_{0,J})$ y $(A_{1,1},...,A_{1,J})$ es estadísticamente significativo; además, al ajustar sobre la nube de J puntos $(A_{1,j},A_{0,j})$ la ecuación de regresión lineal:

$$\hat{A}_{0,j}=B_0+B_1\cdot A_{1,j}=0{,}00167-0{,}00167\cdot A_{1,j}\qquad \mathrm{j}=1,\ldots,50$$

podemos concluir que tanto B_0 como B_1 son significativamente distintos de cero (**Fig. 3**, *sup. dcha.***)**. En otras palabras, podemos concluir que la estructura que subyace en el conjunto $\{Y_{t,j}\}$ es la de haz de rectas.

[6] Así, la comparación entre el número de ocupados en el sector de la construcción en dos instantes temporales diferentes se hará en términos de su cociente (ver Sección 3.5).

[7] En principio sería necesario disponder de la serie que la genera pero, según la **Observación 8**, podemos considerar en su lugar la serie promedio.



r (sig.)	-0,298 (0.036)

	B	Std. error	t	Sig.
B_0	,00167	,001	2,044	,046
B_1	-,00167	,001	-2,162	,036

k	m_k	b_k
1	0,3218	0,0011
2	0,6024	0,0007
3	0,8831	0,0002
4	1,1638	-0,0003
5	1,4445	-0,0007
6	1,7252	-0,0012

Fig. 3 *Sup. Izqda.*: Trayectorias de las series $O_{t,j}$, j = 1,…, J ; *Sup. Dcha.*: Correlación de Pearson entre las secuencias $(A_{0,1},..., A_{0,J})$ y $(A_{1,1},..., A_{1,J})$, coeficientes B_0 y B_1 valores para m_k y b_k ; *Ctro. Izqda.*: Trayectorias de las series $Z_{t,j}$, j = 1,…, J ; *Ctro. Dcha*: Trayectorias de las curvas $C_{t,k}$, k = 1,…,6 ; *Inf. Izqda*: Trayectorias de las curvas $c^4_{t,k}$, k = 1,…,6 ;*Inf. Dcha*: Trayectorias de las curvas $\exp c^4_{t,k}$, k = 1,…,6 .

Fuente: Elaboración propia

PROVINCIA	R_j^2	$A_{0,j}$	$A_{1,j}$	$\alpha = 2{,}966$ α_j
Álava	0,757	0,0006*	0,977	1,973
Albacete	0,797	-0,0020	1,229	2,536
Alicante	0,928	0,0000*	1,516	3,867
Almería	0,877	0,0003*	**1,725**	2,903
Asturias	0,867	0,0000*	0,712	3,519
Ávila	0,641	-0,0017	0,892	2,083
Badajoz	0,627	0,0037	0,419	3,072
Balears (Illes)	0,907	0,0033	0,981	3,622
Barcelona	0,910	-0,0022	1,327	4,986
Burgos	0,788	-0,0004*	0,866	2,464
Cáceres	0,728	-0,0002*	0,686	2,883
Cádiz	0,892	-0,0029	1,573	3,478
Cantabria	0,936	0,0035	0,797	2,897
Castellón de la Plana	0,942	-0,0015	1,528	2,867
Ciudad Real	0,807	0,0022	0,709	3,102
Córdoba	0,897	-0,0028	1,495	3,011
Coruña (A)	0,680	-0,0010*	0,864	3,728
Cuenca	0,826	0,0000*	0,989	2,065
Girona	0,947	0,0004*	1,217	3,311
Granada	0,813	0,0036	0,713	3,205
Guadalajara	0,736	0,0023	0,723	1,929
Guipúzcoa	0,687	0,0000*	0,846	2,883
Huelva	0,888	-0,0023	1,511	2,688
Huesca	0,694	-0,0016	0,685	2,190
Jaén	0,606	-0,0029	0,986	2,987
León	0,830	0,0016	0,529	2,804
Lleida	0,832	0,0034	0,671	2,808
Lugo	0,547	-0,0009*	0,633	2,449
Madrid	0,952	-0,0012	1,334	5,120
Málaga	0,841	0,0004*	1,327	3,847
Murcia	0,964	0,0015	1,465	3,642
Navarra	0,927	0,0017	0,913	2,875
Orense	0,425	-0,0041	**0,322**	2,753
Palencia	0,630	-0,0020	0,791	1,733
Palmas (Las)	0,796	0,0020	1,039	3,371
Pontevedra	0,856	-0,0035	0,981	3,550
Rioja (La)	0,876	0,0009*	1,112	2,165
Salamanca	0,440	0,0001*	0,422	2,562
Santa Cruz de Tenerife	0,882	-0,0001*	1,513	3,437
Segovia	0,875	0,0017	0,843	1,762
Sevilla	0,920	-0,0012	1,435	3,801
Soria	0,647	-0,0009*	0,733	1,182
Tarragona	0,956	-0,0013	1,425	3,422
Teruel	0,809	0,0008*	0,806	1,684
Toledo	0,932	0,0032	0,744	3,244
Valencia	0,954	0,0009	1,404	4,260
Valladolid	0,788	-0,0003*	0,959	2,879
Vizcaya	0,832	-0,0001*	0,840	3,520
Zamora	0,660	0,0001*	0,719	1,944
Zaragoza	0,780	-0,0008*	1,073	3,221

*** No significativo al nivel 0.05**

Tabla 1 Valores de R_j^2, $A_{0,j}$, $A_{1,j}$, α y α_j, $j = 1,\ldots,50$.

