Estadística Española • Volumen 55, número 181 / 2013, pp. 123-148

# El haz de rectas para la comparación gráfica de series temporales geográficas(*)

**Magdalena Ferrán Aranaz**
Departamento de Estadística e I.O. III
Escuela Universitaria de Estadística (U.C.M.)

**Resumen**

**A** pesar de la importancia que tienen las herramientas de representación gráfica en el análisis de datos es sorprendente la escasez de trabajos relativos a la representación de conjuntos masivos de series temporales y, en particular, de series temporales geográficas. En este trabajo se propone una metodología, a la que denominaremos Metodología del haz de rectas, para la comparación gráfica de series temporales que midan un mismo fenómeno o variable sobre distintas regiones geográficas.

*Palabras clave:* Series temporales geográficas; gráficos estadísticos; visualización; análisis exploratorio de datos.

*Clasificación AMS:* 62-09

## The sheaf of straight lines for graphical comparison of geographic time series

**Abstract**

**S**tatistical graphics are essential tools in data analysis but, surprisingly, there has been relatively little work on massive time series data set, particularly in geographic time series. This paper offers a methodology, which we will denominate Sheaf Methodology, for graphical comparison of time series that measure the same phenomenon or variable in different regions.

*Key words*: Geographic time series; statistical graphics; visualization; exploratory data analysis.

*AMS Classification*: 62-09

(*) Agradezco desde aquí a los evaluadores expertos anónimos los comentarios y sugerencias formulados para mejorar este texto.

## 1. Introducción

Una serie temporal es una secuencia de datos medidos en intervalos sucesivos de tiempo y espaciados uniformemente. La recogida de datos estadísticos con estructura de serie temporal es una actividad muy frecuente en la actualidad, no sólo por parte de organismos oficiales como el Instituto Nacional de Estadística, el Banco de España o el Ministerio de Economía y Hacienda, sino también por parte de las grandes o medianas empresas del país. Gran parte de estos datos están asociados con regiones geográficas; una serie temporal geográfica consiste en la recopilación de un conjunto de series temporales regionales en un vector de series temporales.

Una fase esencial de la modelización estadística es el análisis exploratorio de los datos, que implica un estudio y depuración de los datos como paso previo a la decisión del modelo adecuado para el ajuste; en el caso particular de un vector de series regionales el objetivo que se persigue al analizarlas conjuntamente es el de, llegado el momento, conseguir una mayor precisión de las predicciones para cada serie en particular mediante la consideración de la información adicional que ofrecen las relaciones existentes entre todas ellas.

Al interpretar la información que ofrecen los datos el cerebro requiere de más tiempo si éstos vienen presentados en una tabla de números que si vienen representados en un gráfico. Esta es la razón por la que los gráficos son la herramienta más importante para examinar datos longitudinales. La forma estándar de representar una serie temporal es a través de un gráfico de líneas, que consiste en un diagrama cartesiano en el que la abscisa representa el tiempo y la ordenada los valores de la serie, y los puntos consecutivos se unen mediante segmentos formando una línea. La interpretación del gráfico ayudará en la detección de patrones de comportamiento en la trayectoria temporal de los datos. El problema surge cuando se trata de comparar múltiples series temporales, como puede ser el caso de una serie temporal geográfica. Tanto si se representa cada serie regional por separado como si se representan todas ellas conjuntamente mediante su superposición en un único gráfico (bien en la escala original de los datos o bien con los datos transformados a una escala común) la comparación se hace muy complicada en cuanto el número de series es mínimamente elevado. En este trabajo se presenta una metodología para simplificar la comparación gráfica de series temporales geográficas.

La representación de series temporales se remonta a finales del siglo dieciocho y se considera que el pionero, no sólo de los gráficos de líneas, sino también de los gráficos de barras y de sectores fue William Playfair (Tufte, 2001). Desde entonces los gráficos estadísticos han desempeñado un papel fundamental en el análisis exploratorio de los datos, en la formulación de hipótesis o en el desarrollo de modelos, pero la mayor parte de las propuestas se ha centrado en la mejora de los gráficos convencionales; en el caso particular de la representación de series temporales, en la mejora de los gráficos de líneas o de barras (Zhao et al., 2011). En la actualidad las modernas tecnologías informáticas ofrecen magníficas herramientas para el diseño de nuevas técnicas y el desarrollo de software para la visualización de datos espacio-temporales (para una revisión de las mismas véase Andrienko et al., 2003), a pesar de ello sorprende la

escasez de trabajos relativos a la representación de conjuntos masivos de series temporales (Lin et al., 2005). Los tres trabajos más referenciados en este campo posiblemente sean los de Hochheiser and Shneiderman (2001), van Wijk y van Selow (1999) y Weber et al. (2001). La duda que surge es la de si en estas tres propuestas y, en términos generales, en las del campo de la visualización, la estructura del gráfico responde más a un objetivo estético que funcional. Para abrir un debate al respecto, el volumen 22 de la revista Statistical Computing & Graphics Newsletter ofrece dos artículos relativos a la presentación gráfica de datos cuantitativos. En el primero de ellos (Kosara, 2011) se ofrece el punto de vista del experto en informática que construye herramientas y técnicas para la visualización de datos que, además de ofrecer una atractiva presentación, tengan una gran capacidad interactiva; mientras que en el segundo (Gelman y Unwin, 2011), se ofrece el punto de vista del experto en estadística, que utiliza los gráficos como herramienta de ayuda en la toma de decisiones en cualquier punto del proceso de análisis estadístico de los datos.

La herramienta que aquí se propone, denominada "Metodología del haz de rectas", está dirigida al usuario de la estadística, más concretamente al analista de series temporales geográficas. Se trata de una metodología de descripción gráfica que ni se adentra en el terreno de la estadística espacio-temporal (no contempla la posible dependencia espacial entre las series), ni aborda problemas de modelización. Bajo el supuesto de que la estructura que subyace en el conjunto de series regionales objeto de análisis es la de un haz de rectas[1] se procederá, por un lado, a la extracción de un "conjunto de series resumen" y, por otro, a la ordenación de las series regionales en términos de sus similitudes. Cada serie regional se representará por separado, pero la secuencia de gráficos responderá a la ordenación establecida; además, en cada gráfico, junto con la línea regional, se representarán las líneas relativas a las series resumen. Estos dos aspectos facilitarán la interpretación de las similitudes y diferencias entre las distintas regiones.

El trabajo se estructura de la siguiente forma: tras la presentación de la metodología mediante el desarrollo de un ejemplo sencillo, se exponen los fundamentos teóricos para, finalmente, ilustrar el proceso de aplicación mediante la comparación de las cincuenta series temporales relativas al número de ocupados en el sector de la construcción en cada una de las provincias españolas.

## 2. Presentación de la metodología

Antes de proceder a la exposición de un ejemplo introductorio, formalicemos dos conceptos clave en el desarrollo de la Metodología del haz de rectas.

