CAPÍTULO VII

Explorando pautas en series estacionales múltiples mediante técnicas multivariantes

Enrique Martín Quilis*

Se utilizan técnicas multivariantes (modelos fatoriales, análisis de conglomerados) para identificar pautas comunes y específicas en un vector de series temporales de elevada dimensión cuya sección cruzada es de naturaleza espacial. Estas técnicas se aplican en un contexto de series temporales de alta frecuencia caracterizadas por la presencia de diversos componentes (tendencia, ciclo, estacionalidad, efectos de calendario) que implica un notable aumento de la dimensión efectiva del conjunto de datos. La metodología propuesta es aplicada a una base de datos territorial de la economía española cuya cobertura es muy amplia, tanto temporal (1974-2019) como espacial (nivel provincial).

Palabras clave: estacionalidad, ciclos, extracción de señales, alieneamiento dinámico óptimo, análisis de conglomerados, análisis factorial.

* Agradezco la colaboración de Ana Abad, José Antonio Campo y Rafael Frutos en diversas etapas de este proyecto, así como las sugerencias de Daniel Peña, Pilar Poncela y Esther Ruiz. Las opiniones presentadas corresponden al autor, sin que coincidan de forma necesaria con las de la Agencia Tributaria.

1. INTRODUCCIÓN

El análisis de *big data* se caracteriza por "las tres V": volumen, variedad (o diversidad) y velocidad (Sathi, 2012; Kolanovic y Krishnamachari, 2017; Diebold, 2020). Las series temporales de datos económicos de alta frecuencia (por ejemplo, mensual) y muy alta frecuencia (por ejemplo, diaria) están formadas por una multiplicidad de componentes subyacentes muy distintos entre sí: atípicos, efectos de calendario (por ejemplo, fiestas móviles y laboralidad), tendencia, ciclo, estacionalidad e irregularidad. De esta manera y de forma casi automática, adquieren una dimensión muy superior a la de los datos observados, haciéndolo además de forma heterogénea. Aparecen así, rápidamente, las dos primeras V (volumen y variedad), afectando a la tercera V (velocidad), al complicarse el procesamiento de los datos y reducir por consiguiente su velocidad.

En este trabajo se van a utilizar dos técnicas centrales en el análisis de *big data,* análisis de conglomerados y modelos factoriales, en un contexto de series temporales múltiples cuya dimensión transversal tiene un significado geográfico. Este contexto dinámico va a requerir un uso bastante intensivo de técnicas de modelización de series temporales y de extracción de señales para realizar un adecuado procesamiento y descomposición de las series observadas. Asimismo, la elección de un concepto de distancia adaptado a la naturaleza dinámica de los datos es esencial.

La metodología econométrica consta de tres etapas, siguiendo el principio de "divide y conquista" para hacer frente a la maldición de las dimensiones que pesa sobre el análisis de series temporales múltiples. En la primera se realiza un análisis univariante de todas las series consideradas, con el fin de determinar cuál es la transformación de Box-Cox más apropiada, aislar los efectos deterministas (atípicos, fiestas móviles y ciclo semanal) de los estocásticos y realizar una descomposición de estos últimos en tendencia-ciclo, estacionalidad e irregularidad. A su vez, la aplicación de métodos de extracción de señales permite separar la tendencia del ciclo.

La segunda etapa es de carácter bivariante y trata de estimar una matriz de distancia mediante la aplicación del algoritmo de alineamiento óptimo (*DTW*, por *Dynamic Time Warping*) a todos los pares de series. Este algoritmo tiene en cuenta tanto la naturaleza dinámica de los objetos cuya semejanza se desea medir como la posibilidad de que existan desfases temporales entre ellos (por ejemplo, relaciones de adelanto o desfase). La utilización del enfoque DTW permite enlazar de manera consistente las etapas 1 (univariante, dinámica) con la 3 (multivariante, estática).

Finalmente, en la tercera etapa se utilizan métodos de formación de conglomerados (tanto jerárquicos como partitivos) y de análisis factorial para identificar pautas comunes en los componentes estacional y cíclico de las series consideradas. Estos métodos permitirán comprobar en qué medida ambas aglomeraciones son afines, tanto espacial como dinámicamente.

Desde un punto de vista económico, se examina la relación entre la estacionalidad (un componente estructural, no estacionario y candidato natural para incorporar información geográfica) y el ciclo (un componente transitorio aunque persistente, asintóticamente estacionario y muy condicionado por los factores macroeconómicos de corto y medio plazo). En particular, se trata de responder a la pregunta: ¿resuelven los agentes económicos sus programas de optimización tomando la estacionalidad como una restricción exógena o, por el contrario, determinan conjuntamente su comportamiento tanto estacional como no estacional? Esta segunda posibilidad extiende la crítica a la distinción entre tendencia y ciclo basada en los modelos estocásticos de crecimiento óptimo a la diferenciación entre los componentes estacional y no estacional (Prescott, 1986; Todd, 1990). Desde este punto de vista, los agentes económicos resuelven sus programas de optimización de una forma indiferenciada, de manera que los elementos estacional y no estacional de sus decisiones son aspectos distintos de un único proceso de decisión, siendo meramente facetas diferentes de una misma respuesta a impulsos comunes (Barsky y Miron, 1989; Beaulieu, MacKie-Mason y Miron, 1992; Cecchetti, Kashyap y Wilcox, 1997; Gerenew y Gourio, 2018).

Si la agrupación estacional coincide con la cíclica, se obtiene evidencia a favor de este punto de vista y, en particular, de la conveniencia de realizar un análisis del ciclo y de la coyuntura desagregado según las pautas estacionales. Si, por el contrario, la información estacional y la cíclica guardan poca relación, la balanza se inclina a favor de la exogeneidad del fenómeno estacional.

La estructura del trabajo es la siguiente. En la segunda sección se exponen los métodos de estimación de los componentes estacional y cíclico. A continuación, se describe el método de cálculo de la distancia. La cuarta sección presenta la metodología de formación de conglomerados utilizada y, en la quinta, se describen los datos empleados. Los resultados empíricos se ofrecen en la sexta sección. El trabajo termina con un apartado de conclusiones.

2. ESTACIONALIDAD Y CICLOS

El procedimiento utilizado para estimar los componentes estacional y cíclico de las series temporales analizadas consta de tres etapas: corrección de los efectos asociados a las observaciones atípicas y efectos de calendario, extracción basada en modelos ARIMA de las señales estacional y de tendencia-ciclo y, por último, estimación del ciclo por medio de un filtro de paso en banda aplicado a la serie de tendencia-ciclo obtenida en la etapa anterior. A continuación, se describe brevemente cada fase.

Se considera que la serie temporal observada puede ser expresada de acuerdo con la siguiente expresión:

$$z_t = o_t + n_t, \qquad [1]$$

siendo z_t, con $t = 1..n$, la serie observada, posiblemente transformada mediante la función de Box-Cox; o_t es el componente determinista, resultado de la combinación de efectos de

calendario (ciclo semanal, Pascua móvil) e intervenciones ligadas a valores atípicos y n_t es el componente estocástico.

Por su parte, o_t representa una combinación de modelos de intervención asociados a factores de tipo extraordinario que afectan a la serie de manera no recurrente junto con los efectos de calendario vinculados con el ciclo semanal y la Pascua móvil. La expresión completa es:

$$o_t = \Sigma_{h=1}^{k} V_h(B) I_t(t_h) + \Sigma_{j=1}^{7} \beta_j d_{j,t} + \gamma P_t(\tau), \qquad [2]$$

donde $I(t_h)$ es una variable binaria de tipo impulso, siendo t_h el periodo en el que tiene lugar el acontecimiento atípico. El filtro $V_h(B)$ recoge los efectos dinámicos asociados a la observación anómala. En este trabajo se consideran tres posibles tipos de atípicos: aditivos, transitorios y cambios de nivel. Asimismo, $d_{j,t}$ = *[(número de días de tipo j en el mes t) – (número de domingos en el mes t)],* con j = *lunes, ..., sábados* y $d_{7,t}$ es la diferencia entre la duración del mes t y la duración media de todos los meses. Finalmente, $P_t(\tau)$ representa la proporción que representa la semana de Pascua en el mes t, habiéndose considerado que su efecto se registra en los días anteriores al Domingo de Resurrección. En este trabajo se asume $\tau = 8$.

La especificación del componente estocástico sigue una representación autorregresiva, integrada y de medias móviles (ARIMA) de tipo multiplicativo (Box y Jenkins, 1976):

$$n_t = \frac{\theta_q(B)\Theta_Q(B^{12})}{\phi_p(B)\Phi_P(B^{12})(1-B)^d(1-B^{12})^D} a_t, \qquad [3]$$

donde $\phi(B)$ y $\theta_q(B)$ son, respectivamente, polinomios de orden p y q en el operador de desfases B y $\Phi_p(B^{12})$ y $\Theta_Q(B^{12})$ son polinomios de orden P y Q en B^{12}. Los filtros $(1\text{-}B)^d$ y $(1\text{-}B^{12})^D$ son operadores de diferenciación regular y estacional, controlados por los parámetros enteros d y D, respectivamente. Por último, a_t es una secuencia de ruido blanco con esperanza nula y desviación típica constante σ_a.