Fuente: Elaboración propia a partir de datos del INE

3.2 HOMOGENEIZACIÓN DE LA ESCALA

Una vez verificadas las condiciones de aplicación de la metodología y precediendo a la construcción del conjunto de series resumen es necesario homogeneizar la escala de medida de las series $Y_{t,j}$ construyendo las correspondientes series $Z_{t,j}$ (**Fig. 3**, *ctro. izqda.*):

$$Z_{t,j} = Y_{t,j} - \alpha_j + \alpha \qquad j = 1,..., J$$

donde:

$$\alpha_j = \frac{1}{T} \cdot \sum_{t=1}^{T} Y_{t,j} \qquad j = 1,..., J$$

y α es la escala común elegida:

$$\frac{1}{T} \cdot \sum_{t=1}^{T} Z_{t,j} = \alpha \qquad \forall j$$

Si, por ejemplo:

$$\alpha = \frac{1}{T} \cdot \sum_{t=1}^{T} X_t$$

Entonces la media de las J series transformadas será igual a la media de la serie Y_t. La **Tabla 1** ofrece los valores de α y α_j para las cincuenta provincias.

3.3 EXTRACCIÓN DEL CONJUNTO DE CUVAS RESUMEN

Los pasos a seguir en el proceso de construcción del conjunto de series resumen son los siguientes:

Paso 1: Elegir K, número de series resumen.

Paso 2: Estimar por mínimos cuadrados ordinarios los coeficientes B_0 y B_1 de la ecuación de regresión: $\hat{A}_{0,j} = B_0 + B_1 \cdot A_{1,j}$.

Paso 3: Para $k = 1,..., K$, fijar el valor del coeficiente m_k y calcular $b_k = B_0 + B_1 \cdot m_k$. Por ejemplo, elegir dos valores distantes a y b dentro del rango de variación de los valores $A_{1,j}$, $j = 1,..., J$ y considerar: $m_1 = a$ y $m_k = m_{k-1} + \theta$, $k = 2,..., K$, con $\theta = (b - a)/(K - 1)$.

Paso 4: Para $k = 1,..., K$, calcular $g_{t,k} = b_k \cdot t + m_k \cdot Y_t$.

Paso 5: Para $k = 1,..., K$, calcular $\beta_k = \frac{1}{T} \cdot \sum_{t=1}^{T} g_{t,k}$.

Paso 6: Para $k = 1,..., K$, calcular $C_{t,k} = g_{t,k} - \beta_k + \alpha$.

Obsérvese que, si denominamos:

$$\mu_k = -\beta_k + \alpha$$

entonces:

$$C_{t,k} = b_k \cdot t + m_k \cdot Y_t + \mu_k \qquad \text{con} \qquad b_k = B_0 + B_1 \cdot m_k \qquad k = 1,..., K$$

Además:

$$\frac{1}{T}\cdot\sum_{t=1}^{T} C_{t,k} = \frac{1}{T}\cdot\sum_{t=1}^{T}(g_{t,k} - \beta_k + \alpha) = \frac{1}{T}\cdot T\cdot(\alpha - \beta_k) + \frac{1}{T}\cdot\sum_{t=1}^{T} g_{t,k} = (\alpha - \beta_k) + \beta_k = \alpha$$

En otras palabras, el conjunto de series resumen $\{C_{t,k}\}$, $k = 1,...,K$, construido según los pasos 1 a 6 verifica las condiciones [1], [2] y [3] de la **Proposición 1** respecto de Y_t**.**

Por ejemplo, en términos del rango de variación de $A_{1,j}$ (**Tabla 1**), fijamos $K = 6$[8] coeficientes de la forma:

$$m_1 = 0{,}3218 \quad \text{y} \quad m_k = m_{k-1} + \theta \quad \text{para} \quad k = 1,...,6$$

siendo:

$$\theta = (\max_j A_{1,j} - \min_j A_{1,j})/(K-1) = (1{,}7252 - 0{,}3218)/(6-1) = 0{,}2807$$

y, a partir de los valores de B_0 y B_1, calculamos los coeficientes b_k:

$$b_k = 0{,}00167 + (-0{,}00167)\cdot m_k \qquad k = 1,...,6$$

Los valores de los coeficientes m_k y b_k (**Fig. 3**, *sup. dcha.*) sirven para construir el conjunto de series $\{g_{t,k}\}$:

$$g_{t,k} = b_k \cdot t + m_k \cdot Y_t \qquad k = 1,...,6$$

y, a partir de sus medias:

$$\beta_k = \frac{1}{T}\cdot\sum_{t=1}^{T} g_{t,k} \qquad k = 1,...,6$$

el conjunto de series resumen $\{C_{t,k}\}$:

$$C_{t,k} = b_k \cdot t + m_k \cdot X_t + \mu_k \qquad k = 1,...,6$$

donde:

$$\mu_k = -\beta_k + \alpha$$

Así, los conjuntos:

$$Z_{t,j} \quad j = 1,...,50 \qquad \text{y} \qquad C_{t,k} \quad k = 1,...,6$$

se encuentran en la misma escala (**Fig. 3**, *ctro. izqda y dcha.*).

[8] La elección de K puede hacerse a modo de tanteo y, en función de la solución obtenida, corregir su valor si parece necesario y recalcular las series temporales resumen.