Definición 1: *Un conjunto* $\{c_{t,k}\}$ *de* K *series temporales distintas, todas ellas definidas en los mismos instantes,* t=1,…,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:*

[1] Este concepto será formalizado en el siguiente apartado.

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

Obsérvese que:

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

En consecuencia[2]:

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

y, en dicho caso, $c_{t,k} = B_0$. En otras palabras, las K rectas $c_{t,k} = b_k + m_k . x_t$ forman un haz de rectas secantes de vértice $(x_t,\ c_{t,k}) = (-B_1,\ B_0)$.

Definición 2*: Sea* $\{Y_{t,j}\}$, j = 1 ,..., J, *un conjunto de* J *series temporales distintas, todas ellas definidas en los mismos instantes,* t = 1,...,T. *Diremos que* <u>*la estructura que subyace en el conjunto*</u> $\{Y_{t,j}\}$ <u>*es la de un haz de rectas*</u> *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:*

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

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

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

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

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

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

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 $Y_{t,1}$, $Y_{t,2}$, $Y_{t,3}$ *e* $Y_{t,4}$ cuatro series temporales (Fig. 1, A y B) e $Y_t$ su correspondiente serie promedio. Supongamos, por un lado, que para cada una de las cuatro series, el coeficiente $R_j^2$ relativo a la ecuación de regresión:

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

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

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

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

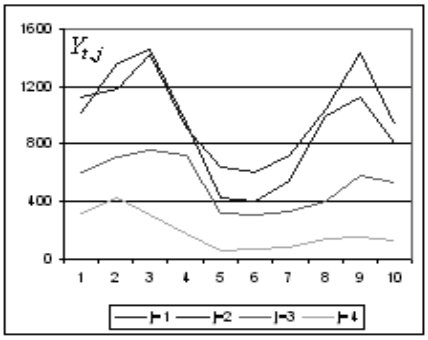
Figura 1

**Ejemplo introductorio**

**A**

| $Y_{t,j}$ | $j = 1$ | $j = 2$ | $j = 3$ | $j = 4$ | $Y_t$ |
|---|---|---|---|---|---|
| $t$ =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 |
| $\bar{Y}_j$ | **904,8** | **999** | **523,6** | **182,2** | $\bar{Y}$ =**652,4** |

**B**



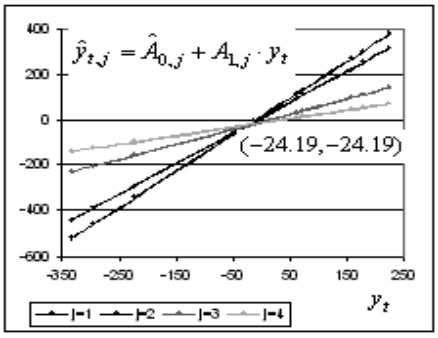

**C**

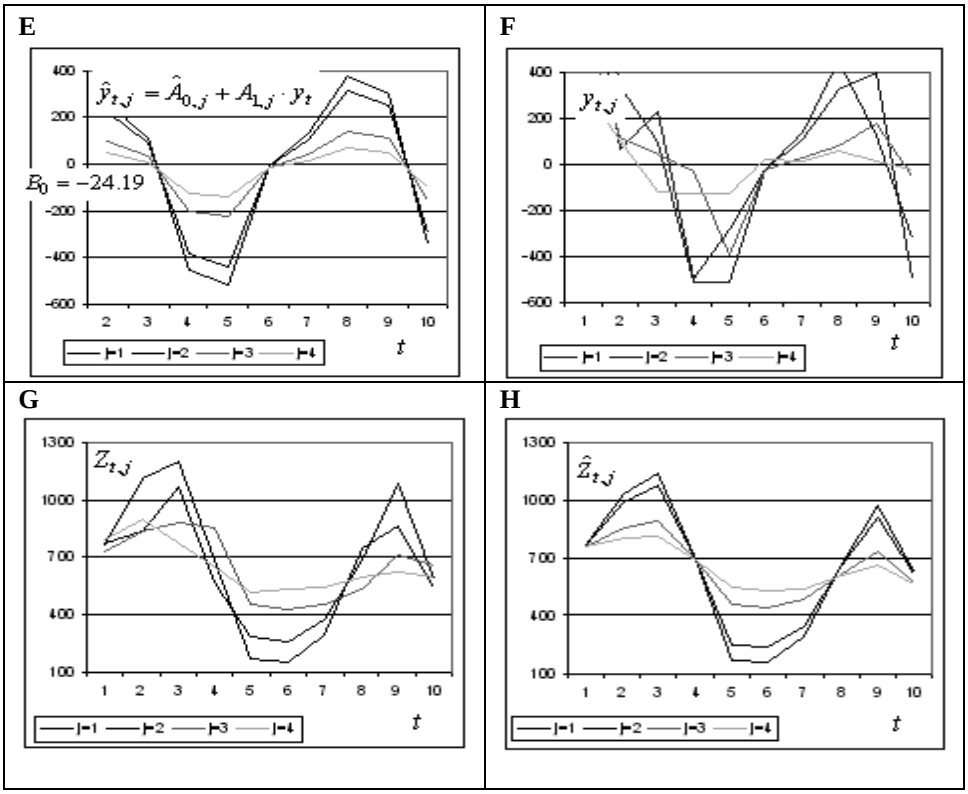
| $j$ | $R_j^2$ | $A_{0,j}$ | $A_{1,j}$ | $A_{2,j}$ | $\hat{A}_{0,j}$ | $\hat{A}_{2,j}$ |
|---|---|---|---|---|---|---|
| 1 | 0,983 | 2,51 | 1,61 | -157,0 | 14,67 | -223,9 |
| 2 | 0,934 | 23,80 | 1,35 | -13,9 | 8,51 | 70,2 |
| 3 | 0,843 | -7,98 | 0,66 | 135,3 | -8,16 | 136,3 |
| 4 | 0,894 | -18,32 |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






Fuente: Elaboración propia

Supongamos también que la asociación lineal entre las dos secuencias de coeficientes $(A_{0,1}, \ldots, A_{0,4})$ y $(A_{1,1}, \ldots, A_{1,4})$ 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):

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

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

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

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 para el que la media de la serie $\hat{Y}_{t,j}$ es igual a la de $Y_{t,j}$ $\left(\bar{\hat{Y}}_t = \bar{Y}_t\right)$, se puede contemplar como una estimación del valor $Y_{t,j}$; por ejemplo (Fig. 1, C):

