
*Máster en Modelización Matemática, Estadística y Computación.*

# Técnicas estadísticas multivariantes de series temporales para la validación de un sistema reconstructor basado en redes neuronales.

Autor: Sergio Luis Suárez Gómez

Directora: Ana Carmen Cebrián Guajardo

Curso 2015-2016

# Índice general

# Capítulo 1

# Introducción

## 1.1. Presentación del problema

La atmósfera terrestre se encuentra en un continuo cambio; se forman turbulencias en la misma debido a las condiciones climatológicas, el viento, su fuerza y dirección, los cúmulos nubosos, la densidad de los componentes que se encuentran en la atmósfera, la presión, la temperatura, la altura de cada una de las distintas posibles capas de turbulencias y su capacidad de mezcla y cambio repentino, son algunos de los factores que influyen en la naturaleza caótica y cambiante de la atmósfera.

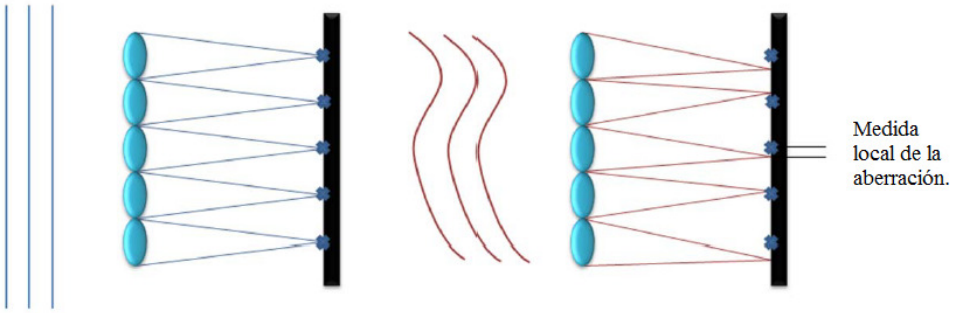
A pesar de las necesarias ventajas que proporciona la existencia de la atmósfera, supone un obstáculo caprichoso para la toma de imágenes astronómicas. Debido a la naturaleza ondulatoria de la luz, cuando se emite de un objeto astronómico lejano (a una distancia infinita, a efectos prácticos), la luz llega a la superficie terrestre como un frente de onda plano. Por causa de las turbulencias presentes en la atmósfera, que provocan un cambio del índice de refracción y con el consecuente cambio en la fase de la luz que atraviesa la atmósfera, los fotones que conforman dicha luz no llegan de manera uniforme a la superficie, provocando que el frente de onda presente irregularidades y deje

de ser plano. Por tanto la imagen que se recoge, ya sea a simple vista, o mediante un instrumento como puede ser un telescopio, presenta una distorsión.

La complejidad de los distintos modelos de las turbulencias, que es debida al número de variables que influyen en cada perfil atmosférico, y la velocidad de cambio de las condiciones de cada turbulencia (en ocasiones, algunos cambios son significativos con variaciones temporales de milisegundos), hacen imposible en la práctica encontrar la expresión analítica de un modelo que represente con exactitud las variaciones que experimenta la luz al atravesar la atmósfera.