3.4 INTERPRETACIÓN DE LA SOLUCIÓN

Según su expresión las fluctuaciones de la sexta serie resumen son las más pronunciadas y las de la primera las más suaves. El orden responde a la relación con la serie Y_t: por la **Observación 1,** dado que el conjunto $\{C_{t,k}\}$ verifica las hipótesis de la **Proposición 1** respecto de Y_t, el conjunto de series $\{c^4_{t,k}\}$ (**Fig. 3**, *inf. izqda.*), donde:

$$c^4_{t,k} = C^4_{t,k} - C^4_{t-4,k} = b_k + m_k \cdot y^4_t \qquad k = 1,...,6 \qquad \text{con} \qquad y^4_t = Y_t - Y_{t-4}$$

define un haz de $K = 6$ rectas respecto de y^4_t de vértice:

$$(y^4_t, c^4_{t,k}) = (-4 \cdot B_1, 4 \cdot B_0) = (0.0067, 0.0067)$$

Así el orden de las series resumen $C_{t,k}$ viene dado por el orden de las series $c^4_{t,k}$ que, a su vez, viene dado por el grado de sensibilidad de éstas frente a y^4_t. Más concretamente, por el apartado **A)** de la **Proposición 1**, puesto que:

$$m_1 < m_2 < m_3 < m_4 < m_5 < m_6$$

entonces la secuencia $C_{t,1}, C_{t,2}, ..., C_{t,6}$ es tal que:

$$d(C_{t,k}, C_{t,k+1}) < d(C_{t,k}, C_{t,k'}) \qquad k = 1,...,4 \qquad k' = k+2,...,6$$

$$d(C_{t,k}, C_{t,k-1}) < d(C_{t,k}, C_{t,k'}) \qquad k = 3,...,6 \qquad k' = 1,..., k-2$$

Además, por la **Observación 7**, en aquellos instantes en que las trayectorias $c^4_{t,k}$ se cortan lo hacen a la altura del valor:

$$s \cdot B_0 = 4 \cdot 0.00167 = 0.0067$$

o, lo que es equivalente, en aquellos instantes en que las trayectorias $\exp c^4_{t,k}$ se cortan (**Fig. 3**, *inf. dcha.*) lo hacen a la altura del valor:

$$e^{0.0067} = 1.0067$$

que se puede interpretar como una estimación de la tendencia media del crecimiento interanual de las series del conjunto $\{C_{t,k}\}$.

Por otro lado, por el apartado **B)** de la **Proposición 1**, en aquellos puntos en los que dos de las trayectorias de series del conjunto $\{C_{t,k}\}$ se cortan también lo hacen las restantes. Así como la trayectoria entre dichos puntos de corte permiten resumir la trayectoria de la serie $Y_{t,j}$ [9], también la trayectoria entre los puntos de corte de las series

[9] Al fin y al cabo la serie promedio es una más del haz de rectas.

del conjunto $\{\exp C_{t,k}\}$ permiten resumir la de la serie $\exp Y_{t,j}$ (**Fig. 4**, *sup. izqda.* y *dcha.*): en el periodo transcurrido entre las dos fechas correspondientes a los dos primeros puntos de corte (entre finales de 1977 y finales de 1991), se produce un incremento del 10,1 por ciento; en el periodo de poco más de siete años transcurrido entre las dos fechas correspondientes al segundo y al tercer puntos de corte (entre finales de 1991 y finales de 1998), el incremento es del 3,9% o, lo que es equivalente, en el periodo de veintiún años transcurrido entre las fechas correspondientes al primer y al tercer puntos de corte (entre finales de 1977 y finales de 1998), es del 14,4 por ciento; finalmente, en el periodo de doce años comprendido entre finales de 1998 y finales de 2010, el incremento es del 7,5 por ciento o, lo que es equivalente, en el periodo de treinta y tres años transcurrido entre finales de 1977 y finales de 2010, del 24 por ciento. Obsérvese que dicho porcentaje coincide (salvo errores de redondeo) con el correspondiente a elevar el valor estimado de la tendencia media de crecimiento interanual al número de años del periodo de observación:

$$1.0067^{33} \cong 1{,}24$$

En definitiva esta relación entre estos cuatro puntos de corte es lo que la serie $\exp Y_t$ y las series $\exp C_{t,k}$, $k = 1,...,6$ presentan en común. Lo que las diferencia es la trayectoria seguida entre ellos: mientras que entre cualquier par de puntos de corte la serie $\exp C_{t,1}$ se desvía muy poco de lo que correspondería a un crecimiento interanual constante e igual a la estimación de la tendencia media, la serie $\exp C_{t,6}$ se desvía mucho; la trayectoria de la serie $\exp Y_t$ se encuentra en una posición intermedia.

El análisis de la trayectoria de una provincia concreta se hará en estos términos; por ejemplo, la trayectoria de la serie $\exp Z_{t,j}$ correspondiente a Madrid (**Fig. 4**, *inf. izqda.*) se corta con la de las distintas series del conjunto $\{\exp C_{t,k}\}$, si no exactamente en los cuatro puntos de corte comunes, en posiciones muy próximas, por lo que podemos afirmar que los incrementos entre los cuatro valores correspondientes resumen con bastante precisión la tendencia de la ocupación en el sector de la construcción en Madrid a lo largo del periodo de observación. Por otro lado, en las tres etapas delimitadas por los cuatro puntos de corte su trayectoria es de fuerte sensibilidad frente a la serie promedio, aunque en la tercera de forma menos acusada.

En otras palabras, bajo el supuesto de que existe un factor responsable de las fluctuaciones del conjunto de las series de ocupación en el sector de la construcción en las distintas provincias, en el sentido de que la menor o mayor volatilidad de las fluctuaciones depende del menor o mayor grado de sensibilidad frente a variaciones de dicho factor, podemos afirmar que la provincia de Madrid se asocia con un alto grado de sensibilidad.