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

Si homogeneizamos la escala de las cuatro series, por ejemplo, al nivel de la serie promedio, $\bar{Y}$, (Fig. 1, G):

$$Z_{t,j} = Y_{t,j} - \bar{Y}_j + \bar{Y} \qquad \text{j} = 1, \ldots, 4$$

también el valor $\hat{Z}_{t,j} = \hat{Y}_{t,j} - \bar{Y}_j + \bar{Y}$ se puede contemplar como estimación del valor $Z_{t,j}$ (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 una conjunto de series temporales $\{Y_{t,j}\}$ (Fig. 2, *sup. izqda.*) que miden un mismo fenómeno o variable en distintas regiones geográficas, como paso previo al proceso de construcción de las series resumen comprobaremos si la metodología es aplicable. La comparación regional 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 $(A_{1,\text{j}}, \hat{A}_{0,j})$ situado en un segmento de la recta $y = B_0 + B_1.x$ (Fig. 2, *centro dcha.*), las series resumen se construirán a partir de K puntos representativos de este mismo segmento (Fig. 2, *inf. izqda.)*:

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

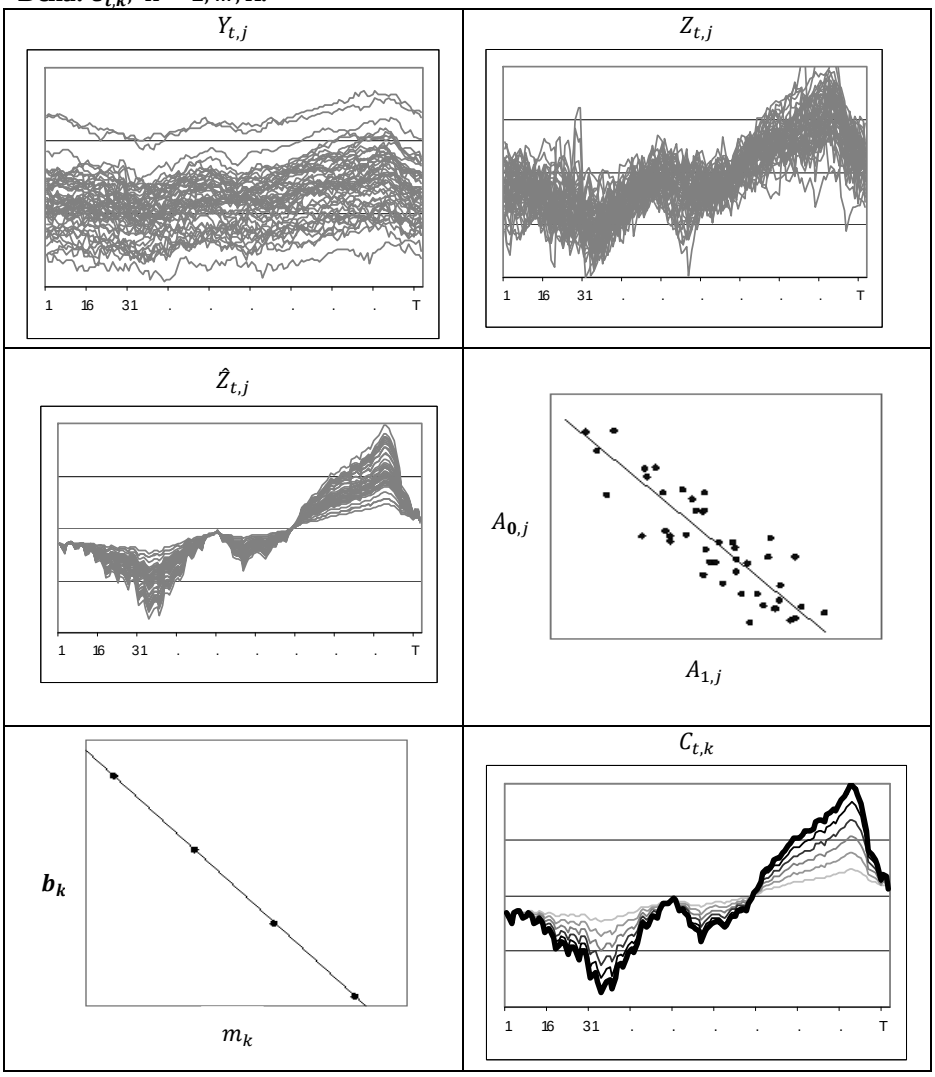
En definitiva, la expresión de las series resumen (Fig. 2, *inf. dcha.)* vendrá dada por:

$$C_{t,k} = b_k.t + m_k.Y_k + \mu_k \qquad \text{k} = 1, \ldots, \text{k};$$

siendo $\mu_k$ aquel valor que sitúe la k -ésima serie resumen en la misma escala que las J series homogeneizadas.

Figura 2

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




Fuente: Elaboración propia

## 3. Fundamentos teóricos

Expongamos a continuación una serie de resultados teóricos relacionados con la construcción del conjunto de series resumen.

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

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

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

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

siendo:

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

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

A) Si $C_{t,q}, C_{t,y}, y\ C_{t,s}$, son tres series temporales cualesquiera del conjunto $\{C_{t,k}\}$ tales que $m_q < m_r < m_s$ entonces $d(C_{t,q}, C_{t,y}) < d(C_{t,q}, C_{t,s})$, donde $d$ es la distancia euclídea.

B) Para cualquier par de series temporales del conjunto $\{C_{t,k}\}$ existe al menos un punto en su trayectoria[3] 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): Las expresiones [2] y [3] implican que:

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

$$= (B_0 + B_1 \cdot m_k) \cdot \overline{t} + m_k \cdot \overline{X} + \mu_k = B_0 \cdot \overline{t} + (B_1 \cdot \overline{t} + \overline{X}) \cdot m_k + \mu_k$$

entonces, por [1]: $\alpha = B_0 \cdot \overline{t} + (B_1 \cdot \overline{t} + \overline{X}) \cdot m_k + \mu_k \qquad \forall k$

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

Luego: $$\mu_k = D_0 + D_1 \cdot m_k \qquad \forall k \qquad [4]$$

donde $D_0 = \alpha - B_0.\bar{t}$ y $D_1 = -B_1.\bar{t} - \bar{X}$. En consecuencia, [2] también puede expresarse como:

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

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

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

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

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; es decir, si suponemos[4] $m_1 < m_2 < ... < m_K$ 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 las trayectorias de $C_{t,k}$ y $C_{t,k'}$ no tuvieran ningún punto de corte una de ellas tomaría siempre valores mayores que la otra, $C_{t,k} > C_{t,k'}$ $\forall t$, y en dicho caso:

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



[4] Ver nota a pie de página 1.

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[5], 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[6] $m_k \neq m_{k'} \; \forall k' \neq k$ entonces en el instante *t* se verifica $B_1.t + X_t + D = 0$ y, en consecuencia:

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

Luego, por [5], en el instante *t* se verifica $C_{t,k} > C_{t,k'} \; \forall k' \neq k$. 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 $B_1.t + X_t + D = 0$ y, por [5], $C_{t,k} = D_0 + B_1.t, \forall k$. Luego, si $t'$ es otro instante de corte, la diferencia $C_{t,k} - C_{t',k} = B_0.(t - t')$ es independiente de la media de las series temporales.