A su vez, el término estocástico n_t admite una descomposición, según la hipótesis de los componentes subyacentes, en tendencia-ciclo (p_t), estacionalidad (s_t) e irregularidad (i_t):

$$n_t = p_t + s_t + i_t. \qquad [4]$$

Una vez estimado el modelo ARIMA con análisis de intervención (AI) descrito en [1]-[3], es posible extraer tanto una señal estacional como una de tendencia-ciclo aplicando filtros de error cuadrático medio mínimo compatibles con dicho modelo ARIMA. De esta manera, se obtiene una estimación de los componentes subyacentes adaptada a las propiedades dinámicas de la serie y, merced al principio de descomposición canónica, libre de elementos irregulares de tipo ruido blanco. Una descripción muy completa de los métodos y modelos para el tratamiento de observaciones atípicas, efectos de calendario, modelización univariante y descomposición de series temporales se encuenta en Peña, Tiao y Tsay (2001).

La expresión general de este proceso de filtrado para la estacionalidad es:

$$\hat{s}_t = V_s(B,F)\hat{n}_t = k_s \Pi(B)\Pi(F)\Psi_s(B)\Psi_s(F)\hat{n}_t. \quad [5]$$

Para la tendencia-ciclo:

$$\hat{p}_t = V_p(B,F)\hat{n}_t = k_p \Pi(B)\Pi(F)\Psi_p(B)\Psi_p(F)\hat{n}_t, \quad [6]$$

donde k_s y k_p son parámetros que normalizan las funciones de ganancia de los filtros, $\Pi(B)$ es la expansión autorregresiva del modelo ARIMA de n_t, $\Psi_i(B)$ $(i = s, p)$, es la expresión de medias móviles del modelo teórico de los componentes y $\hat{n}_t$ es la estimación del componente estocástico obtenida al eliminar de la serie observada z_t sus elementos deterministas o_t (Maravall, 1987).

La señal de tendencia-ciclo así obtenida permite la estimación de un componente cíclico independiente al aplicar a aquella un filtro de paso en banda diseñado desde el dominio de la frecuencia. Dicho filtro es de tipo Butterworth, especificado para aproximar con una precisión dada a uno cíclico de tipo ideal (Pollock, 1999; Oppenheim y Schaffer, 1989; Bógalo y Quilis, 2003; Proakis y Manolakis, 2006).

De esta forma, el componente cíclico se obtiene según la siguiente expresión:

$$\hat{c}_t = H_c(B,F)\hat{p}_t = H_c(B,F)k_p \Pi(B)\Pi(F)\Psi_p(B)\Psi_p(F)\hat{n}_t, \quad [7]$$

donde $H_c(B,F)$ es el filtro cíclico de paso en banda antes mencionado y c_t es la señal cíclica.

Este método bietápico puede ser interpretado de forma bayesiana, ya que combina información *a priori* (un filtro fijo de tipo Butterworth) con la contenida en la tendencia estimada a partir de la muestra (por medio de un filtro adaptable de tipo Wiener-Kolmogorov). De esta manera, se obtiene la información *a posteriori*: una serie de fluctuaciones alrededor de la tendencia de periodicidad comprendida entre dos y ocho años.

Las principales ventajas de este enfoque son:

- La estimación de los componentes es compatible con las propiedades agregadas de las series, de forma que se evita la inducción de fenómenos espurios como, por ejemplo, la estimación de un componente estacional en una serie que carece de estacionalidad.
- El filtro usado en la estimación se adapta a las características de la serie observada, de forma que series con características estacionales distintas tendrán asimismo filtros distintos.
- El preprocesamiento de las series mediante el modelo ARIMA-AI permite estimar los componentes estocásticos sin la influencia distorsionadora asociada a las observaciones atípicas y a los efectos de calendario, lo que redunda en una mejor estimación de los mismos.

3. MEDIDA DE DISTANCIA

La formación de conglomerados se basa en dos elementos principales: una matriz de distancia (o similitud) entre los objetos que se desea agrupar y un procedimiento o algoritmo que, a partir de dicha distancia, determina la relación de pertenencia de los objetos respecto a los grupos. Cuando los objetos que se desea agrupar son series temporales, la definición de un concepto de distancia deviene especialmente difícil, debido a su multidimensionalidad implícita asociada a la existencia de componentes subyacentes (por ejemplo, tendencia, estacionalidad, etc.) y a la posibilidad de que existan relaciones dinámicas entre ellas (por ejemplo, relaciones de adelanto, coincidencia o retraso).

Por todo ello, se han presentado diversas medidas de distancia para series temporales. Estas medidas pueden basarse en un enfoque no paramétrico, bien en el dominio del tiempo o de la frecuencia (Caiado, Crato y Peña, 2006; Caiado, Crato y Poncela, 2020), o basado en modelos (Piccolo, 1990). La elección del enfoque es objeto de un amplio debate, muy condicionado por la naturaleza de las series temporales y el objetivo último de la agrupación (Galeano y Peña, 2000; Liao, 2005; Wang, Smith y Hyndman, 2006; Rani y Sikka, 2012).

En este trabajo se utiliza el algoritmo de alineamiento dinámico temporal DTW para calcular una matriz de distancia entre los elementos de un vector de series temporales (Sakoe y Chiba, 1978). DTW es un enfoque no paramétrico, basado directamente en la semejanza entre perfiles temporales (*shape-based*) y que se adapta muy bien a las características de los objetos que se desea clasificar: componentes subyacentes estimados mediante técnicas de extracción de señales.

El procedimiento DTW consta de dos etapas: cálculo de una matriz de distancia inicial y algoritmo de alineación óptima mediante programación dinámica. La matriz de distancia final entre todos los pares de series temporales se obtiene aplicando dicho algoritmo a las correspondientes matrices iniciales. A continuación, se describen ambas etapas.

3.1. Matriz inicial de distancia

Asumiendo que $z_{i,t}$ y $z_{j,t}$ son dos componentes de un vector de series temporales de dimensión k, $Z_t = (z_{1,t}...z_{k,t})'$, la matriz de distancia inicial toma como punto de partida la diferencia absoluta entre todas las observaciones efectuadas en t y en s de las dos series[1]:

$$C_{t,s} = |z_{i,t} - z_{j,s}| \qquad t, s = 1...n. \qquad [8]$$

La información contenida en la matriz anterior se va acumulando sobre todos los pares (t,s) mediante la siguiente recursión:

$$D_{t,s} = C_{t,s} + min\,[D_{t-1,s}, D_{t-1,s-1}, D_{t,s-1}] \qquad t, s = 2...n. \qquad [9]$$

[1] Con el fin de aligerar la notación, se omiten los índices *i* y *j*.

La condición inicial para la recursión [9] es $D_{1,1} = C_{1,1}$. Si el número de observaciones es elevado, la matriz D puede adquirir un gran tamaño, resultando su cálculo computacionalmente costoso. Por esta razón y con el fin de reducir la carga numérica, se suele acotar su cálculo a una ventana de proximidad entre las observaciones (p.e. una distancia entre t y s inferior al 10 % del tamaño muestral).

La matriz [9] representa, para cada par temporal (t,s), una medida de la similitud entre las series $z_{i,t}$ y $z_{j,t}$ acumulada hasta dicho par. Para obtener una medida sintética de distancia entre ambas series es necesario transformar dicha matriz en un escalar. Esta transformación se realiza mediante el algoritmo de alineación óptima que se describe a continuación.

3.2. Algoritmo de alineación óptima

El primer paso de este algoritmo consiste en determinar un emparejamiento entre las observaciones de $z_{i,t}$ y $z_{j,t}$ de manera que la similitud entre ellas, cuantificada mediante [9], sea máxima. Como dicha cuantificación se realiza de forma acumulativa desde $t = 1$ hasta $t = n$, la determinación de dicho emparejamiento se realiza en sentido inverso, $t = n$ hasta $t = 1$, siguiendo el enfoque de la programación dinámica.

De esta manera, tomando $D_{n,n}$ como condición inicial, se retrocede hasta la observación inicial y se emparejan las observaciones de $z_{i,t}$ y $z_{j,t}$ para las que D es mímina, utilizando el mismo entorno que se ha considerado en la recursión [9] mediante la que ha sido calculada. Así, avanzando en sentido temporal inverso y minimizando D en cada paso, se determina una secuencia para ambas variables, $i(t)$ y $j(t)$, de forma que su similitud es máxima. A continuación se detalla el procedimiento.

La condición inicial del algoritmo se sitúa al final de la muestra temporal: $t,s = n$; $i(t) = j(t) = n$ y $c = 1$. A partir de ahí, se aplica un bucle inverso desde $t = n$ hasta $t = 1$:

$$a,b = arg\ min_{\alpha,\beta}(D_{t-\alpha,s}, D_{t-\alpha,s-\beta}, D_{t,s-\beta}). \qquad [10]$$

En este paso se determinan los índices temporales de las dos series $z_{i,t}$ y $z_{j,t}$ de forma que se emparejan de forma óptima, minimizando la distancia en el mismo entorno sobre el que ha sido calculada según [9].

Una vez que se ha determinado el par (a,b), se actualiza el pivote temporal: $t = t\text{-}a$ y $s = s\text{-}b$ y se asigna este par temporal a la alineación óptima $i(t)$ y $j(t)$. Nótese que tanto a como b solo pueden adoptar los valores 0 o 1, en virtud de la recursión [9] usada para calcular D. Finalmente, se incrementa el contador de iteraciones, $c = c + 1$, y se repite [10] hasta alcanzar el par inicial $t = s = 1$.