Alternativamente, para describir la evolución del número de ocupados en el sector de la construcción podemos utilizar la representación de la serie de incrementos interanuales (**Fig. 4**, *inf. dcha.*):

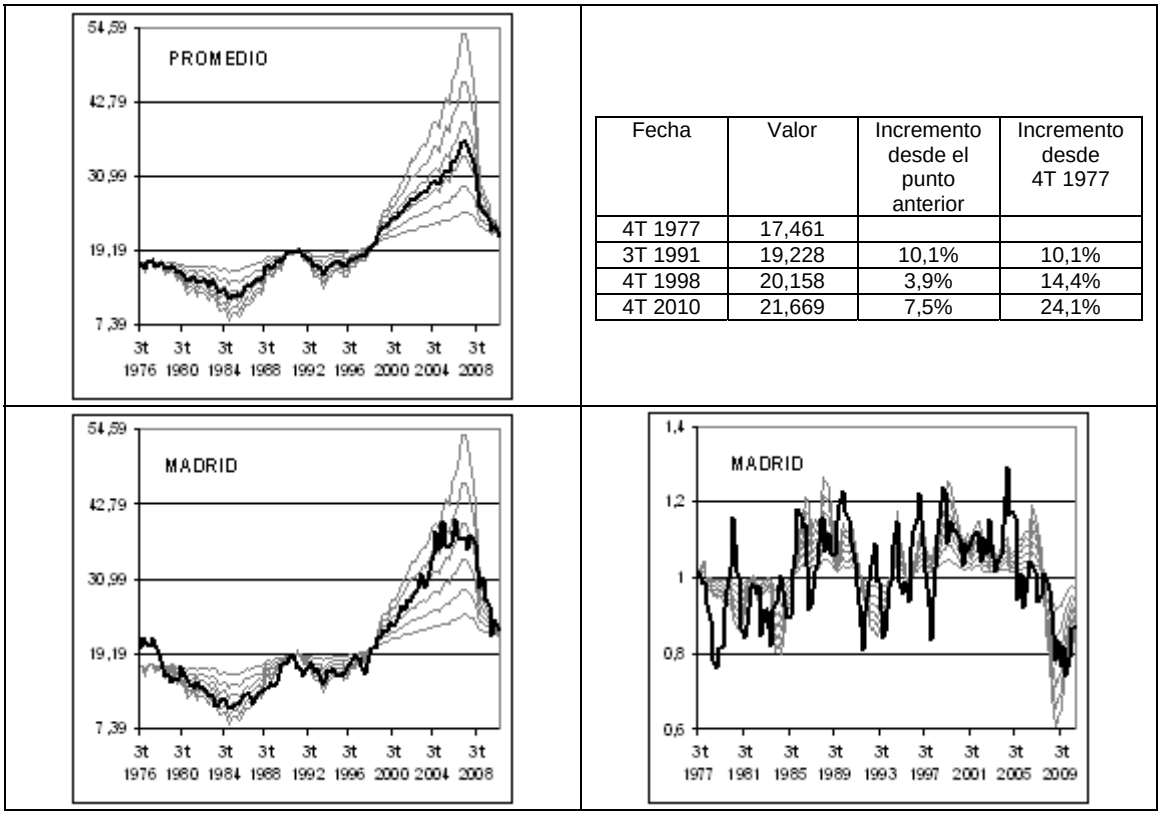
$$o_{t,j}^{4} = \frac{O_{t,j}}{O_{t-4,j}} \qquad t = 1,...,T$$

sobre el conjunto $\{\exp c_{t,k}^{4}\}$ dado que, por definición de $Z_{t,j}$:

$$\ln\left(\frac{O_{t,j}}{O_{t-4,j}}\right) = Z_{t,j} - Z_{t-4,j}$$

Sin embargo, la interpretación sería claramente más compleja, razón por la que la metodología se aplica directamente sobre las series observadas en lugar de sobre las correspondientes series de incrementos.

Aunque la representación de cada serie $\exp Z_{t,j}$ sobre la solución de series resumen $\{\exp C_{t,k}\}$ simplifica la descripción de su trayectoria en comparación con las restantes, para establecer las similitudes y diferencias entre todas ellas sería necesario comparar las J representaciones. Veamos en lo que sigue cómo aplicar el fundamento teórico que subyace en el proceso de construcción del conjunto $\{C_{t,k}\}$ para simplificar esta comparación.



Fecha	Valor	Incremento desde el punto anterior	Incremento desde 4T 1977
4T 1977	17,461		
3T 1991	19,228	10,1%	10,1%
4T 1998	20,158	3,9%	14,4%
4T 2010	21,669	7,5%	24,1%

Fig. 4 *Sup. Izqda.*: Serie $\exp Y_t$ representada sobre el conjunto $\{\exp C_{t,k}\}$, k = 1,…,6 ; *Sup. Dcha.*: Puntos de cruce de las trayectorias de $\exp C_{t,k}$; *Inf. Izqda.*: Serie $\exp Z_{t,j}$ para Madrid representada sobre el conjunto $\{\exp C_{t,k}\}$, k = 1,…,6 ; *Inf. Dcha*: Serie de incrementos interanuales de ocupados en el sector de la construcción en Madrid representada sobre el conjunto $\{\exp c^4_{t,k}\}$.