*Observación 1:* Si denominamos $c_{t,k}^s = C_{t,k} - C_{t-s,k}$ y $x_t^s = X_t - X_{t-s}$, por [2]:

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

donde, por [3], $s.b_k = s.B_0 + s.B_1.m_k \; \forall k$. En consecuencia, respecto de $x_t^s$, el conjunto $\{x_{t,k}^s\}$ tiene estructura de haz de K rectas de vértice $(x_t^s, c_{t,k}^s) = (-s.B_1,\ s.B_0)$.

*Observación 2:* Si $C_t$ es la serie promedio, la hipótesis [2] implica:

$$C_t = \frac{1}{K} \cdot \sum_{k=1}^{K} C_{t,k} = \frac{1}{K} \sum_{k=1}^{K} (b_k \cdot t + m_k \cdot X_t + \mu_k) = \bar{b} \cdot t + \bar{m} \cdot X_t + \bar{\mu}$$

Si denominamos $b = \bar{b}, m = \bar{m}, y \; \mu = \bar{\mu}$ entonces:

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

Luego la serie $c_t^s = C_t - C_{t-s}$ es tal que existen dos coeficientes *b* y *m* tales que:

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

[5] 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.

[6] Ver nota a pie de página 1.

donde, por [3]: $$b = B_0 + B_1.m \quad [9]$$

*Observación 3*: Por [6] y [8]:

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

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

Entonces, por [3] y [9]: $$s \cdot b_k^c = s \cdot B_0 - s \cdot B_0 \cdot m_k^c \quad [12]$$

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

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

En tal caso, por [10] y [12], el vértice del haz es $(c_t^s, c_{t,k}^s) = (s.B_0,\ s.B_1)$.

*Observación 4:* Por la Observación 3, 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 $s.B_0$; además, en dichos instantes y a dicha altura, se cortan las restantes trayectorias así como la de $c_t^s$.

*Observación 5:* 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, aunque en la práctica, para construir el conjunto de series resumen, no será necesario conocerla; en su lugar se considerará la serie promedio $Y_t$.

## 4. La metodología del haz de rectas: un estudio de caso

### 4.1 Condiciones de aplicación

Sea $\{O_{t,j},\ t = 1, \ldots T = 138\}$, $J = 1, \ldots, 50$, 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[7] y sea $Y_t$ la correspondiente serie promedio:

[7] 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 4.5).

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

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[8]. Para comprobarlo calcularemos los coeficientes $A_{0,j}$, $A_{1,j}$ y $A_{2,j}$ mediante el ajuste de la ecuación de regresión lineal:

$$\hat{Y}_{t,j} = A_{0,j} \cdot t + A_{1,j} \cdot Y_t + A_{2,j} \qquad \text{j} = 1, \ldots, 50$$

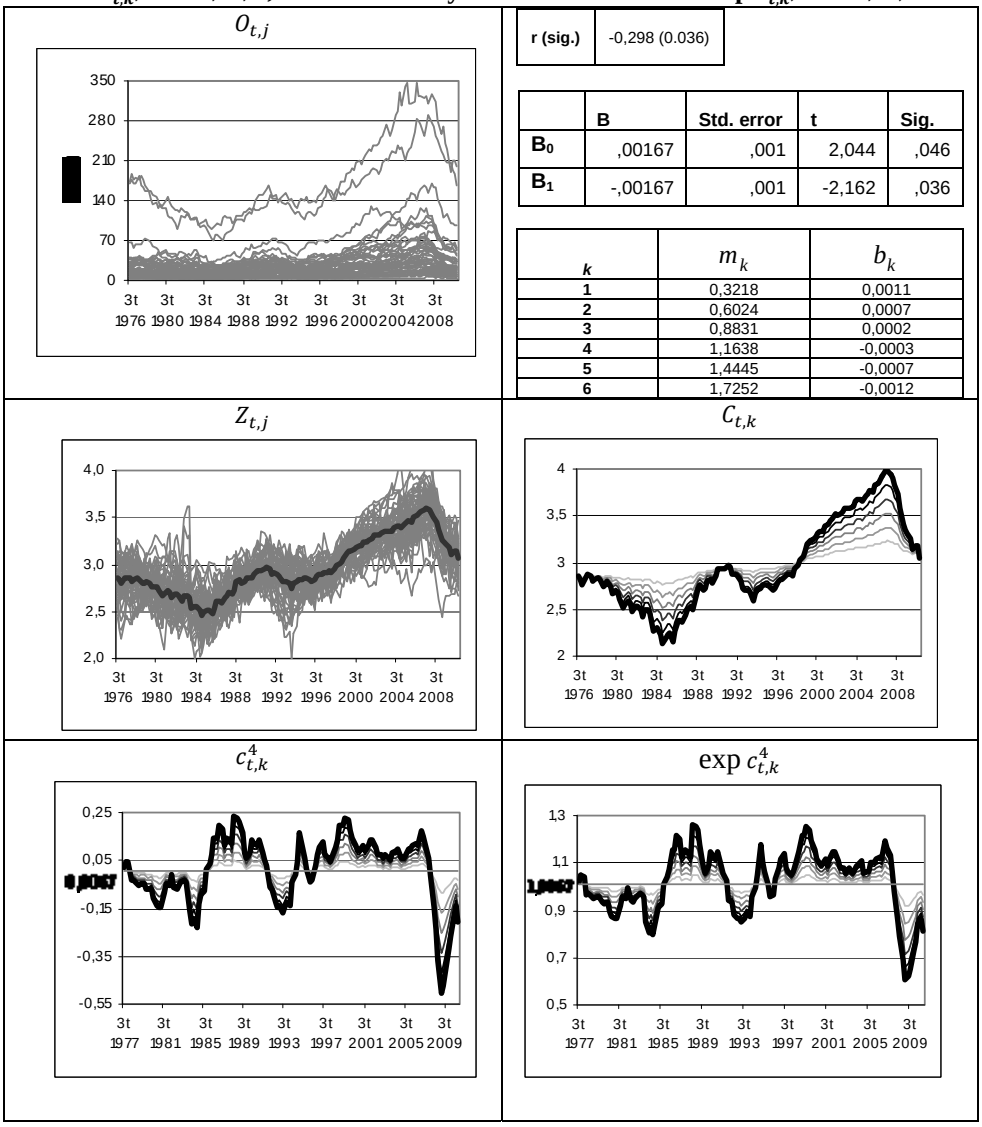
en cada una de las cincuenta provincias.



[8] En principio sería necesario disponer de la serie que la genera pero, según la **Observación 5**, podemos considerar en su lugar la serie promedio.