Una vez que se ha determinado el emparejamiento óptimo, $i(t)$ y $j(t)$, la distancia final agregada entre las series $z_{i,t}$ y $z_{j,t}$ se calcula sumando todos los valores de la matriz D que han sido asociados mediante dicho emparejamiento:

$$DTW_{i,j} = \Sigma_{t=1}^{n} D_{i(t),j(t)}. \qquad [11]$$

Como ya se ha señalado, la matriz de distancia final [11] se calcula sobre los componentes subyacentes de un vector de series temporales[2]. Asimismo, estas matrices servirán de métrica para los algoritmos de formación de conglomerados que se describen en la siguiente sección.

4. FORMACIÓN DE CONGLOMERADOS

Una vez estimados los componentes estacionales de las series que se desea analizar, se procede a formar grupos de series con patrones estacionales similares. La formación de estas agrupaciones se realiza mediante el análisis de conglomerados (Everitt *et al.,* 2011; Kassambara, 2017).

El análisis de conglomerados es una técnica estadística no paramétrica, de tipo exploratorio, que agrupa los objetos atendiendo a su semejanza. Se puede considerar también como un método de aprendizaje no supervisado, mediante el que se busca la identificación de patrones a través de la detección de pautas o características similares, sobre las que se dispone de poca o ninguna información *a priori* (Sathi, 2012).

Estas dos características proporcionan al análisis de grupos una gran flexibilidad, ya que se adapta muy bien a situaciones que requieren la identificación de regularidades empíricas como paso previo a la elaboración de modelos estadísticos explícitos. Asimismo, el análisis de conglomerados resulta especialmente útil cuando se desea reducir drásticamente la dimensión de grandes masas de información, de forma que se selecciona un representante de cada grupo en lugar de la población completa. En este sentido, este tipo de análisis se asemeja a un muestreo endógeno en el que los propios objetos determinan su representante.

Existen diversos métodos de formación de conglomerados basados en criterios de optimización. En general, se trata de formar grupos con la máxima heterogeneidad entre ellos y la mínima dentro de cada uno de ellos. Naturalmente, el mayor problema de estos métodos radica en la explosión combinatoria a que da lugar una búsqueda exhaustiva. Con el fin de dotar al procedimiento de optimización de un contenido operativo en situaciones reales, se han propuesto diversos algoritmos de formación de un número dado de conglomerados. En este trabajo se utiliza el de las k-medias por su relativa eficiencia y por la solidez de su planteamiento teórico (Faber, 1994; Everitt *et al.,* 2011).

Se ha preferido un algoritmo partitivo en lugar de uno jeráquico por su mayor flexibilidad, escalibilidad y porque asegura la idoneidad de los respresentantes de cada grupo (centroides), propiedad muy importante si se desea utilizar el análisis de conglomerados para reducir la muestra a procesar. Como ya se ha señalado, la matriz de distancia sobre la que

[2] En particular, los componentes estacional y cíclico estimados mediante la metodología descrita en la sección segunda.

se aplica este algoritmo es la que ha sido calculada mediante el proceso DTW descrito en la sección anterior.

El algoritmo de las k-medias opera de la siguiente forma. En primer lugar, se realiza una agrupación aleatoria de los objetos en *G* conglomerados. En este trabajo los objetos que se desea agrupar son las series temporales asociadas a los componentes estacional y cíclico de un vector de series temporales. El número de conglomerados se asume como dado[3].

A continuación, se selecciona una serie temporal como representante (centroide) de cada grupo, buscando aquella que se encuentra, en promedio, más próxima a todas las demás que forman parte del mismo conglomerado. Una vez determinados los centroides, se revisa la asignación de series a grupos. De esta manera, se considera que el objeto *m* pertenece al grupo *h* si la distancia que lo separa del correspondiente centroide de los *G* grupos considerados es la menor posible:

$$m \in h \quad \Leftrightarrow h = arg\, min_g\, (DTW_{m,g}) \quad g = 1..G. \qquad [12]$$

Después de revisar la asignación de las series temporales a los grupos, se determinan los nuevos centroides (representantes) siguiendo el mismo criterio de distancia mínima antes expuesto, modificándose asimismo la asignación según el criterio definido en la expresión [12].

El proceso de cálculo de centroides y reasignación continúa hasta que se satisface algún criterio de convergencia (por ejemplo, que la variación en valor absoluto de todos los centroides sea inferior a un umbral predeterminado).

El algoritmo de las k-medias requiere que el número de conglomerados a formar sea conocido. En muchas aplicaciones, como la presente, esta información preliminar no está disponible y ha de ser obtenida mediante alguna investigación previa. En este trabajo se ha utilizado una aglomeración jerárquica mediante el método de Ward (1963) con el fin de disponer de una estimación del número de grupos. Se ha seleccionado este método porque incorpora explícitamente una función objetivo compatible con los criterios de optimización antes expuestos.

En resumen, el proceso de formación de conglomerados consta de dos etapas: 1) determinación preliminar del número de grupos *G* mediante el examen del dendrograma generado por el método (jerárquico) de Ward y, 2) aplicación del algoritmo de las k-medias, tomando *G* como número apropiado de conglomerados.

5. DATOS

Los datos utilizados en este trabajo son las pernoctaciones en establecimientos hoteleros procedentes de la *Encuesta de Ocupación Hotelera (EOH),* elaborada por el Instituto Nacional

[3] Más adelante se detalla cómo se determina *G*.

de Estadística (INE, 2019). Esta encuesta está dirigida a hoteles y acampamentos, sobre la base del marco que proporcionan los directorios de las consejerías de turismo de las comunidades autónomas, cuya actualización se realiza de forma continua. La muestra está diseñada mediante un muestreo estratificado, siendo los estratos la provincia y la categoría hotelera. La investigación es exhaustiva excepto en aquellos estratos con un número grande de establecimientos, para los cuales se selecciona una muestra.

La recogida de la información es mensual, contestando cada establecimiento durante un período de siete días seguidos elegidos de manera que entre todos los establecimientos de cada estrato se cubren todos los días del mes. Dentro del marco conceptual de la *EOH,* se entiende por pernoctación la ocupación por una persona de una plaza o una cama supletoria dentro de una jornada hotelera y en un mismo establecimiento. La ocupación por una persona en el mismo día de dos o más plazas en establecimientos distintos da lugar a más de una pernoctación.

El período muestral utilizado es 1976-2019. De esta manera, se dispone de un conjunto de series temporales de gran longitud (más de 40 años), de forma que es posible realizar un análisis muy completo de su comportamiento cíclico. Esta amplitud temporal tiene un coste, bajo la forma de falta de homogeneidad en la información de base, especialmente en el caso de la sustitución, a partir de 1999, de la *Encuesta sobre Movimientos de Viajeros en Establecimientos Hoteleros* por la actual *Encuesta de Ocupación Hotelera.* Aunque ambas comparten los mismos objetivos y conceptos, la nueva encuesta amplió el ámbito de la investigación y cambió el marco poblacional, introduciendo un cambio de nivel de signo positivo. Análogamente, en 1993 se dejó de encuestar a los establecimientos hoteleros de menor categoría, lo que produjo un efecto escalón de signo contrario.

Afortunadamente, estos cambios metodológicos de tipo permanente son identificados y corregidos mediante el análisis de intervención descrito en la sección 2, haciendo que variables exógenas de tipo cambio de nivel recojan las variaciones metodológicas cuando estas resultan significativas. Por lo que se refiere a las características de las series elegidas que más interesan en este trabajo, destacan la periodicidad mensual y su elaboración a nivel provincial. Ambas características permiten estimar la pauta estacional de cada una de las provincias y su posterior clasificación de acuerdo a la misma. Además, estas series son un buen indicador del ciclo de la actividad turística debido a su cobertura y a su objeto de investigación.

Los datos correspondientes al año 2020, severamente afectados por la crisis sanitaria debida a la COVID-19, no han sido considerados en el estudio porque, dada su naturaleza extraordinariamente atípica y su gran impacto (INE, 2020), requieren un estudio específico que está claramente fuera del objetivo de este trabajo.

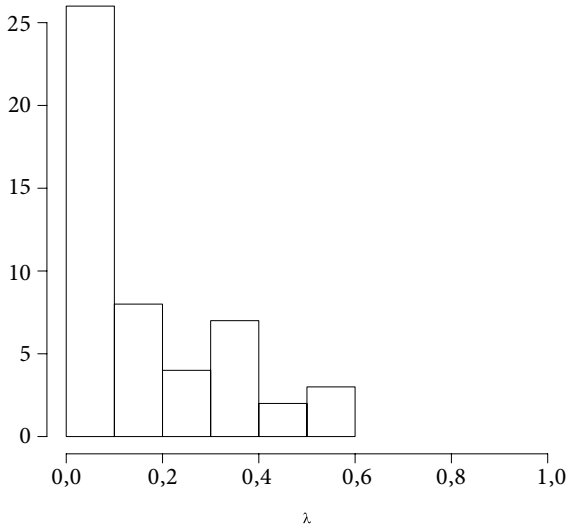
6. RESULTADOS EMPÍRICOS

A continuación se examinan los resultados obtenidos mediante la aplicación de los métodos expuestos en las secciones anteriores a los datos disponibles para este estudio.

La determinación de la transformación Box-Cox más apropiada se ha realizado utilizando el criterio de Guerrero (1993), implementado en el paquete R forecast (Hyndman y Khandakar, 2008). Como puede apreciarse en la figura 1, la distribución del parámetro óptimo λ de dicha transformación está marcadamente concentrado en torno a cero, por lo que se ha aplicado la transformación logarítmica a todas las series.

Figura 1.

Distribución del parámetro Box-Cox óptimo



Fuente: Elaboración propia.