Fuente: Elaboración propia a partir de datos del INE

3.5 COMPARACIÓN DE LAS TRAYECTORIAS

Dado el conjunto de series $\{C_{t,k}\}$, consideremos la distancia entre cada par de ellas (Fig. 5, *sup. izqda.*):

$$d_{k,k'} = d^2(C_{t,k}, C_{t,k'}) \qquad k, k' = 1,..., K$$

siendo d la distancia euclídea. Por el apartado **A)** de la **Proposición 1**, la serie más próxima a la primera (línea más clara) es la segunda, seguida de la tercera y así sucesivamente, lo que implica una secuencia de distancias creciente; por otro lado, la más próxima a la sexta (línea de mayor grosor) es la quinta, seguida de la cuarta y así sucesivamente, lo que implica una secuencia de distancias decreciente. En lo que se refiere a cualquier otra serie, la secuencia será decreciente hasta cero (distancia consigo misma) y creciente hasta la sexta. En otras palabras, la secuencia de distancias de la primera serie temporal a cada una de las restantes está positivamente correlada con la secuencia de distancias de la segunda que, a su vez, está positivamente correlada con la secuencia de distancias de la tercera, y así sucesivamente hasta la secuencia de distancias de la quinta que está positivamente correlada con la secuencia de distancias de la sexta. Además, esta secuencia de seis correlaciones parte de un valor alto y positivo que se va debilitando terminando en un valor alto en términos absolutos y negativo. Como consecuencia de esta relación entre las distancias, si aplicamos un Análisis de Componentes Principales sobre la correspondiente matriz de distancias y representamos la solución en el espacio de los dos primeros componentes (**Fig. 5**, *sup. dcha.*) observamos el denominado efecto Guttman[10].

Así, al aplicar un Análisis de Componentes Principales sobre la matriz de distancias entre cada par de series del conjunto $\{Z_{t,j}\}$:

$$d_{j,j'} = d^2(C_{t,j}, C_{t,j'}) \qquad j, j' = 1,..., J$$

en la representación de la solución sobre los dos primeros componentes (**Fig. 5**, *inf.*) puede observarse que si trazáramos una línea recorriendo la nube de puntos desde Almería hasta Orense obtendríamos una curva muy aproximada a un arco de parábola. La correcta interpretación de la posición de los puntos en el espacio factorial dependerá de su calidad de representación. Bajo el supuesto de que dicha calidad es alta, que el ángulo que forman desde el origen dos puntos-provincia sea muy pequeño implica que las dos correspondientes columnas de distancias están muy correladas positivamente y, en consecuencia, que las dos provincias son parecidas entre sí; que el ángulo sea próximo a los 180 grados, implica que las dos correspondientes columnas de distancias están muy correladas negativamente y, en consecuencia, que las dos provincias son distintas entre sí; y finalmente, que el ángulo sea próximo a los 90 grados, que las dos correspondientes columnas de distancias están muy incorreladas y, en consecuencia, que las dos provincias no son ni muy parecidas ni muy distintas.

[10] El efecto Guttman se obtiene cuando, al representar las filas o las columnas de una matriz en el espacio de las dos primeras componentes de la solución factorial, la nube de puntos correspondiente tiene forma de arco de parábola.

En términos globales los distintos puntos-provincia están bien representados[11], por lo que la interpretación de su posición es bastante fiable y, en consecuencia, también lo es la ordenación de las provincias que ofrece la línea imaginaria que recorre la nube de puntos desde Almería hasta Orense.