Figura 3

**Sup. Izqda.: Trayectorias de las series $O_{t,j}$ , j = 1, … , J ; Sup. Dcha.: Correlación de Pearson entre las secuencias $(A_{0,1}, \dots, A_{0,J})$ y $(A_{1,1}, \dots, A_{1,J})$, 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 $c_{t,k}^4$, k = 1, … , 6; Inf. Dcha: Trayectorias de las curvas exp $c_{t,k}^4$, k = 1, … , 6.**




| 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 |

Fuente: Elaboración propia

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.

Tabla 1

**Valores de $R_j^2, A_{0,j}, A_{1,j}, \alpha\ y\ \alpha_j, j = 1, \ldots 50$**

| *Provincia* | $R_j^2$ | $A_{0,j}$ | $A_{1,j}$ | $\frac{\alpha = 2{,}966}{\alpha_j}$ |
|---|---|---|---|---|
| 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 |

| *Provincia* | $R_j^2$ | $A_{0,j}$ | $A_{1,j}$ | $\frac{\alpha = 2{,}966}{\alpha_j}$ |
|---|---|---|---|---|
| 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

Fuente: Elaboración propia a partir de datos del INE

Por otro lado (Fig. 3, *sup. dcha.*), el coeficiente de correlación de Pearson entre las secuencias $(A_{0,1}, \ldots, A_{0,J})$ y $(A_{1,1}, \ldots, A_{1,J})$, es estadísticamente significativo; además, al ajustar sobre la nube de *J* puntos $(A_{1,j}, \ldots, A_{0,j})$ la ecuación de regresión lineal:

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

podemos concluir que tanto $B_0$ como $B_1$ son significativamente distintos de cero. En otras palabras, podemos concluir que en el conjunto $\{Y_{t,j}\}$ subyace una estructura de haz de rectas.

### 4.2 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 $Z_{t,j} = Y_{t,j} - \alpha_j + \alpha \quad j = 1, \ldots, J$ (Fig. 3, *ctro. izqda.*), donde $\alpha_j = \bar{Y}_j$ y $\alpha$ es la escala común elegida $(\bar{Z}_j = \alpha \ \forall j)$. Si, por ejemplo, elegimos $\alpha = \bar{Y}$ entonces las medias de las *J* series transformadas serán iguales a la media de la serie promedio[9]. La Tabla 1 ofrece los valores de $\alpha$ y $\alpha_j$ para las cincuenta provincias.

[9] Obsérvese que estamos utilizando la misma nomenclatura que en la expresión [1]; así, como veremos en el siguiente apartado, al construir un conjunto de series resumen que verifique las hipótesis de la Proposición 1 su escala, $\alpha$ , será la del conjunto de series homogeneizadas.

### 4.3 Extracción del conjunto de curvas resumen

A partir de los valores de $B_0$ y $B_1$ (Fig. 3, *sup. dcha.*) los pasos a seguir en el proceso de construcción del conjunto de series resumen son los siguientes:

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

*Paso 2*: Para $k = 1, \dots, K$, calcular $C_{t,k} = g_{t,k} - \beta_k + \alpha$ , donde $g_{t,k} = b_k.t + m_k.Y_t$ y $\beta_k = \bar{g}_k$

Obsérvese que si denominamos $\mu_k = -\beta_k + \alpha$, entonces:

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

Además:

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

En otras palabras, el conjunto de series resumen $\{C_{t,k}\},\ k = 1,\dots,K$ , construido según los pasos 1 y 2 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), fijemos *K* = 6[10] coeficientes de la forma $m_1$ = 0,3218 y $m_k = m_{k-1} + \theta$ para *K* = 2,…,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 $B_0$ y $B_1$, calculemos los coeficientes $b_k = 0{,}00167 + (-0{,}00167).m_k$, $k = 1, \dots, 6$. Para cada par de valores $m_k$ y $b_k$ (Fig. 3, *sup. dcha.*) calculemos, según el *Paso 2*, la serie $g_{t,k}$ y su correspondiente media $\beta_k = \bar{g}_k$. La serie resumen viene dada por:

$$C_{t,k} = g_{t,k} - \beta_k + \alpha_k$$

Así, las series del conjunto $\{C_{t,k}\}$ están en la misma escala que las del conjunto $\{Z_{t,j}\}$ (Fig. 3, *ctro. izqda y dcha.*).

### 4.4 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

[10] 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.

respecto de $Y_t$, las seis series $c_{t,k}^4 = c_{t,k}^4 - c_{t-4,k}^4 = b_k + m_k . y_t^4$, con $y_t^4 = Y_t - Y_{t-4}$, forman un haz de rectas respecto de $y_t^4$ (Fig. 3, *inf. izqda.*) de vértice:

$$(y_t^4, c_{t,k}^4) = (-4 \cdot B_1, 4 \cdot B_0) = (0.0067, 0.0067).$$

Así el orden de las series resumen $C_{t,k}$ viene dado por el orden de las series $c_{t,k}^4$ que, a su vez, viene dado por su grado de sensibilidad frente a $y_t^4$. Más concretamente, por el apartado A) de la Proposición 1, dado que $m_1 < m_2 < m_3 < m_4 < m_5 < m_6$ entonces la secuencia $C_{t,1}, C_{t,2}, \ldots, C_{t,6}$ es tal que:

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

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

Además, por la Observación 4, en aquellos instantes en que las trayectorias $c_{t,k}^4$ 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 $\exp c_{t,k}^4$ 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 trayectorias de series del conjunto $\{C_{t,k}\}$ se cortan también lo hacen las restantes. Así como la línea que conecta los puntos de corte permite resumir la trayectoria de la serie $Y_t$[11], también la línea que conecta los puntos de corte de las series del conjunto $\{\exp C_{t,k}\}$ permite resumir la de la serie $\exp Y_t$ (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

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

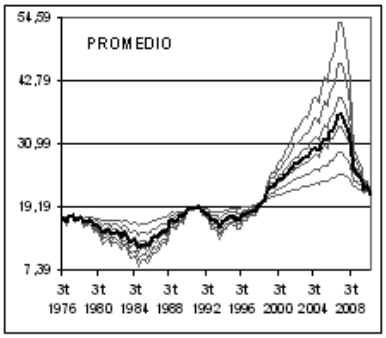
que dicho porcentaje coincide (salvo errores de redondeo) con el correspondiente a elevar el valor estimado de la tendencia media de crecimiento interanual al número de años del periodo de observación:

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

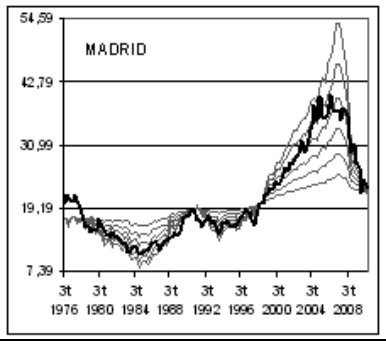
En definitiva esta relación entre los puntos de corte es lo que la serie exp $Y_t$ y las series exp $C_{t,k}$, $k = 1,\ldots,6$, presentan en común. Lo que las diferencia es la trayectoria seguida entre ellos: mientras que entre cualquier par de puntos de corte la serie exp $C_{t,1}$ se desvía muy poco 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.