La descomposición de las series log-transformadas, descontados los posibles efectos deterministas vinculados con observaciones anómalas y efectos de calendario, se ha realizado mediante el programa X13-ARIMA-SEATS[4] (U.S. Census Bureau, 2017; Sax y Eddelbuettel, 2018). En la figura 2 se presenta, a título de ejemplo, la descomposición completa de las pernoctaciones registradas en la provincia de Alicante, incluyendo la estimación de una tendencia secular mediante el filtro de Hodrick-Prescott y, como residuo, una señal cíclica.

En este trabajo la estimación de una señal cíclica independiente se obtiene directamente, mediante un filtro de paso en banda de tipo Butterworth (MathWorks, 2013). Este filtro está diseñado para extraer las fluctuaciones comprendidas entre dos y ocho años. La señal cíclica obtenida mediante este filtro muestra un perfil similar al obtenido con el de Hodrick-Prescott pero es mucho menos irregular[5]. Esta mayor suavidad y precisión hace más fiable el cálculo

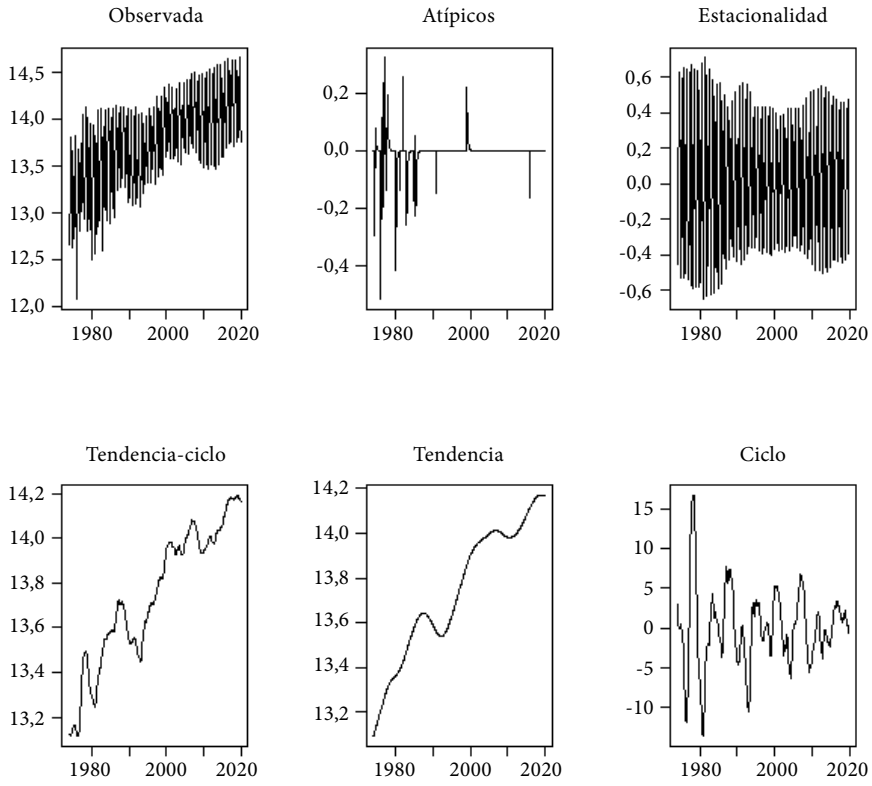
[4] Usando la opción de descomposición basada en modelos ARIMA.

[5] El filtro de Butterworth (paso en banda) elimina la irregularidad pero el de Hodrick-Prescott (paso bajo) la mantiene intacta.

de la distancia entre las series consideradas, especialmente dada la naturaleza no paramétrica y basada en perfiles del método DTW. En la figura 3 se comparan las señales cíclicas del ejemplo considerado (Alicante).

Figura 2.

Descomposición de la serie de pernoctaciones en Alicante

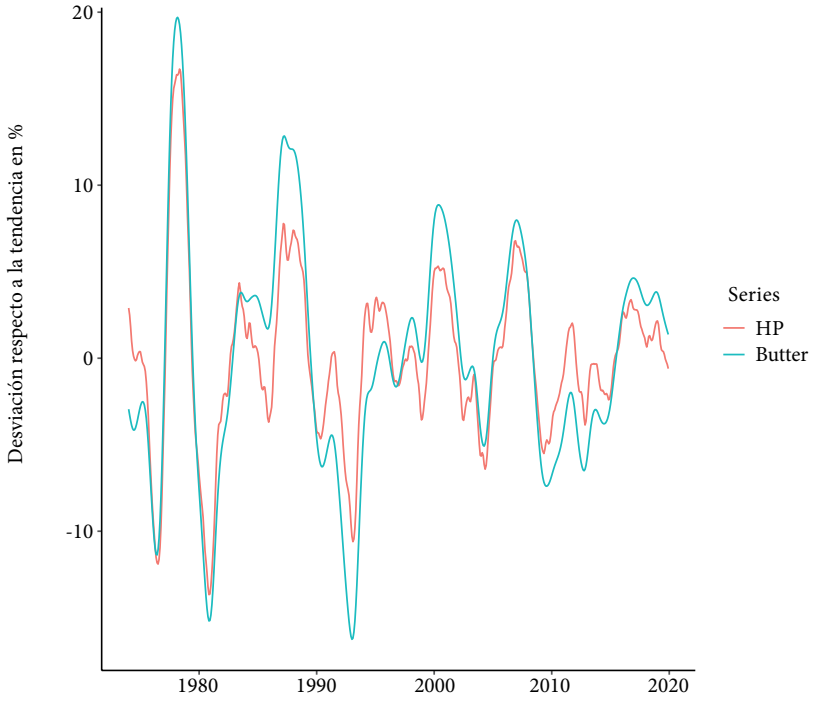


Fuente: Elaboración propia.

La estimación de la matriz de distancia mediante DTW y el análisis de conglomerados se ha realizado mediante diversos paquetes programados en el lenguaje R (R Core Team, 2019), destacando dtwclust (Sardá-Espinosa, 2019), cluster (Maechler *et al.,* 2017), factoextra (Kassambara, 2020) y dendextend (Galili, 2015). Finalmente, la estimación de los modelos factoriales estáticos se ha efectuado con la librería factorLib, programada en Matlab (Quilis, 2019).

Figura 3.

Señal cíclica (Alicante): estimaciones alternativas



Fuente: Elaboración propia.

6.1. Aglomeración estacional

La aglomeración generada por el método de Ward aplicado a la matriz de distancias DTW de los cincuenta factores estacionales puede ser examinada mediante el índice silueta (Rousseeuw, 1987). La figura 4 muestra los índices correspondientes para cada provincia, considerando un número de grupos comprendido entre dos y siete.

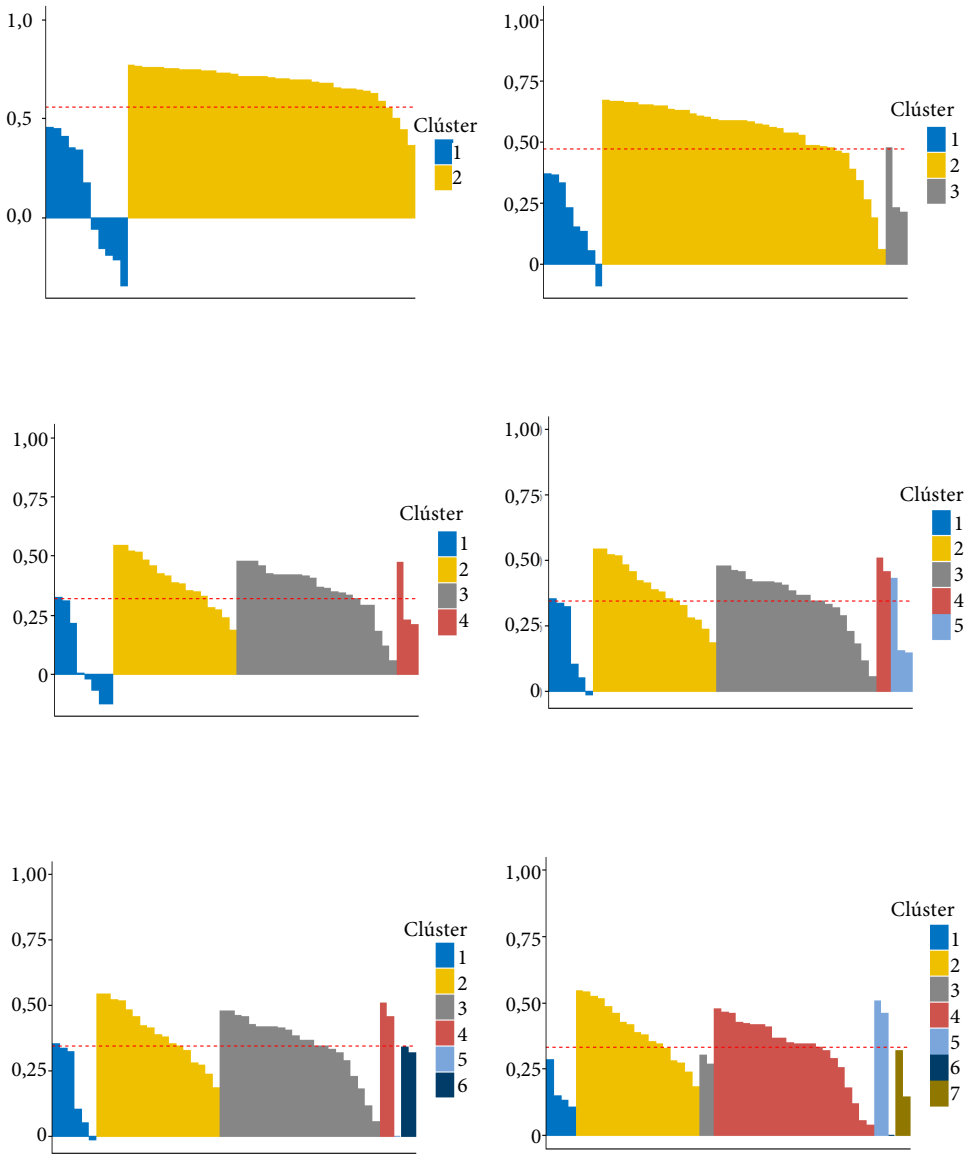
Buscando reducir al máximo posible el número de asignaciones inadecuadas[6], el número de grupos apropiado se sitúa entre cinco y siete. El valor medio del índice es muy similar en los tres casos, por lo que resulta difícil usarlo como criterio de selección.