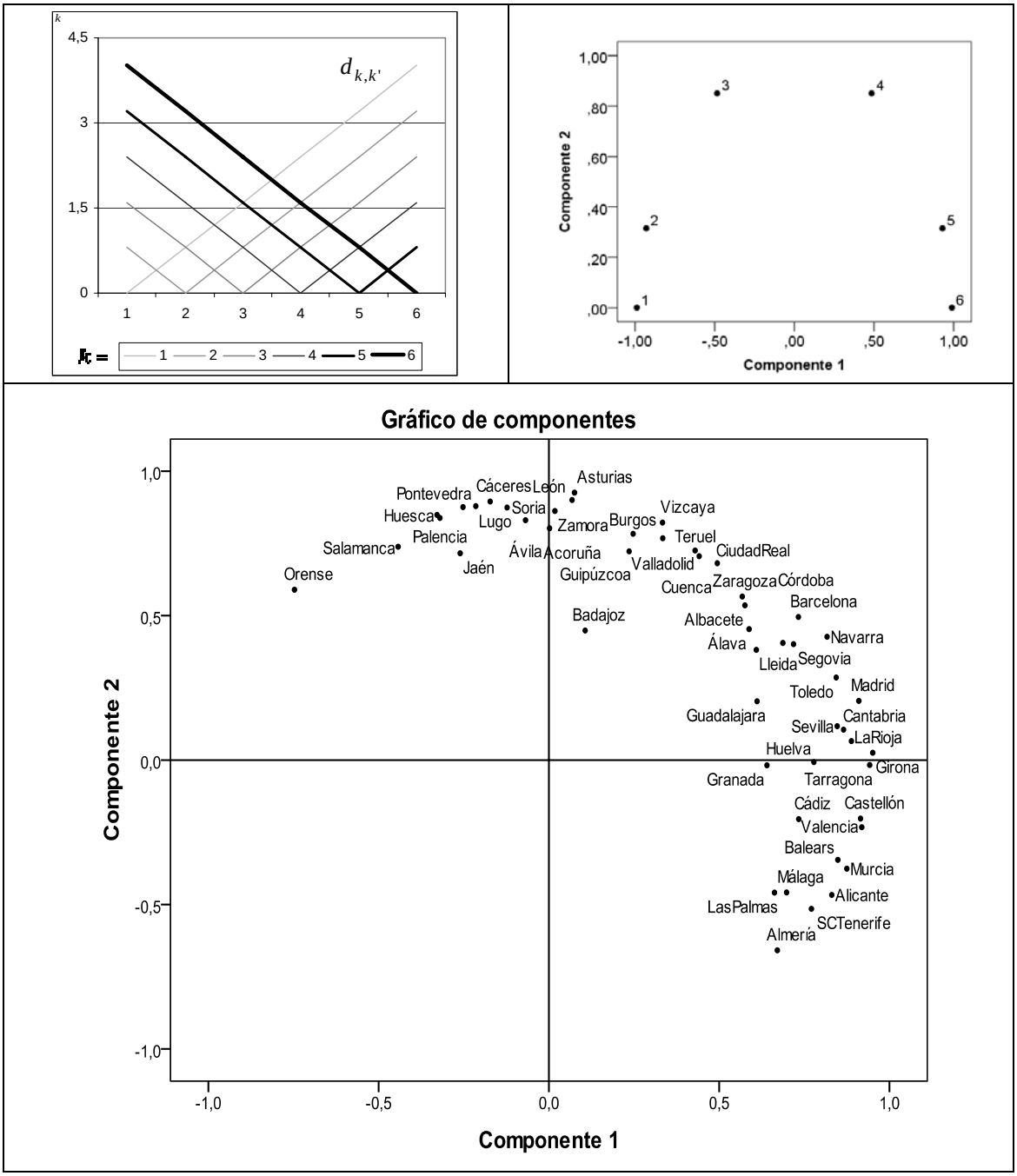


Fig. 5 *Sup. Izqda.*: Distancias entre las series $C_{t,k}$, k = 1,…,6 ; *Sup. Dcha.*: Representación factorial de las distancias; *Inf.*: Representación factorial de las distancias entre las series $Z_{t,j}$, j = 1,…,50 .

Fuente: Elaboración propia a partir de datos del INE

[11] La calidad de representación de un punto viene dada por su distancia al origen, que a lo sumo puede tomar el valor 1.

La matriz de distancias está calculada en términos del conjunto $\{Z_{t,j}\}$, por lo que, a efectos de comparar provincias, la representación gráfica de sus trayectorias debería hacerse en esta escala que, en definitiva, es la del conjunto $\{C_{t,k}\}$. Alternativamente, si consideramos las series del conjunto $\{Y_{t,j}\}$, para interpretar la trayectoria de cada una de ellas en comparación con la de las restantes, podemos expresar el conjunto de series resumen en su misma escala:

$$C^j_{t,k} = C_{t,k} - \alpha + \alpha_j \qquad k = 1,...,6$$

Por el apartado **C)** de la **Proposición 1** la diferencia entre los puntos de corte no depende de la escala en que representemos el conjunto $C_{t,k}$, luego si t y t' son dos instantes correspondientes a dos puntos de corte entonces:

$$C^j_{t',k} - C^j_{t,k} = C_{t',k} - C_{t,k} = C_{t',k'} - C_{t,k'} = C^j_{t',k'} - C^j_{t,k'}$$

Además, al no depender de j, esta diferencia también es la misma en todas las provincias.

Aunque la representación del número de ocupados en el sector de la construcción en cada provincia, $O_{t,j} = \exp Y_{t,j}$, (**Fig. 6**) se realizará con referencia al conjunto $\{\exp C^j_{t,k}\}$ (eje izquierdo), para comparar las distintas provincias utilizaremos como referencia el conjunto $\{\exp C_{t,k}\}$ (eje derecho). La secuencia de gráficos de la **Figura 6** responde al trazado de la línea imaginaria que va desde Almería hasta Orense en la **Figura 5**, así las primeras provincias representadas son tales que su trayectoria va pareja a la serie resumen de mayor sensibilidad frente a la serie promedio (la sexta), como por ejemplo Málaga y Las Palmas (**Fig. 7**, *sup*. *izqda*.), y la de las últimas, pareja a la serie resumen de menor sensibilidad (la primera), como por ejemplo Palencia y Huesca (**Fig. 7**, *inf*. *dcha*.); el resto de provincias, tales como Navarra y Segovia (**Fig. 7**, *sup*. *dcha*.) o como Vizcaya y Valladolid (**Fig. 7**, *inf*. *izqda*.) se encuentra entre ambos extremos.

4. CONCLUSIONES

A partir de la información en un conjunto grande de series temporales, y bajo el supuesto de que la estructura que subyace en él es la de un haz de rectas, hemos construido un conjunto pequeño de series resumen. La representación sobre este conjunto nos ha permitido interpretar gráficamente la trayectoria de cada serie temporal en comparación con la de las restantes. En definitiva hemos comprobado que, a la hora de construir un modelo estadístico o económico que persiga objetivos de tipo explicativo o predictivo para un conjunto grande de series temporales, la aplicación de la metodología del haz de rectas puede ser una herramienta gráfica de gran utilidad como parte del estudio exploratorio previo.