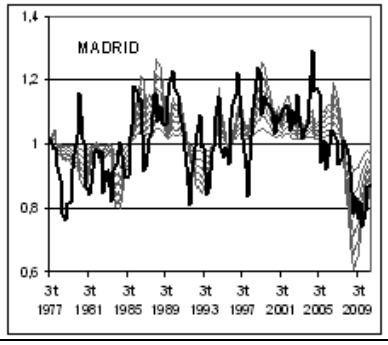
Figura 4

***Sup. Izqda.*: Serie exp $Y_t$ representada sobre el conjunto exp $C_{t,k}$, $k = 1,\ldots,6$; *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 $\{\exp C_{t,k}\}$, $k = 1,\ldots,6$; *Inf. Dcha*: Serie de incrementos interanuales de ocupados en el sector de la construcción en Madrid representada sobre el conjunto $\{\exp c_{t,k}^4\}$**




| 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% |




Fuente: Elaboración propia a partir de datos del INE

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 en comparación con el comportamiento medio.

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.*) sobre el conjunto $\{\exp c_{t,k}^4\}$; 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.

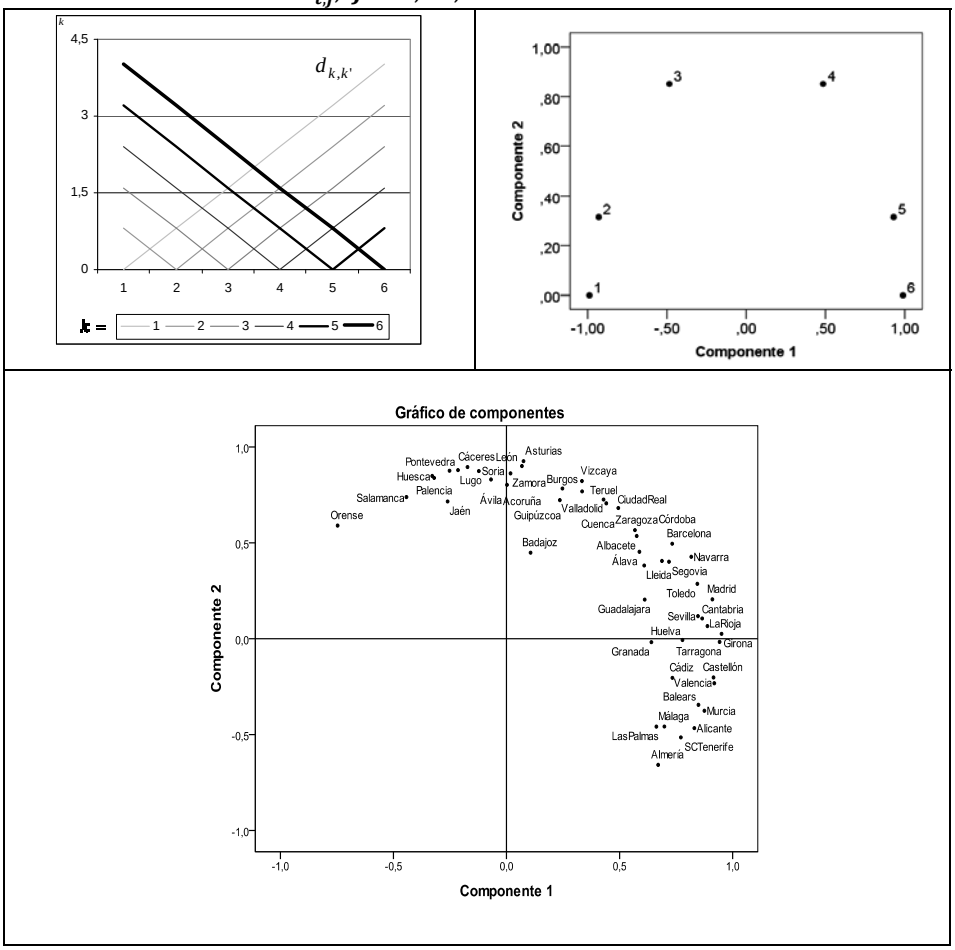
### 4.5 Comparación de las trayectorias

Dado el conjunto de series $\{C_{t,k}\}$, consideremos la distancia entre cada par de ellas (Fig. 5, *sup. izqda.*): $d_{k,k'} = d^2(C_{t,k}, C_{t,k'}),\ k, k' = 1, \ldots, K$, 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[12].

Figura 5

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




Fuente: Elaboración propia a partir de datos del INE

[12] 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.

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

En términos globales los distintos puntos-provincia están bien representados[13], 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.

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:

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

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:

$$C_{t',k}^{j} - C_{t,k}^{j} = C_{t',k} - C_{t,k} = C_{t',k'} - C_{t,k'} = C_{t',k'}^{j} - C_{t,k'}^{j}$$

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

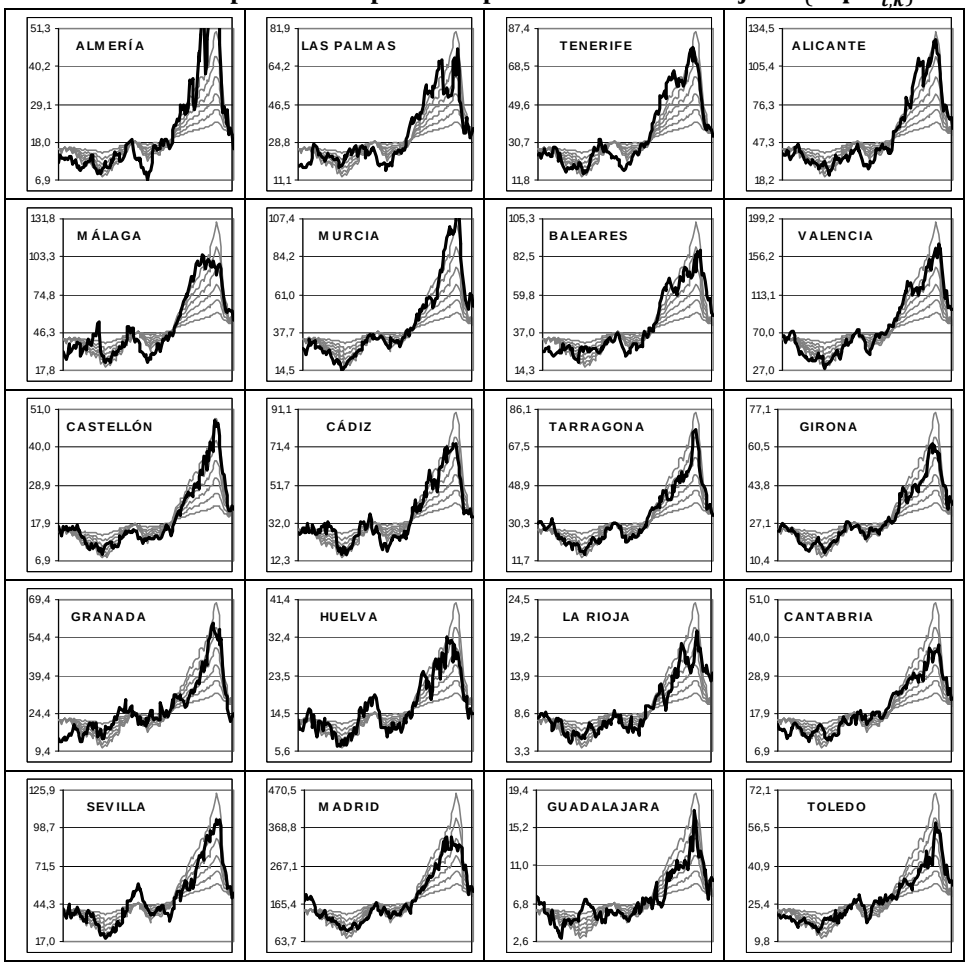
Aunque la representación del número de ocupados en el sector de la construcción en cada provincia, $O_{t,j} = \exp Y_{t,j}$, (Fig. 6) se realizará con referencia al conjunto $\{\exp C_{t,k}^{j}\}$ (eje izquierdo), para comparar las distintas provincias utilizaremos como referencia el