Con el fin de refinar el análisis, se parte de seis grupos y se comparan las agrupaciones correspondientes con la distribución de las cargas en un modelo con dos factores estáticos. La figura 5 muestra dicha comparación.

[6] Representadas por valores negativos del índice silueta.

Figura 4.

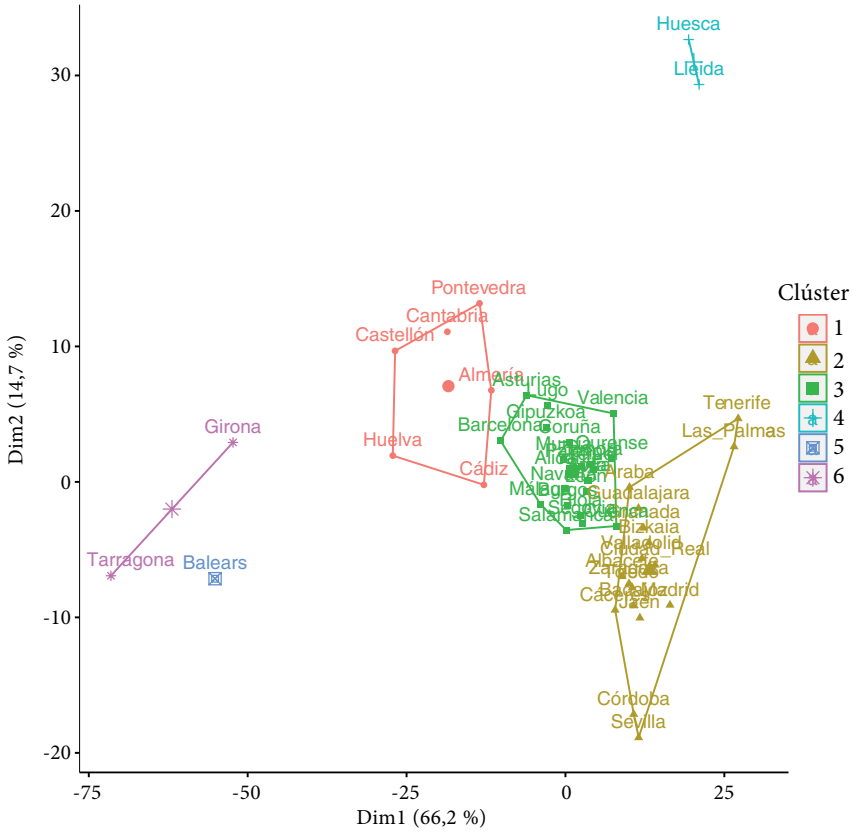
Estacionalidad: agrupación jerárquica (Ward). Índice silueta



Fuente: Elaboración propia.

Figura 5.

Estacionalidad: agrupación jerárquica (Ward) vs. cargas factoriales



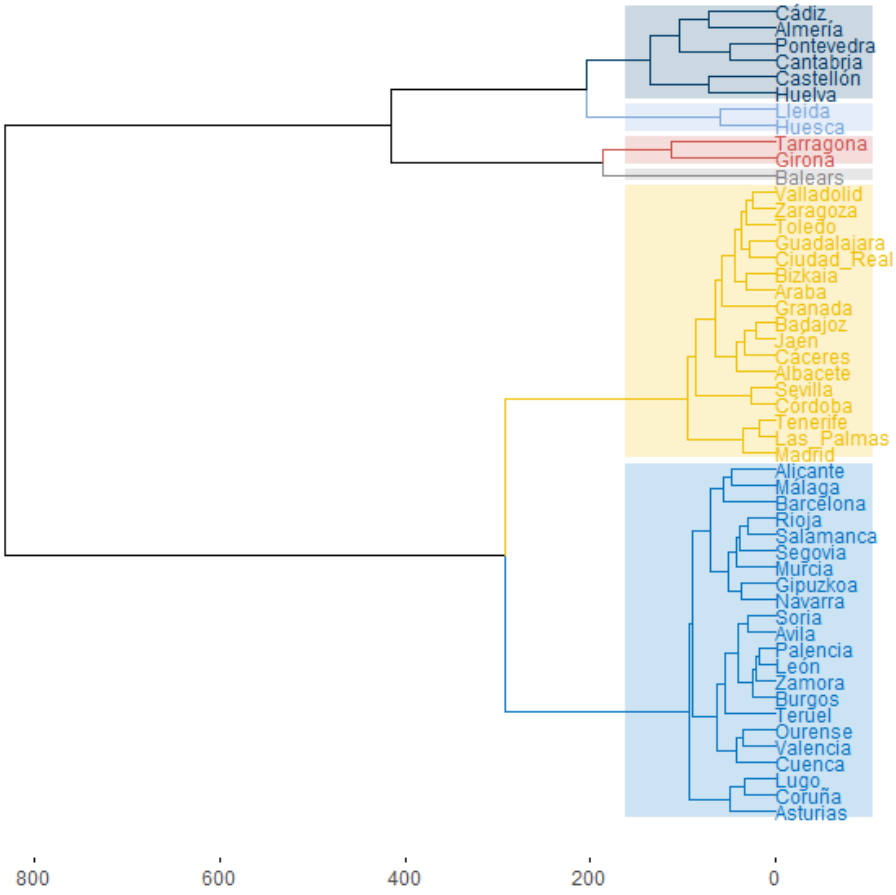
Fuente: Elaboración propia.

En primer lugar, se aprecian dos grupos muy diferenciados y geográficamente afines (Huesca y Lleida, por una parte, y Balears, Girona y Tarragona, por otra). En segundo lugar, aparecen otros tres de mayor tamaño mostrando una mayor continuidad en su aglomeración respecto a los dos factores. Dentro de estos tres grupos destaca el central, muy cohesionado y relativamente poco identificado con el primer factor. El paso de cinco a seis grupos separa Balears y el paso de seis a siete particiona uno de los grupos contiguos en dos. Se ha considerado que esta segunda partición es poco informativa pero no así la primera, debido a tres características especiales de Balears: su insularidad, su peculiar tendencia y su destacado peso[7].

[7] Baleares representa, en promedio, el 23 % del total de pernoctaciones, considerablemente más que las dos siguientes provincias (Las Palmas y Tenerife, con un 9 % y un 8 %, respectivamente).

Figura 6.

Estacionalidad: agrupación jerárquica (Ward). Dendrograma



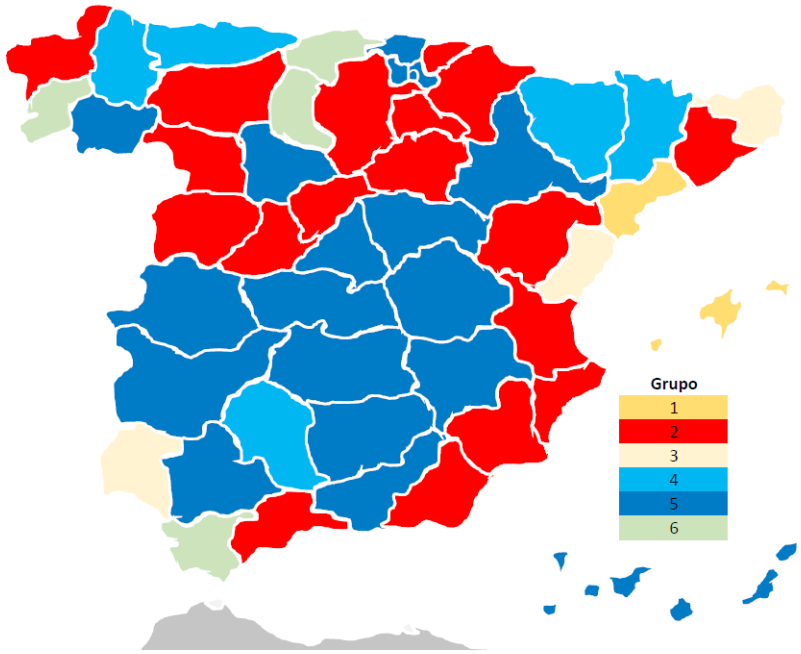
Fuente: Elaboración propia.

Considerando seis grupos, el dendrograma correspondiente ofrece la siguiente aglomeración:

En la figura 6 destacan dos grandes grupos con una distancia de fusión similar y cuatro grupos relativamente pequeños y heterogéneos, a tenor de su mayor distancia de fusión. Dentro de estos grupos, las provincias pirenaicas (Huesca y Lleida) muestran una mayor afinidad.

Figura 7.

Estacionalidad: agrupación particional (k-medias). Distribución geográfica



Fuente: Elaboración propia.