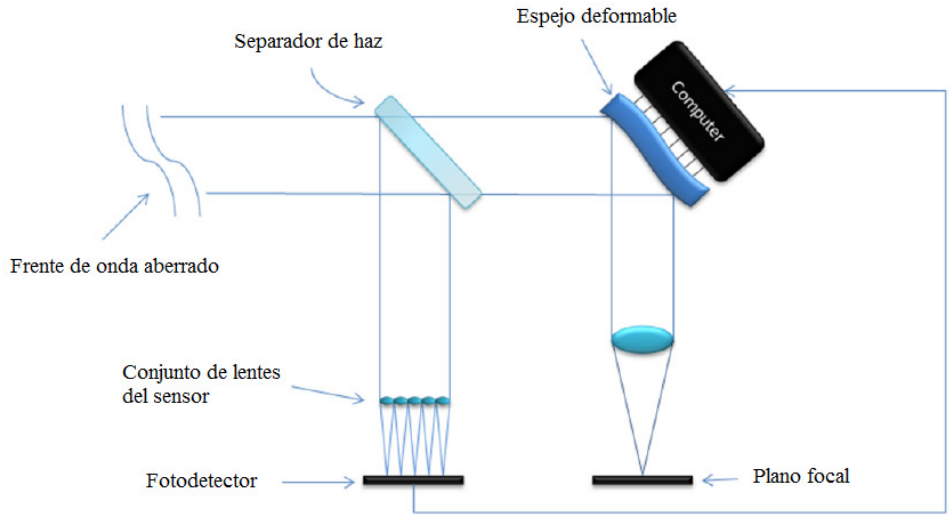
Para solucionar éste problema, aparece el campo de la óptica adaptativa, cuyo fin último es intentar compensar, en tiempo real, las distorsiones que sufren las señales que atraviesan la atmósfera. La óptica adaptativa que se ha ido desarrollando en los últimos años afronta el problema de la siguiente forma: por un lado, para medir las turbulencias en la luz, se utilizan sensores de frente de onda, en concreto el más utilizado es el sensor Shark-Hartmann [de Cos Juez et al., 2012], cuya superficie está dividida en un conjunto de lentes y asociadas a cada una, tiene un fotodetector. De ésta forma, el frente de onda de la luz incidente se divide en zonas, una asociada a cada lente, y por tanto se puede medir la inclinación en cada zona comprobando la desviación del enfocado con el fotodetector (esquema 1).

Esquema 1: Funcionamiento de un Shark-Hartmann



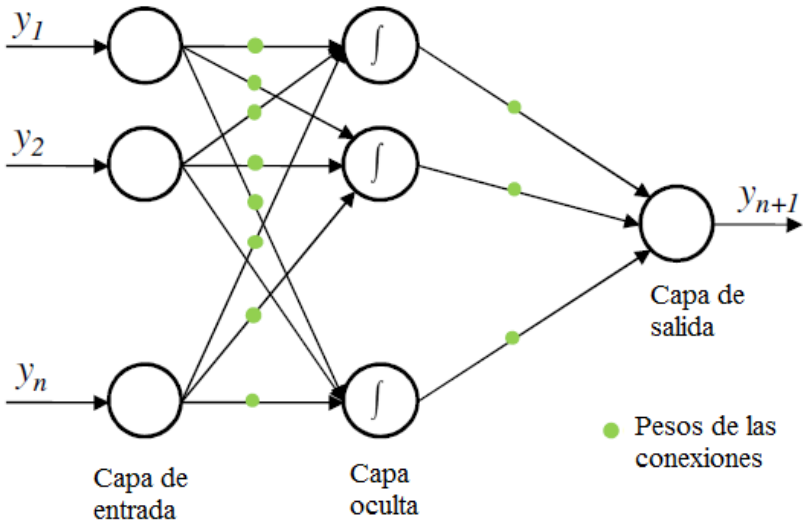

Las medidas recibidas por el sensor de frente de onda son utilizadas para, mediante el cálculo numérico de un reconstructor por ordenador, dar una señal a un espejo deformable que modificaría las fases de la luz de forma local corrigiendo así la imagen obtenida (esquema 2). Uno de los problemas en los que se está trabajando actualmente es encontrar un reconstructor que sea robusto frente a variaciones en las condiciones de la atmósfera y que el tiempo requerido para su cálculo sea lo más cercano a realizar los cálculos en tiempo real.

Esquema 2: sistema de óptica adaptativa



Una de las herramientas que dispone la óptica adaptativa para tratar éste problema son las redes neuronales. Las redes neuronales son modelos que son útiles en el campo de la predicción de sistemas complejos que se basan en un gran número de elementos procesadores (neuronas) interconectados. Las neuronas están agrupadas en capas y conectadas a las neuronas de capas adyacentes por medio de coeficientes variables (pesos) que representan la influencia de cada neurona en la siguiente capa. Los pesos se ajustan por diversos métodos de retropropagar el error respecto del patrón que se pretende que aprenda o modele. La señal de salida que se obtiene de una red neuronal es la respuesta que ofrece la red ante el patrón que se aplica en las señales de entrada en función del sistema que la red esté modelando (esquema 3).

Esquema 3: estructura de una red neuronal




Una red neuronal puede ser entrenada de forma que aprenda a procesar datos procedentes de una turbulencia y que proporcione una salida con las correcciones que haya que aplicar. Éste método ha sido utilizado en [Osborn et al., 2012], [Osborn et al., 2014], [de Cos Juez et al., 2012].