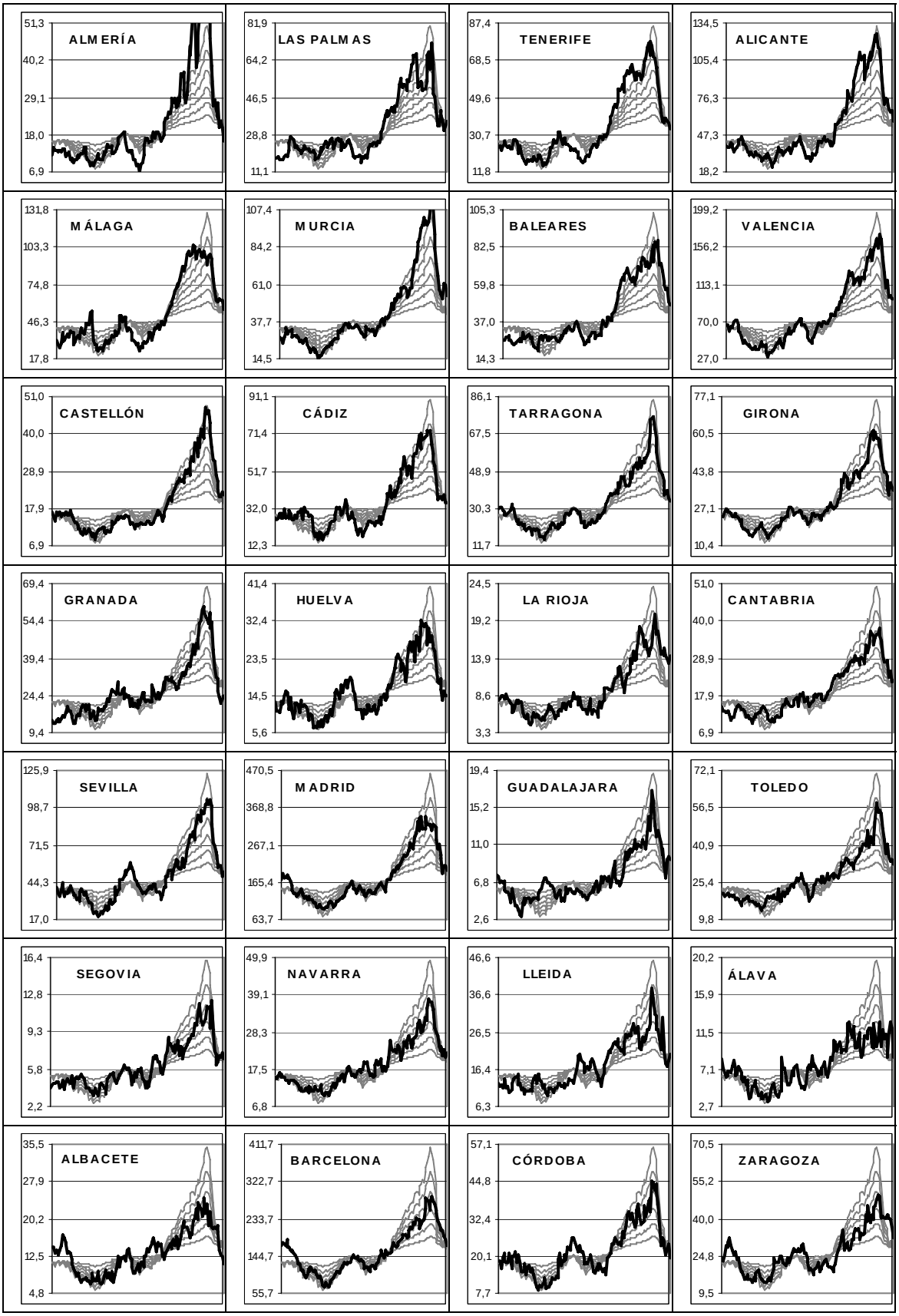


Fig. 6 Trayectoria del número de ocupados en el sector de la construcción (en miles) en cada una de las cincuenta provincias españolas representada sobre el conjunto $\{\exp C_{t,k}^{j}\}$.