Considerando $G = 6$, el algoritmo de las k-medias descrito en la sección anterior da lugar a la distribución geográfica de los grupos que puede observarse en el mapa de la figura 7.

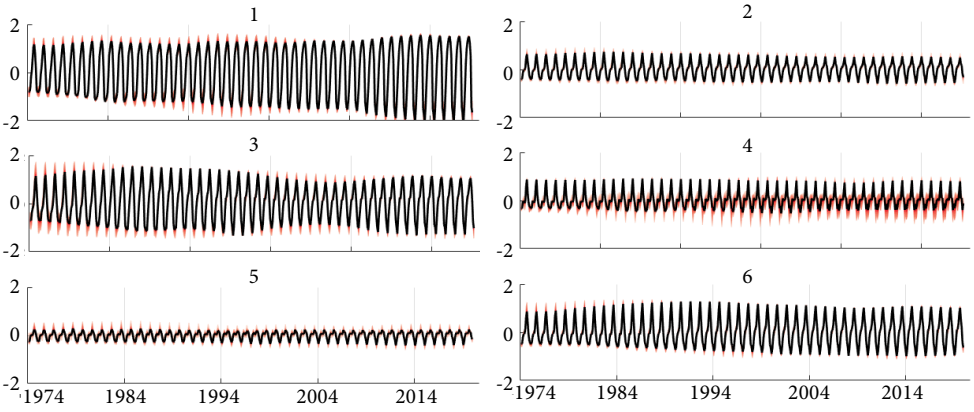
Las agrupaciones generadas son muy similares a las generadas con el método jerárquico de Ward, destacando algunas variaciones: Balears ha sido fusionada con Tarragona en un grupo y las provincias pirenaicas (Huesca y Lleida) han sido agrupadas junto a otras tres relativamente distantes (Asturias, Lugo y Huelva).

Para interpretar adecuadamente la pauta territorial es conveniente examinar los perfiles estacionales de los seis grupos, cuyo promedio temporal aparece en la figura 8.

El grupo 1 (Baleares y Tarragona) se caracteriza por su intenso patrón estacional, que ha ido ganando amplitud con el tiempo y cuyo perfil intraanual muestra una extensa estación alta (desde junio hasta septiembre). El segundo grupo es más numeroso (19 provincias) y su patrón estacional es bastante estable. Este grupo posee un perfil concentrado en el trimestre estival y, especialmente, en el mes de agosto. Geográficamente, abarca casi todas las provincias costeras mediterráneas pero no se identifica exclusivamente con esa característica, ya que

Figura 8.

Factores estacionales agrupados



Fuente: Elaboración propia.

la mayor parte del grupo está formado por provincias interiores, fundamentalmente del cuadrante noroccidental de la península.

El grupo 3 es un grupo pequeño (Girona y Tarragona), geográficamente afín y con una estacionalidad evolutiva. Su perfil intraanual marca un máximo en agosto, en fuerte contraste con los meses valle (diciembre y enero). El cuarto grupo es territorialmente heterogéneo (dos provincias pirenaicas, dos cantábricas y una interior), factor que puede explicar su heterogeneidad transversal. El perfil intraanual muestra el contraste básico entre verano e invierno de los dos grupos anteriores, si bien de forma menos acusada.

El grupo 5, uno de los dos más grandes y relativamente homogéneo internamente, se caracteriza por su reducido rango de variación estacional y su localización geográfica predominantemente interior, con la notable excepción de las dos provincias canarias. Finalmente, el sexto grupo es muy heterogéneo geográficamente. Su pauta estacional es bastante intensa y evolutiva.

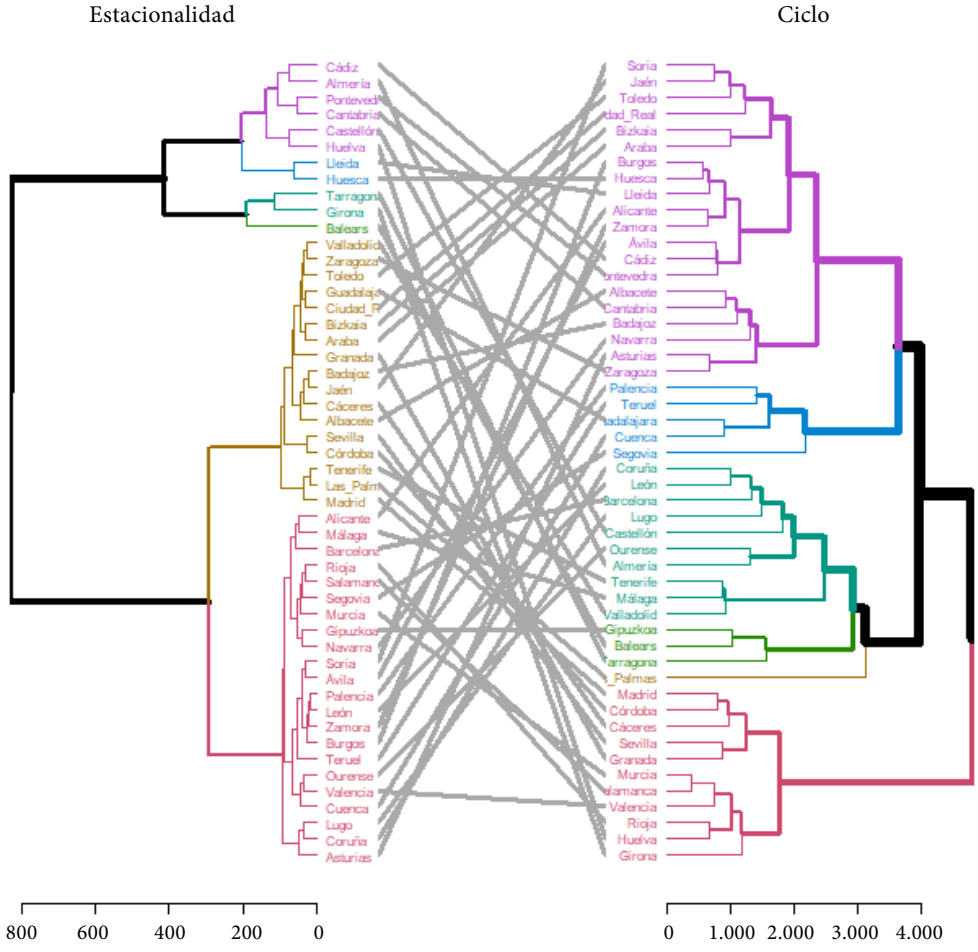
6.2. Conformidad cíclica

Una vez realizada la agrupación de las series provinciales según su pauta estacional, se puede responder a las preguntas planteadas en la introducción, comprobando si esta agrupación se reproduce al considerar su comportamiento cíclico.

En primer lugar, se ha realizado una comparación directa entre los dengrogramas obtenidos mediante el método de Ward aplicado a las matrices de distancia estacional y cíclica. La figura 9 representa ambos grafos y su correspondencia:

Figura 9.

Agrupación jerárquica: correspondencia entre dendrogramas



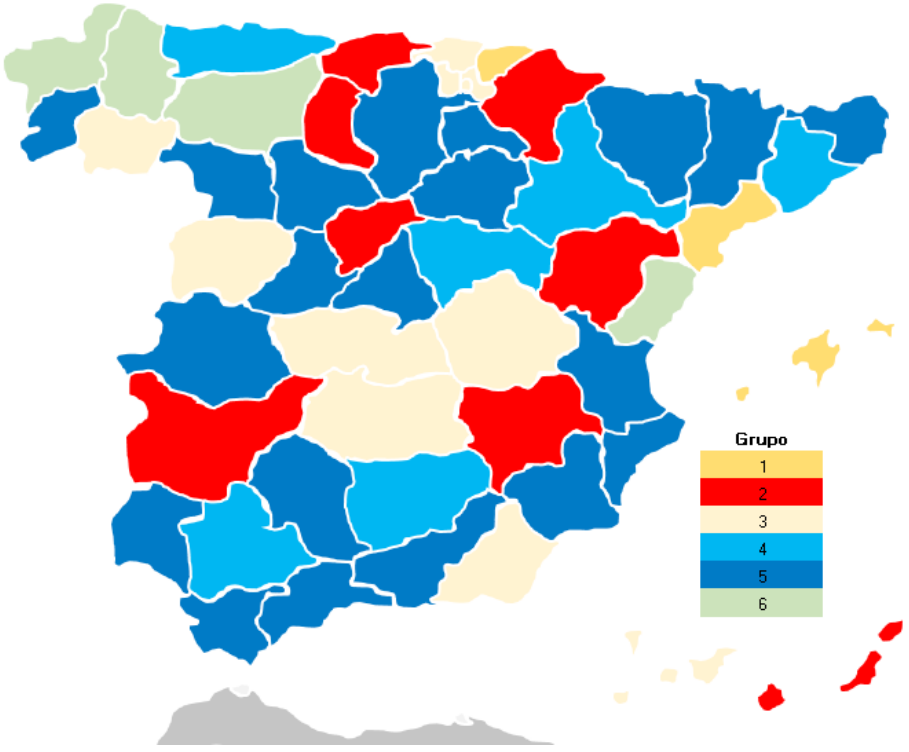
Fuente: Elaboración propia.

La coincidencia entre ambos grafos es moderada, a tenor del valor intermedio del índice de entrelazamiento (*entanglement*): 0,56. Por otra parte, la comparación geográfica entre ambas agrupaciones confirma el resultado anterior, pudiendo apreciarse una menor conexión territorial entre los seis grupos cíclicos así como un solapamiento muy contenido entre ambos. Esta menor conexión se muestra en el mapa de la figura 10.