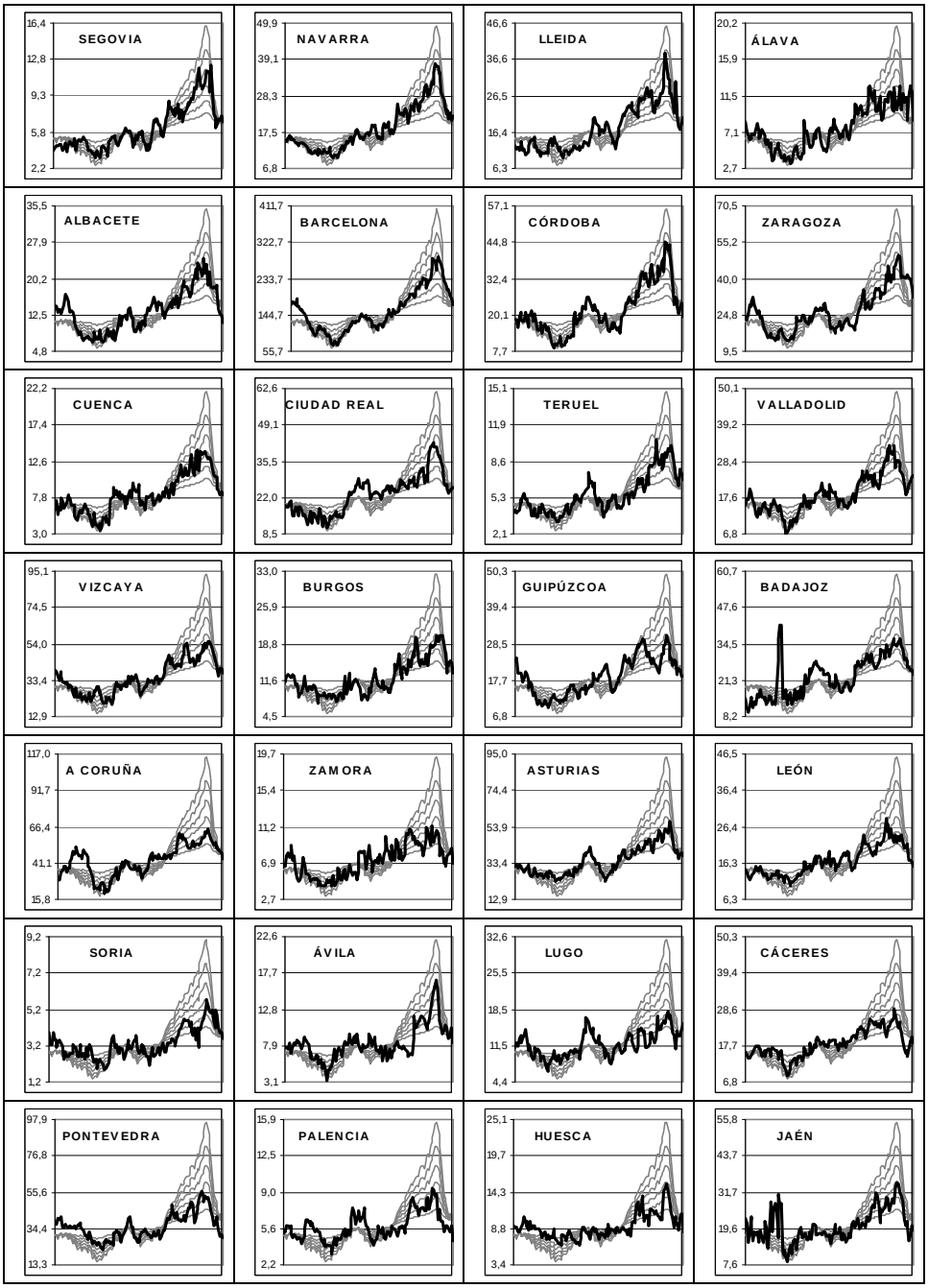
[13] La calidad de representación de un punto viene dada por su distancia al origen, que a lo sumo puede tomar el valor 1.

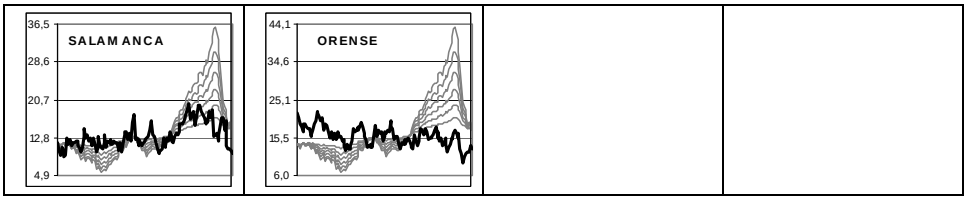
conjunto $\{\exp C_{t,k}\}$ (eje derecho). La secuencia de gráficos de la Figura 6 responde al trazado de la línea imaginaria que va desde Almería hasta Orense en la Figura 5, así las primeras provincias representadas son tales que su trayectoria va pareja a la serie resumen de mayor sensibilidad frente a la serie promedio (la sexta), como por ejemplo Málaga y Las Palmas, y la de las últimas, pareja a la serie resumen de menor sensibilidad (la primera), como por ejemplo Palencia y Huesca; el resto de provincias, tales como Navarra y Segovia o como Vizcaya y Valladolid se encuentra entre ambos extremos.