Éste es el punto de partida del trabajo: se dispone de un vector de 1152 observaciones del sensor, medido cada dos milisegundos, que conforman las señales de entrada del sistema, formado por una red neuronal en continuo reentrenamiento, y un vector de medidas en el tiempo correspondientes a los valores proporcionados por la red neuronal en cada instante, que forman las señales de salida del sistema que se utilizan para corregir las imágenes.

Se pretende comprobar que el procedimiento realizado es válido, dado que, aunque se conoce en profundidad el correcto funcionamiento de las redes neuronales, se desconoce si utilizar redes de la misma topología pero que van actualizando su entrenamiento en un proceso continuo produce resultados coherentes. Se quiere comprobar qué ocurre con la estructura temporal de las salidas, es decir, si el sistema es capaz de mantener la estructura temporal de las señales de entrada en las señales de salida; para ello se pretende analizar los datos proporcionados por el sistema con la ayuda de modelos VARMA y de técnicas de clustering de series temporales.

## 1.2. Presentación de los datos

*Datos de entrada.*

Los datos son tomados por un sensor Shark-Hartmann. Se corresponden con un sensor de 144 fotodetectores que dan valores de las posiciones en $x$ y en $y$ del centro del punto enfocado (centroide), correspondientes a cuatro estrellas guía láser [1], dando un total de 1152 variables por cada instante de tiempo. Se realizaron medidas cada 2 milisegundos hasta tener un total de 2500 medidas.

Se puede considerar, por tanto, que se dispone de un vector de 1152 series temporales, de las cuales se tienen 2500 observaciones. Las 1152 series se agrupan en 4 vectores de 288 series cada uno, que se corresponden a las entradas de cada una de las estrellas guía. Éstas series temporales muy posiblemente estén relacionadas entre ellas y no serán independientes, por lo que un modelo multivariante será más adecuado que los modelos univariantes separados. Al igual que ocurre en el caso de los datos tratados aquí, ésto es habitual cuando los datos provienen de la observación de procesos económicos, físicos, etc.

*Datos de salida.*

Respecto de la señal de salida, está constituida por un vector de 288 series, de las que de forma análoga al caso anterior, se han obtenido valores correspondientes al estado del modelo cada dos milisegundos.

[1] En muchas ocasiones para la toma de imágenes se utiliza como referencia una estrella artificial creada por láser cuando no se puede disponer de una estrella natural [de Cos Juez et al., 2012].

# Capítulo 2

# Modelos univariantes

## 2.1. Análisis univariante de series temporales

En ésta sección se introducirán los procesos que son utilizados en el modelado de series temporales univariantes y alguna de sus propiedades más características. Los procesos de series temporales ARMA, ARIMA y SARIMA son los más básicos que permiten entender y modelar el comportamiento temporal de grandes conjuntos de datos observados en el tiempo.

Para empezar, los modelos ARMA definen una familia paramétrica de procesos estacionarios, conocidos como procesos autorregresivos de media móvil (*autoregressive moving average*). Éstos modelos son definidos como sigue:

El proceso $\{X_t, t = 0 \pm 1, \pm 2, \dots\}$ se dice que es un proceso de tipo ARMA$(p, q)$ si $\{X_t\}$ es estacionario y si para todo $t$,

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \qquad (2.1)$$

donde al proceso $\{Z_t\}$ se le llama *white noise* (proceso de ruido blanco) de observaciones incorreladas de media cero y varianza $\sigma^2$, escrito de la forma $\{Z_t\} \sim WN(0, \sigma^2)$.

La ecuación (1.1) se puede escribir de forma más compacta como:

$$\phi(B)X_t = \theta(B)Z_t, \qquad t = 0, \pm 1, \pm 2, \ldots \tag{2.2}$$

donde $\phi$ y $\theta$ son los polinomios de grados $p$ y $q$ respectivamente:

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \tag{2.3}$$

y

$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q \tag{2.4}$$

siendo $B$ el operador despazamiento hacia atrás (*backward shift operator*) definido por

$$B^j X_t = X_{t-j}, \qquad j = 0, \pm 1, \pm 2, \ldots \tag{2.5}$$

Los procesos conocidos como ARIMA son generalizaciones de los modelos ARMA, a los que se incorporan la posibilidad de ser series no estacionarias por tener una componente de tendencia. Si un proceso ARIMA es diferenciado un número finito de veces (adecuado para eliminar la tendencia), el modelo se reduce a un proceso ARMA. Éstos modelos se definen de la siguiente forma: si $d$ es un entero no negativo, entonces $\{X_t\}$ se dice que es un proceso ARIMA$(p, d, q)$ si el proceso satisface la ecuación:

$$\phi(B)(1 - B)^d X_t = \theta(B)Z_t \tag{2.6}$$

donde $\phi(z)$, $\theta(z)$ son polinomios de grados $p$ y $q$ respectivamente con $\phi(z) \neq 0$ para $|z| \leq 1$ y $\{Z_t\} \sim WN(0, \sigma^2)$.

Los modelos conocidos como SARIMA (seasonal ARIMA), permiten aleatoriedad en el patrón estacionario de un ciclo al siguiente. Por ser un proceso derivado de un modelo ARIMA, también mantiene la posibilidad de diferenciación. Éstos modelos son definidos como sigue:

Si $d$ y $D$ son enteros no negativos, entonces $\{X_t\}$ se dice que es un proceso seasonal ARIMA$(p,d,q) \times (P,D,Q)_s$ con periodo $s$ si el proceso diferenciado $Y_t := (1-B)^d(1-B^s)^D X_t$ es un proceso ARMA tal que:

$$\phi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)Z_t \tag{2.7}$$

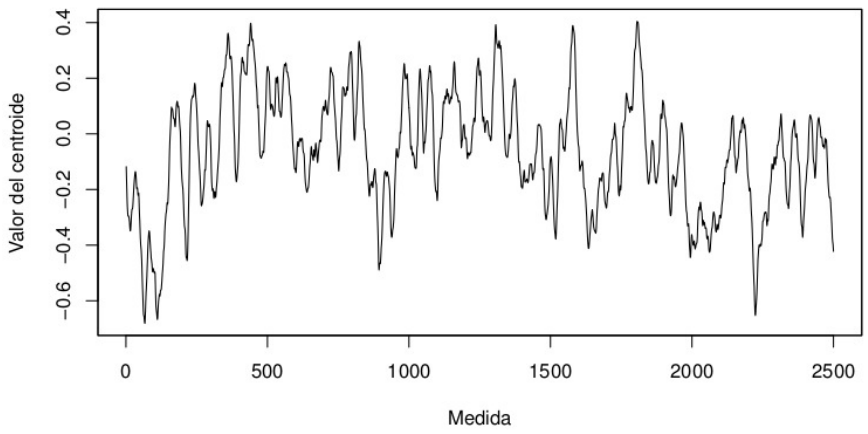
Donde $\{Z_t\} \sim WN(0,\sigma^2)$, y los polinomios son definidos como: $\phi(z) = 1-\phi_1 z-\cdots-\phi_p z^p$, $\Phi(z) = 1-\Phi_1 z-\cdots-\Phi_P z^P$, $\theta(z) = 1+\theta_1 z+\cdots+\theta_q z^q$ y $\Theta(z) = 1+\Theta_1 z+\cdots+\Theta_Q z^Q$.

## 2.2. Análisis univariante de los datos.

*Series de entrada.*

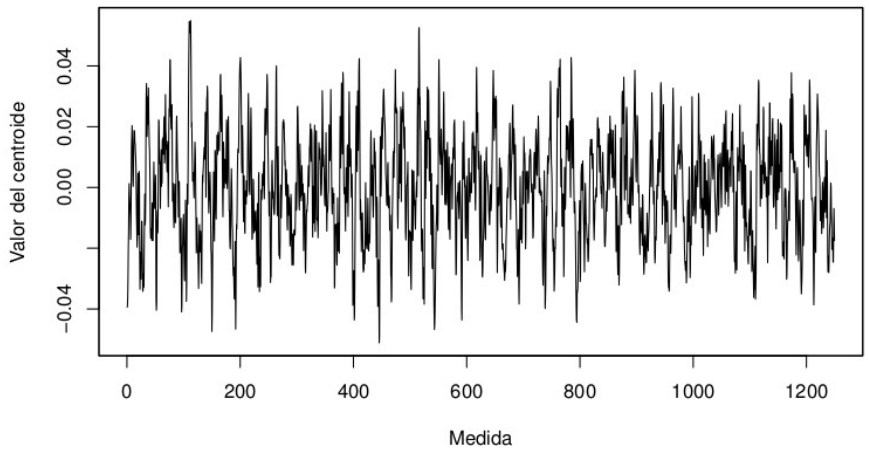
Un análisis univariante exploratorio de algunas de las series da una idea de las características de los datos con los que se va a trabajar. Por ejemplo, el rango de valores y la evolución temporal de los datos de la primera serie temporal del vector se puede observar en la Figura 1.