Fuente: Elaboración propia a partir de datos del INE

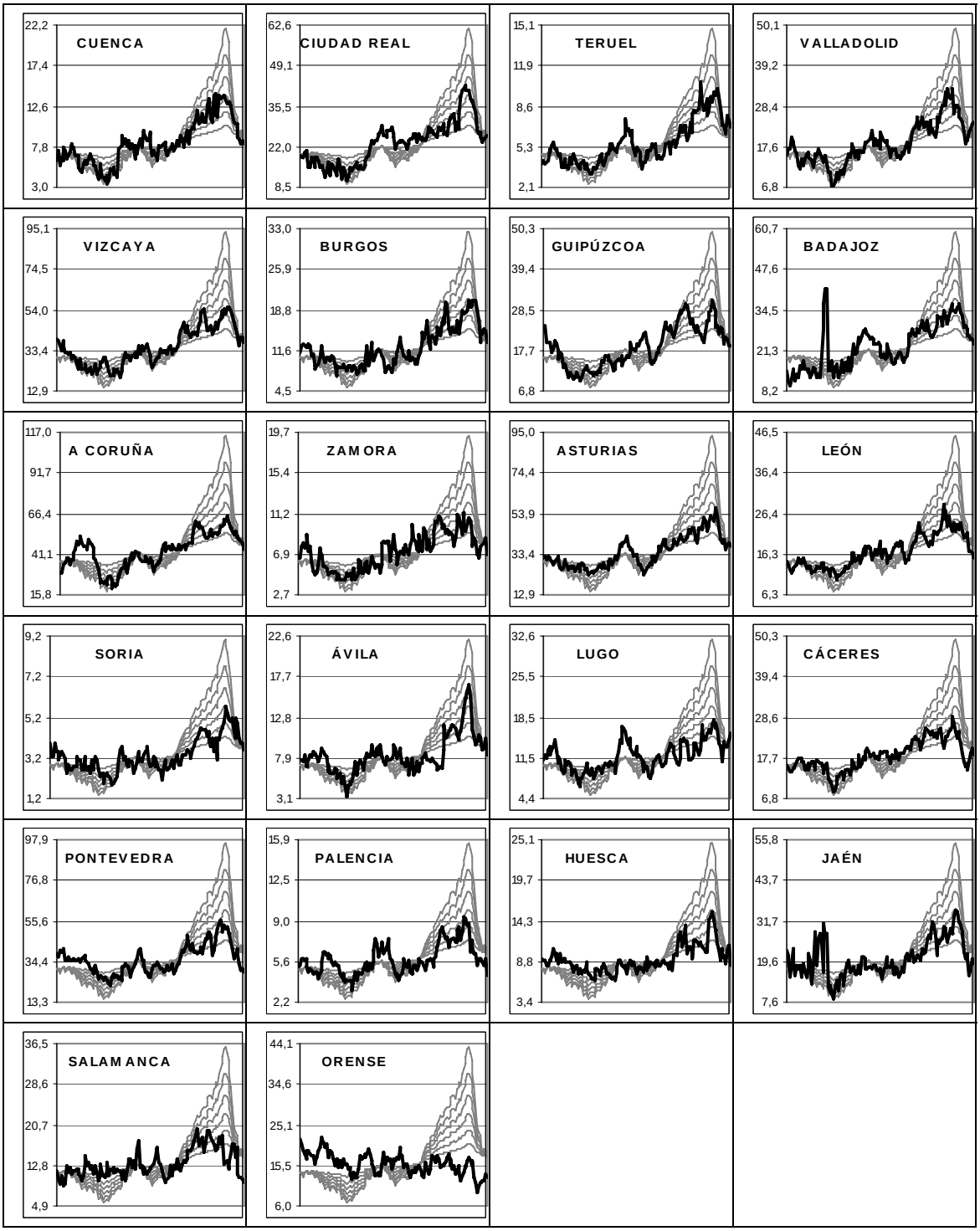


Fig. 6 (Cont) Trayectoria del número de ocupados en el sector de la construcción (en miles) en cada una de las cincuenta provincias españolas representada sobre el conjunto $\{\exp C_{t,k}^{j}\}$.

Fuente: Elaboración propia a partir de datos del INE

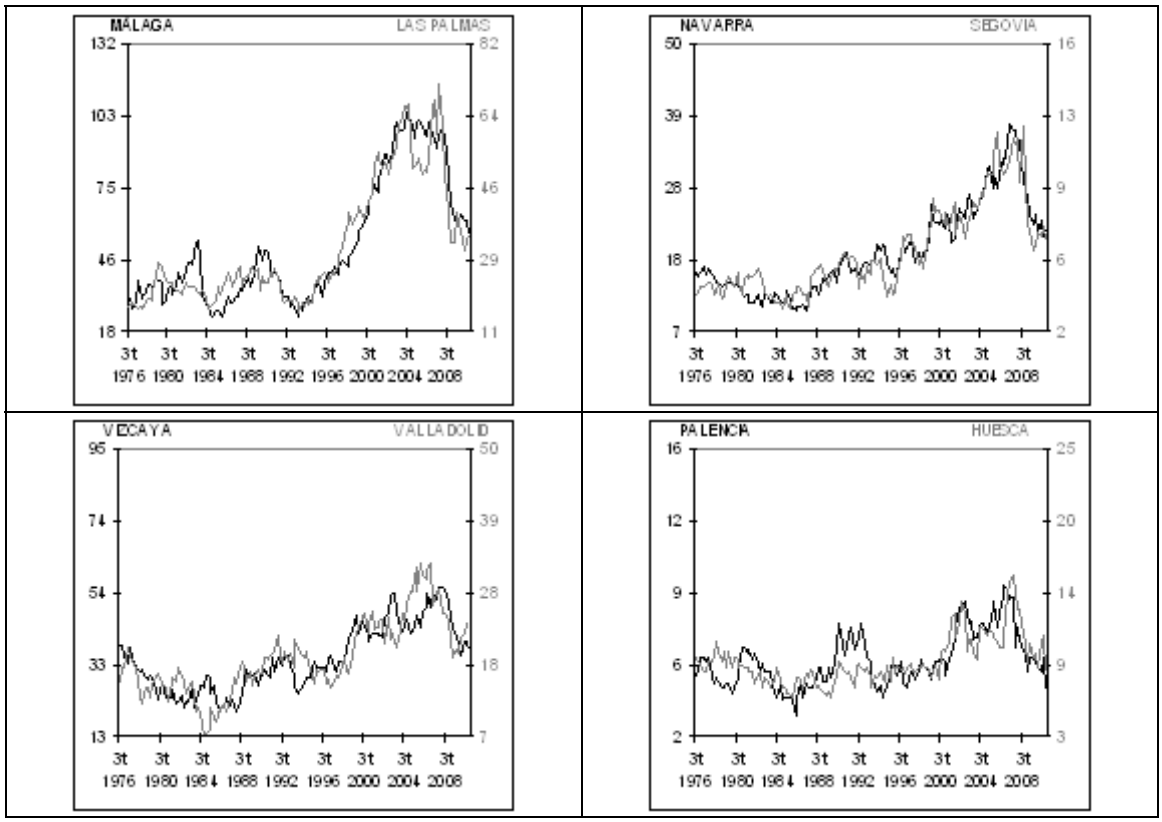


Fig. 7 Ocupados en el sector de la construcción residencial (en miles) en Málaga y Las Palmas (*Sup. Izqda*), Navarra y Segovia (*Sup. Dcha.*),Vizcaya y Valladolid (*Inf. Izqda.*) y Palencia y Huesca (*Inf. Dcha.*).

Fuente: Elaboración propia a partir de datos del INE