Figura 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 $\{\exp C_{t,k}^{j}\}$**




SEGOVIA
16,4 12,8 9,3 5,8 2,2
NAVARRA
49,9 39,1 28,3 17,5 6,8
LLEIDA
46,6 36,6 26,5 16,4 6,3
ÁLAVA
20,2 15,9 11,5 7,1 2,7
ALBACETE
35,5 27,9 20,2 12,5 4,8
BARCELONA
411,7 322,7 233,7 144,7 55,7
CÓRDOBA
57,1 44,8 32,4 20,1 7,7
ZARAGOZA
70,5 55,2 40,0 24,8 9,5
CUENCA
22,2 17,4 12,6 7,8 3,0
CIUDAD REAL
62,6 49,1 35,5 22,0 8,5
TERUEL
15,1 11,9 8,6 5,3 2,1
VALLADOLID
50,1 39,2 28,4 17,6 6,8
VIZCAYA
95,1 74,5 54,0 33,4 12,9
BURGOS
33,0 25,9 18,8 11,6 4,5
GUIPÚZCOA
50,3 39,4 28,5 17,7 6,8
BADAJOZ
60,7 47,6 34,5 21,3 8,2
A CORUÑA
117,0 91,7 66,4 41,1 15,8
ZAMORA
19,7 15,4 11,2 6,9 2,7
ASTURIAS
95,0 74,4 53,9 33,4 12,9
LEÓN
46,5 36,4 26,4 16,3 6,3
SORIA
9,2 7,2 5,2 3,2 1,2
ÁVILA
22,6 17,7 12,8 7,9 3,1
LUGO
32,6 25,5 18,5 11,5 4,4
CÁCERES
50,3 39,4 28,6 17,7 6,8
PONTEVEDRA
97,9 76,8 55,6 34,4 13,3
PALENCIA
15,9 12,5 9,0 5,6 2,2
HUESCA
25,1 19,7 14,3 8,8 3,4
JAÉN
55,8 43,7 31,7 19,6 7,6




Fuente: Elaboración propia a partir de datos del INE

## 5. Conclusiones

La representación gráfica de múltiples series temporales relativas a distintas zonas geográficas se está convirtiendo en un problema cada vez más habitual desde el momento en que, para poder observar la evolución de los distintos indicadores económicos, se hacen necesarias comparaciones longitudinales regionales. Uno de los principales inconvenientes radica en la falta de homogeneidad en la escala de las distintas series, razón por la que estas comparaciones han implicado tradicionalmente la construcción de complejas medidas de similitud. E.J. Keogh y S. Kasetty (2002) realizan una revisión exhaustiva de las aportaciones realizadas en el campo de la comparación masiva de series temporales con el objetivo de advertir a la comunidad científica en este campo de investigación sobre el hecho de que gran parte de las propuestas tienen muy poca generalización a otros problemas distintos de aquel para el que han sido diseñadas. Afirman que en la mayor parte de los trabajos el objetivo perseguido es el de construir nuevas medidas de similitud y que las propuestas realizadas no ofrecen mejores resultados que la sencilla y bien conocida distancia euclídea. Para justificar esta afirmación realizan un experimento de agrupación de series temporales sobre dos conjuntos de datos ampliamente referenciados en la literatura[14] y para los que el grupo de pertenencia es conocido. Tras comparar los resultados de la agrupación con doce medidas de similitud distintas observan que ninguna de ellas mejora la solución proporcionada por la distancia euclídea; más aún, comprueban que la mayoría de las distancias proporcionan el mismo resultado que si la asignación hubiera sido realizada aleatoriamente. En definitiva concluyen que, aunque el resultado de este experimento no implica necesariamente que las medidas de similitud no tengan su interés, la contribución de una nueva medida que no pueda demostrar su utilidad debe ser cuestionada.

Como parte de la fase exploratoria de los datos y como paso previo a la construcción de un modelo estadístico o económico que persiga objetivos de tipo explicativo o predictivo, la principal utilidad de las representaciones gráficas de series temporales geográficas debe radicar en su claridad a la hora de facilitar la interpretación de las similitudes y diferencias entre las distintas regiones. Bajo este punto de vista podemos concluir que la metodología del haz de rectas es una herramienta gráfica muy eficaz para la comparación provincial de la ocupación en el sector de la construcción. Sin embargo, a la hora de demostrar la utilidad de una metodología, no basta con que

[14] Cylinder-Bell-Funnel y Control-Chart.

solucione el problema para el que ha sido diseñada sino que debe ser generalizable. Recordemos que la Metodología del haz de rectas se fundamenta en la hipótesis de que la estructura que subyace en conjunto de series objeto de análisis es la de un haz de rectas. Es fácil comprobar[15] que son numerosos los conjuntos de datos cuya estructura puede ser resumida mediante un haz de rectas, tanto los procedentes del mercado laboral o del sector de la construcción, como los procedentes de otros sectores económicos en general o de otros campos de investigación, y para los que, en consecuencia, el resultado de la aplicación de la metodología puede ser muy esclarecedor.

## Referencias

ANDRIENKO, N.; ANDRIENKO, G. Y GATALSKY, P. (2003), «Exploratory spatio-temporal visualization: an analytical review», *Journal of Visual Languages & Computing*, 14, 503-541.

GELMAN, A. Y UNWIN, A. (2011), «Visualization, Graphics and Statistics», *Statistical Computing & Graphics Newsletter*, 22, 9-12.

HOCHHEISER, H. Y SHNEIDERMAN, B. (2001), «Interactive exploration of time series data», In: The 4th International conference on Discovery Science (Washington, DC), Springer-Verlag, Berlin, 441-446.

KEOGH, E.J. Y KASETTY, S. (2002), «On the need for time series data mining benchmarks: A survey an empirical demonstration», Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining (KDD'02), 102-111.

KOSARA, R. (2011), «Visualization: is more than Pictures!», *Statistical Computing & Graphics Newsletter*, 22, 5-8.

LIN, J.; KEOGH, E. Y LONARDI, S. (2005), «Visualizing and discovering non-trivial patterns in large time series databases». Information Visualization, 4, 61-82.

TUFTE, E.R. (2001), «The visual display of quantitative information», Cheshire, Conn: Graphic Press.

VAN WIJK, J.J. Y VAN SELOW, E.R. (1999), «Cluster and calendar based visualization of time series data», In: 1999 IEEE Symposium on Information Visualization (San Francisco, CA), 4-9.

WEBER, M.; ALEXA, M. Y MULLER, W. (2001), «Visualizing time series on spirals», In: 2001 IEEE Symposium on Information Visualization (San Diego, CA), 7-14.

ZHAO, J; CHEVALIER, F.; PIETRIGA, E. Y BALAKRISHNAN, R. (2011), «Exploratory Analysis of Time-Series with ChronoLenses», IEEE Transaction on visualization and computers graphics, 17(12), 2422-2431.

[15] Véase el Apartado 4.1.