**Figura 1: componente x del primer sensor.**



El test KPSS (Kwiatkowski-Phillips-Schmidt-Shin), permite contrastar si una serie es estacionaria (hipótesis nula). En éste caso se obtiene un p-valor de 0.01 y por tanto la serie no es estacionaria, pero su serie diferenciada (Figura 2) sí lo es.

**Figura 2: serie diferenciada decomponente x del primer sensor.**




La existencia de correlación se puede comprobar utilizando herramientas gráficas como pueden ser los correlogramas de la función de autocorrelación (ACF) o de la función de autocorrelación parcial (PACF). Véase la figura 3 y 4.

**Figura 3: ACF de los datos diferenciados.**

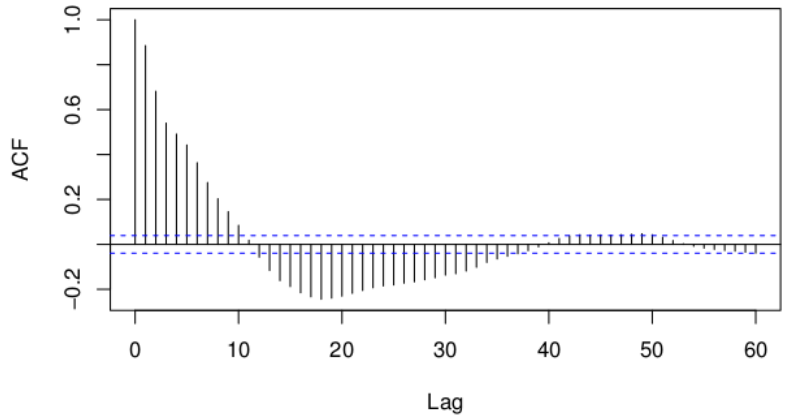
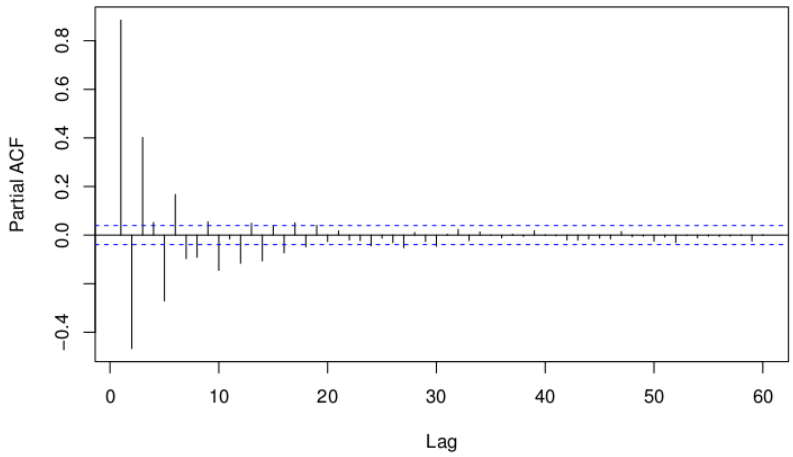