Otra forma de examinar la conformidad entre ambas agrupaciones consiste en comparar el factor común de las señales cíclicas de cada uno de los seis grupos considerados con el

Figura 10.

Ciclo: agrupación jerárquica (Ward). Distribución geográfica



Fuente: Elaboración propia.

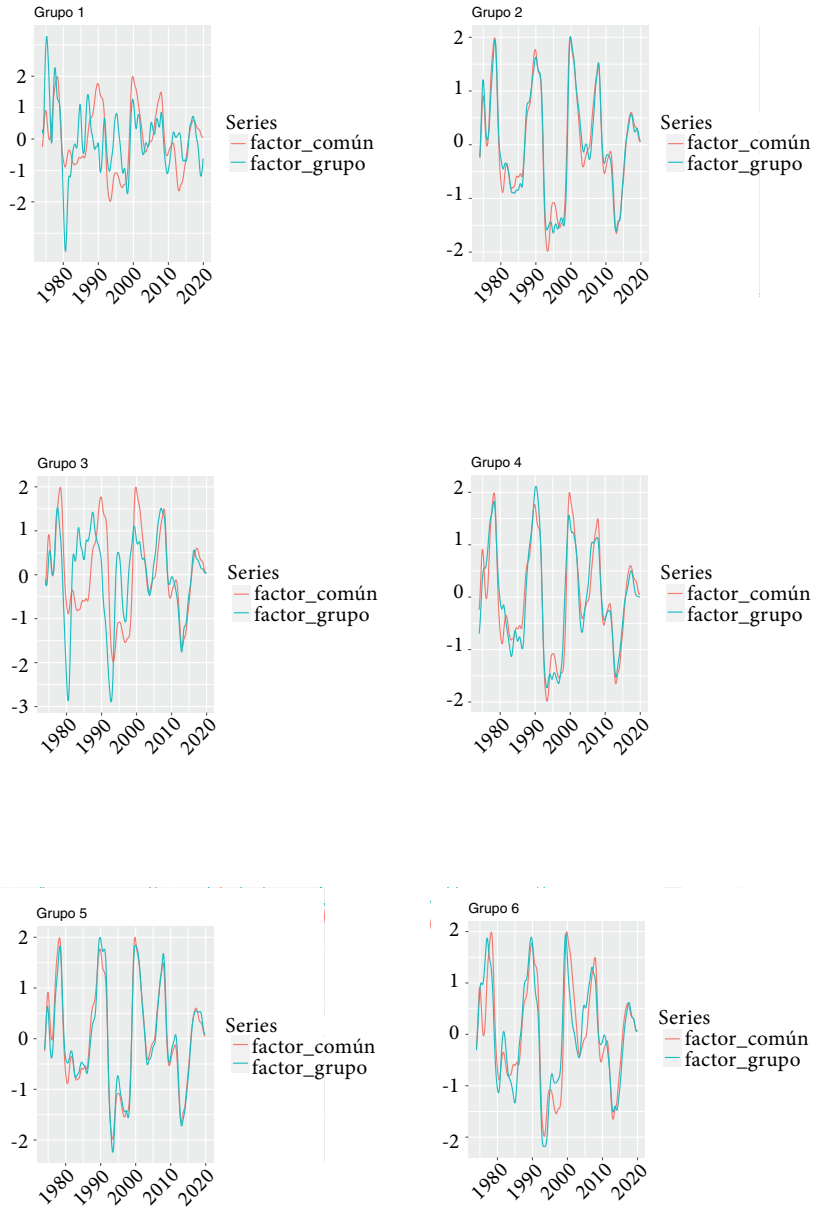
factor común de las cincuenta series provinciales. Los primeros representan la pauta cíclica específica de cada grupo mientras que el segundo sintetiza el ciclo nacional. La figura 11 muestra las series temporales correspondientes.

La comparación entre ambos factores muestra, por lo general, una elevada conformidad. Únicamente las provincias de los grupos 1 (Baleares y Tarragona) y 3 (Girona, Castellón y Huelva) presentan una correlación moderada con el factor nacional, sugiriendo un solapamiento entre las pautas estacionales más específicas de ambos grupos y la idiosincrasia de su comportamiento cíclico. La información proporcionada por las correspondientes funciones de correlación cruzada confirma el diagnóstico gráfico anterior, tal y como puede apreciarse en la figura 12.

En general, predomina una pauta dinámica esencialmente coincidente entre el factor nacional y el específico de cada grupo. Nuevamente, sólo el grupo 3 (Girona, Castellón y Huelva) muestra un cierto adelanto respecto al ciclo común a las cincuenta provincias españolas.

Figura 11.

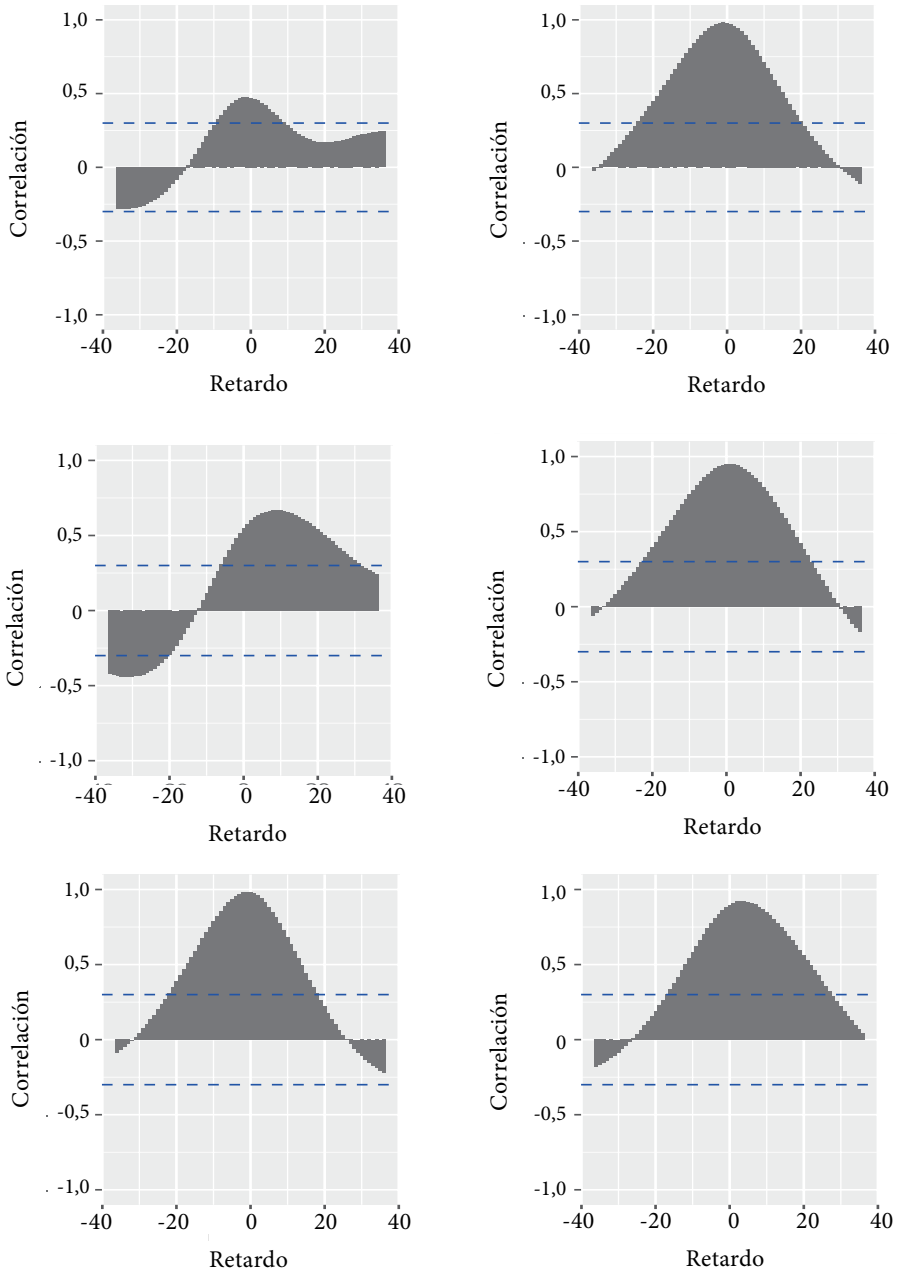
Factores cíclicos específicos vs. factor común



Fuente: Elaboración propia.

Figura 12.

Factores cíclicos específicos vs. factor común



Fuente: Elaboración propia.

7. CONCLUSIONES

El análisis realizado en este trabajo permite identificar una aglomeración de los factores estacionales modulada, principalmente, por la amplitud de sus perfiles intraanuales. Esta aglomeración se concentra en dos grandes grupos que muestran una notable conexión territorial. Se identifican también otros cuatro grupos de menor tamaño y más heterogéneos. En general, la pauta geográfica de los grupos es relativamente compleja, descartándose agrupaciones basadas en rasgos geográficos simples (por ejemplo, costa vs. interior).

La traslación de esta agrupación a los componentes cíclicos muestra un solapamiento muy moderado, de forma que series con comportamientos estacionales diferentes no poseen, por lo general, un patrón cíclico igualmente diferente. Reforzando el resultado anterior, el comportamiento cíclico a nivel provincial muestra un elevado grado de comunalidad, aportando la información estacional un elemento diferenciador marginal. Estos rasgos son bastante robustos frente a la muestra y al cálculo de la matriz de distancia (Frutos y Quilis, 2000).