**Figura 4: Partial ACF de los datos diferenciados.**

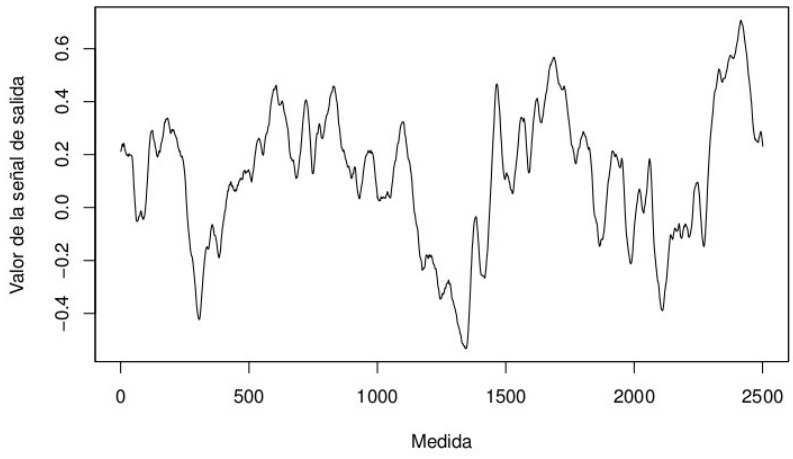



Para confirmar la existencia de correlación serial, se ha utilizado el test de Box-Pierce cuya hipótesis nula es que la correlación de orden 1 es nula. En todas las series se han obtenido p-valores menores que 0.05, lo que indica que en todas las series existe correlación serial.

*Series de salida.*

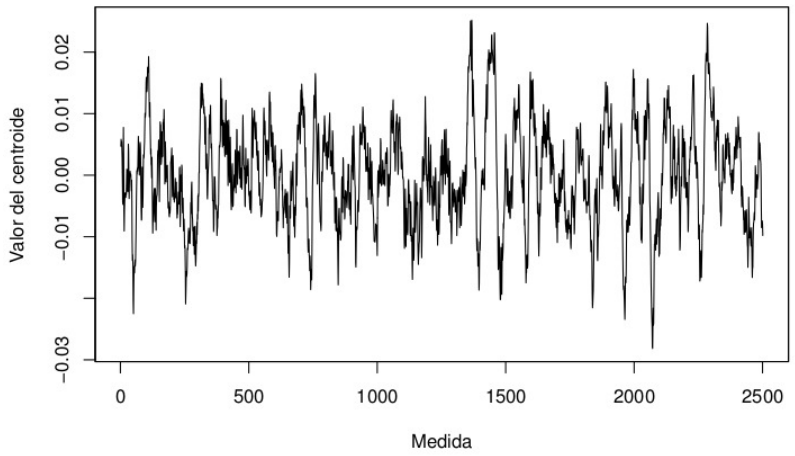
Como ejemplo de una serie de salida, en la figura 5 se muestra el rango de valores y la variación temporal de la serie univariante de la primera componente de la señal de salida.

**Figura 5: primera componente de la salida.**




Utilizando el test KPSS, se tiene como resultado un p-valor de 0.0163 y por tanto la serie no presenta un comportamiento estacionario, y así, es necesario realizar una diferenciación de los datos para tener una serie estacionaria como se puede observar en la figura 6.

**Figura 6: primera componente de la serie diferenciada de las salidas.**




Para entender mejor éstas series, se puede comprobar la correlación de los datos que las conforman. Para ello, nuevamente se utilizan herramientas gráficas que permitan observar la función de autocorrelación (ACF) o la función de autocorrelación parcial (PACF). Véase la figura 7 y 8.

**Figura 7: ACF de los datos diferenciados de las salidas.**

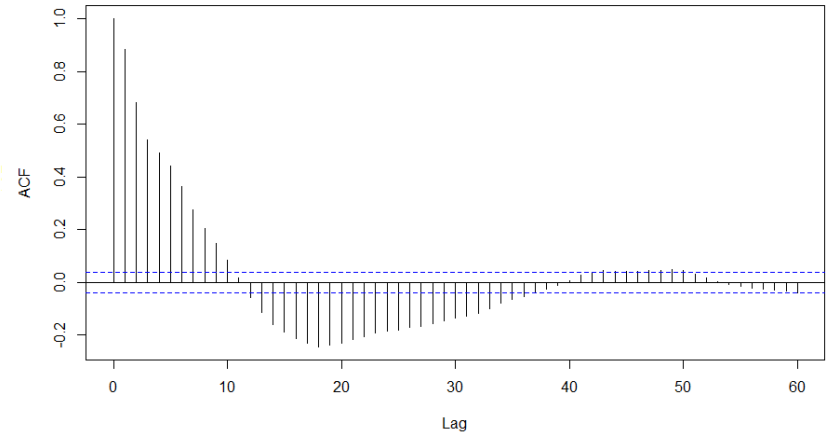
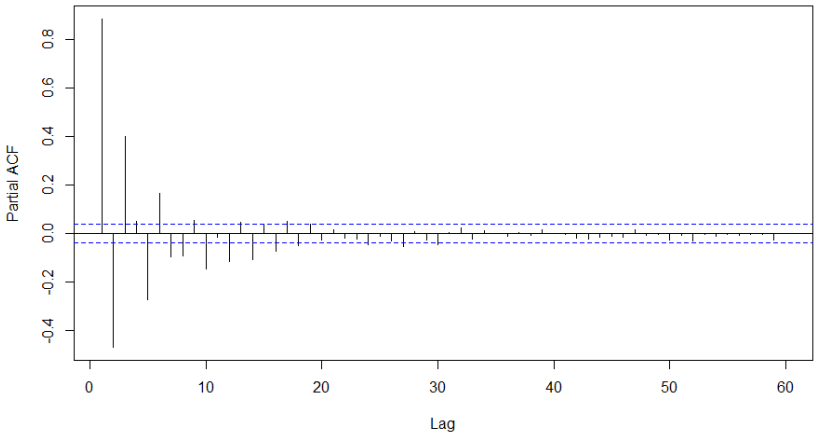


**Figura 8: Partial ACF de los datos diferenciados de las salidas.**




Al igual que con los datos de entrada, el test de Box-Pierce indica la existencia de correlación serial.

# Capítulo 3

# Modelos TSFA y VARMA

Debido a la necesidad de estudiar las series agrupadas en vectores, ya que no se puede desestimar la dependencia que existe entre ellas, se debe considerar un marco de modelización multivariante.

Dentro de las opciones de modelos multivariantes existentes, se han descartado, por motivos de complejidad de modelo y complejidad computacional, algunos tipos de modelos como ARMAX (auto-regressive moving average with exogenous inputs), modelos ARMA con variables exógenas que pueden modelar las dependencias entre las series y factores externos que afecten al comportamiento de las mismas, y sus generalizaciones vectoriales VARMAX.

Dado que el volumen de series para el que están pensados éstos modelos es bajo (habitualmente menos de 10 en la literatura), se ha optado por los procesos VARMA, la generalización directa a una dimensión mayor de procesos tipo ARMA, y por los modelos basados en factores, que permiten reducir el número de variables a tener en cuenta basándose en sus características principales.

En el caso de los datos considerados, parece razonable comparar los modelos que ajusten el comportamiento de los datos procedentes de las estrellas y de la salida de la red neuronal por separado, para buscar sus similitudes y diferencias, dando lugar a cinco conjuntos de datos: los cuatro primeros, uno por cada estrella guía, cuyas estructuras temporales sería razonable que se comportasen de manera similar, y el quinto correspondiente a las salidas del sistema de redes neuronales.

Ésto supondría hacer cinco modelos VARMA de vectores de 288 series, lo cual no es realizable, dado que en la práctica, los modelos tan grandes de éste tipo son poco manejables, además de poco precisos, y requieren un tiempo de computación inviable.

Como alternativa, se puede utilizar la información obtenida de un modelo factorial de series temporales (TSFA), es decir, en vez de modelar directamente las series de cada conjunto de datos, se procederá a reducir la dimensión de cada vector, expresándolos en términos de un número menor de factores. Ésto supone una reducción de número de variables a considerar y sigue permitiendo realizar una comparación entre las características más importantes de las series.

Para la realización práctica de éste estudio y su comparación, se utilizarán distintos paquetes del software R, que se detallan a lo largo del capítulo.

## 3.1. Análisis factorial para series temporales

El análisis factorial es una técnica estadística de reducción de datos cuyo objetivo es explicar las correlaciones entre las variables observadas en términos de un número menor de variables no observadas, que se denominan factores. En concreto se busca representar las variables observadas como combinaciones lineales de dichos factores más un cierto término de error.

Cuando los datos observados tienen una estructura temporal, es decir, son series temporales, el análisis factorial estándar no es adecuado ya que requiere la hipótesis de observaciones incorreladas en cada variable. Para solventar éstos inconvenientes se desarrollaron los modelos TSFA (Time Series Factor Analysis), cuyo objetivo es resumir la mayor parte de información de un número elevado de series temporales en un número menor de factores subyacentes (variables no observadas).