A nivel teórico, la evidencia empírica obtenida resulta consistente con una visión del proceso de optimización de los agentes económicos en el que la estacionalidad aparece como una restricción exógena, de manera que los agentes la descuentan de forma sistemática a la hora de tomar sus decisiones.

En este sentido, la práctica habitual de analistas del ciclo y coyunturistas, consistente en trabajar con series ajustadas de estacionalidad e ignorando por tanto la pauta estacional, puede considerarse apropiada.

Desde un punto de vista metodológico, la combinación secuencial de técnicas muy diversas ha ofrecido un resultado coherente, permitiendo la aplicación de métodos esencialmente estáticos en un contexto de series temporales múltiples estacionales. En este sentido, el recurso a métodos univariantes de extracción de señales ha sido esencial, al permitir un adecuado tratamiento de la variedad subyacente en los datos y reducir sustancialmente la complejidad del análisis. Estos métodos, muy ligados al ajuste estacional, ofrecen una base sólida y muy consolidada para mitigar la "maldición de la dimension" inherente a la modelización de un vector de series temporales, especialmente si son estacionales.

Por otra parte, el procedimiento utilizado para el cálculo de las matrices de distancia (DTW) posibilita un enlace muy adecuado entre la primera etapa (univariante, basada en modelos dinámicos) y la tercera (multivariante, centrada en técnicas estáticas). El método DTW, al tener en cuenta la naturaleza dinámica de los objetos cuya distancia se desea medir y basarse en la semejanza entre sus perfiles, es especialmente apropiado para el contexto de agrupación de componentes subyacentes de un vector de series temporales. Adicionalmente, el método DTW debidamente modificado puede operar sobre vectores de series de dimensión muy elevada (Rakthanmanon *et al.,* 2012).

La escalabilidad del proceso está asegurada para las dos primeras etapas pero, para la tercera, requiere utilizar como procedimiento de aglomeración uno de tipo no jerárquico. La combinación del método *k-means* con procedimientos de optimización a gran escala (por ejemplo, basados en algoritmos genéticos) es una línea prometedora.

Finalmente, este trabajo puede ser ampliado en diversas direcciones. Combinar la metodología esencialmente exploratoria de este trabajo con un enfoque confirmatorio basado en modelos factoriales dinámicos (Nieto, Peña y Saboyá, 2016) o en modelos estructurales multivariantes (Harvey y Koopman, 1997) es una de ellas.

Otro desarrollo interesante consiste en aplicar la metodología utilizada en este trabajo a series económicas diarias, cuya estructura estacional múltiple plantea importantes retos estadísticos (Cuevas, Ledo y Quilis, 2020).

Referencias

Barsky, R. B. y Miron, J. A. (1989). The seasonal cycle and the business cycle. *Journal of Political Economy,* 97(3), pp. 503–534.

Beaulieu, J., MacKie-Mason, K. y Miron, J. A. (1992). Why do countries and industries with large seasonal cycles also have large business cycles? *Quarterly Journal of Economics,* 107(2), pp. 621–656.

Bógalo, J. y Quilis, E. M. (2000). Estimación del ciclo económico mediante filtros de Butterworth. Instituto Nacional de Estadística. *Boletín Trimestral de Coyuntura,* 87, pp. 73-93.

Box, G. E. P. y Jenkins, G. M. (1976). *Time Series Analysis, forecasting and control.* Holden Day.

Caiado, J., Crato, N. y Peña, D. (2006). A periodogram-based metric for time series classification. *Computational Statistics & Data Analysis,* 50, pp. 2668–2684.

Caiado, J., Crato, N. y Poncela, P. (2020). A fragmented-periodogram approach for clustering big data time series. *Advances in Data Analysis and Classification,* 14, pp. 117–146.

Cecchetti, S., Kashyap, A. y Wilcox, D. (1997). Interaction between the seasonal and business cycles in production and inventories. *American Economic Review,* 87(5), pp. 84–92.

Cuevas. A., Ledo, R. y Quilis, E. M. (2020). Nowcasting the Spanish economy using very high frequency tax data. *SSRN Working Paper.*

Diebold, F. X. (2020). On the origin(s) of the term 'Big Data'. *arXiv Working Paper.*

Everitt, B. S., Landau, S., Leese M. y Stahl D. (2011). *Cluster Analysis.* John Wiley and Sons.

Faber, V. (1994). Clustering and the continuous k-means algorithm. *Los Alamos Science,* 22, pp. 138–144.

Franses, P. H. y de Bruin, P. (2000). Seasonal adjustment and the business cycle in unemployment. *Studies in Nonlinear Dynamics & Econometrics,* 4(2), pp. 1-14.

Frutos, R. y Quilis, E. M. (2000). Estacionalidad y ciclos en las series de pernoctaciones. Instituto Nacional de Estadística. *Boletín Trimestral de Coyuntura,* 76, pp. 65-75.

Galeano, P. y Peña, D. (2000). Multivariate analysis in vector time series. *Resenhas do Instituto de Matematica e Estatstica da Universidade de Sao Paulo,* 4(4), pp. 383-403.

Galili, T. (2015). dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics,* 31(22), pp. 3718–3720.

Geremew, M. y Gourio, F. (2018). Seasonal and business cycles of U.S. employment. Federal Reserve Bank of Chicago. *Economic Perspectives,* 3, pp. 1-28.

Guerrero, V. M. (1993). Time-series analysis supported by power transformations. *Journal of Forecasting,* 12, pp. 37–48.

Harvey, A. C. y Koopman, S. J. (1997). Multivariate structural time series models. En: C. Heij, H. Schumacher, B. Hanzon, y C. Praagman (eds.). *System Dynamics in Economic and Financial Models.* John Wiley and Sons.

Hyndman, R. J. y Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software,* 27(3), pp. 1-22.

INE (2019). *Encuesta de Ocupación Hotelera (EOH).* Metodología.

INE (2020). Un retrato de nuestros turistas. *Cifras INE,* julio.

Kassambara, A. (2017). *Practical Guide to Cluster Analysis* in R. STHDA.com.

Kassambara, A. (2020). *Factoextra R package: easy multivariate data analyses and elegant visualization.* STHDA.com.

Kolanovic, M. y Krishnamachari, R. T. (2017). *Big Data and AI Strategies.* JP Morgan, Global Quantitative & Derivatives Strategy.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. y Hornik, K. (2019). Cluster: Cluster analysis basics and extensions. R package version 2.1.0. CRAN.

Maravall, A. (1987). Descomposición de series temporales. Especificación, estimación e inferencia. *Estadística Española,* 29(114), pp. 11–69.

Montero, P. y Vilar, J. A. (2014). TSclust: an R package for time series clustering. *Journal of Statistical Software,* 62(1), pp. 1–43.

Nieto, F. H., Peña, D. y Saboyá, D. (2016). Seasonality in multivariate time series. *Statistica Sinica,* 26(4), pp. 1389-1410.

Oppenheim, A. V. y Schaffer, R. W. (1989). *Discrete Time Signal Processing.* Prentice Hall.

Peña, D., Tiao, G. C. y Tsay, R. S. (2001). *A Course in Time Series Analysis.* John Wiley and Sons.

Piccolo, D. (1990). A distance measure for classifying ARIMA models. *Journal of Time Series Analysis,* 11(2), pp. 153–164.

Pollock, D .S . G. (1999). *A Handbook of Time Series Analysis, Signal Processing and Dynamics.* Academic Press.

Prescott, E. C. (1986). Theory ahead of business cycle measurement. Federal Reserve Bank of Minneapolis. *Quarterly Review,* 10(4), pp. 9–22.

Proakis, J. G. y Manolakis, D. K. (2006). *Digital Signal Processing.* Pearson New International.

Quilis, E. M. (2019). *FactorLib: a Matlab library for static factor analysis.* Matlab, Central File Exchange.

R Core Team (2019). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

Rakthanmanon, T., Campana, B., Mueen, A., Batista, G., Westover, B., Zhu, Q., Zakaria, J. y Keogh, E. (2012). Searching and mining trillions of time series subsequences under Dynamic Time Warping (DTW). SIGKDD, pp. 262–270.

Rani, S. y Sikka, G. (2012). Recent techniques for clustering of time series data: a survey. *International Journal of Computer Applications,* 52(15), pp. 1–9.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics,* 20, pp. 53–65.

Sakoe, H. y Chiba, S. (1978), Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, IEEE Transactions on Speech and Signal Processing,* 26(1), pp. 43–49.

Sardá-Espinosa, A. (2019) Time series Clustering in R using the dtwclust package. *The R Journal,* 11(1), pp. 22–43.

Sathi, A. (2012). *Big Data Analytics.* MC Press.

Sax, C. y Eddelbuettel, D. (2018). Seasonal adjustment by X-13ARIMA-SEATS in R. *Journal of Statistical Software,* 87(11), pp. 1-17.

Todd, R. M. (1990). Periodic linear-quadratic methods for modeling seasonality. *Journal of Economic Dynamics and Control,* 14(3–4), pp. 763–795.

U.S. Census Bureau, Time Series Research Staff (2017). X-13ARIMA-SEATS Reference Manual. U.S. Census Bureau.

Wang, X., Smith, K. y Hyndman, R. (2006). Characteristic-based clustering for time series data. *Data Mining and Knowledge Discovery,* 13, pp. 335-364.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association,* 58(301), pp. 236-244.