Una de las mayores fortalezas de la estimación de modelos TSFA reside en que no requiere de hipótesis demasiado restrictivas, por ejemplo, no requiere establecer a priori un número determinado de factores, pudiéndose hacer una estimación del número más adecuado de los mismos.

Un modelo TSFA se especifica de la forma siguiente:

Los $k$ procesos no observados de interés (denominados *factores*) para una secuencia de $T$ periodos temporales se denotrán por $\xi_{it}$, con $t = 1, \ldots, T$, siendo $\xi_t = (\xi_{1t}, \ldots, \xi_{kt})$. Los procesos observados (las series temporales, que se denominarán *indicadores*) serán denotados por $y_t = (y_{1t}, \ldots, y_{Mt})$ donde $t = 1, \ldots T$. Los indicadores y los factores correspondientes a un periodo $t$ son recogidos en los vectores columna $y_t$ y $\xi_t$ respectivamente. El modelo que

relaciona los indicadores con los factores vendrá dado por

$$y_t = B\xi_t + \varepsilon_t \tag{3.1}$$

con $B$ la matriz de tamaño $M \times k$ de pesos del modelo (*factor loadings*) a ser estimados y $\varepsilon_t$ un vector aleatorio de tamaño $M$ en el que se recogen los errores, perturbaciones, o factores poco relevantes (por ejemplo, no comunes a todas las series).

En esencia, la estimación de los pesos se puede llevar a cabo con cualquiera de los métodos de estimación de uso habiltual en modelos FA (en concreto, máxima verosimilitud) que permitirán, además de obtener los pesos B, la matriz de covarianza de los factores, denotada por $\Phi$ y la matriz de covarianza de los errores, denotada por $\Omega$. Los valores en el tiempo de los factores $\xi$, denominados *puntuaciones* (*factor scores*), se pueden calcular utilizando el estimador *predictor de Bartlett* (*Bartlett predictor*) el cual se puede expresar como

$$\hat{\xi_t^B} = (\hat{B}'\hat{\Omega}^{-1}\hat{B})^{-1}\hat{B}'\hat{\Omega}^{-1}y_t \tag{3.2}$$

La utilización de éste estimador tiene la ventaja de que en su expresión todos los términos se obtienen como resultado de las estimaciones habituales en los modelos FA.

La contrucción de los modelos TSFA y su estimación puede encontrarse de forma detallada en las referencias [Gilbert & Meijer, 2005] y [Forni et al. , 2005].

## 3.2. Análisis utilizando modelos TSFA.

### 3.2.1. Paquete 'TSFA' del software estadístico R

El paquete 'TSFA' extiende el análisis de factores estándar (FA) para poder ser utilizado con datos provenientes de series temporales.

Las funciones principales son :

- *estTSFmodel()* ajusta un modelo utilizando un estimador de modelos FA basado en la matriz de correlaciones. Utiliza el estimador de Bartlett para calcular los factor scores. Ésta función devuelte un objeto de la clase TSFmodel en el que están almacenados el modelo estimado y los datos utilizados.

- *DstandardizedLoadings()* tiene como argumento de entrada un objeto de la clase TSFmodel del cual extrae y devuelve como salida los valores de los pesos estandarizados.

- *factors()* tiene como argumento de entrada un objeto de la clase TSFmodel del cual extrae y devuelve como salida las puntuaciones del factor, es decir, devuelve valores en el tiempo de cada factor.

- *tfplot()* es una herramienta gráfica que permite reprensentar los valores de los factor scores extraidos de la función factors().

Otras funciones interesantes que se implentan dentro de éste paquete permiten, por ejemplo, hacer la estimación de las puntuaciones con predictores distintos a los de Bartlett. También incluye funciones que permiten hacer la

simulación de los datos originales de las series temporales (los indicadores), para un modelo factorial especificado.

### 3.2.2. Resultados de modelos TSFA

Para poder plantear un modelo TSFA, primero se debe hacer una selección adecuada del número de factores a extraer. Para ello, se dispone de varios criterios, como el criterio AIC (criterio de información de Akaike), otros basados en los valores propios, etc.

Uno de los criterios más extendidos, ver [Gilbert & Meijer, 2005], consiste en escoger tantos factores como valores propios, correspondientes a la matriz de autocorrelación de las series, mayores que 1 se hayan calculado.

Los valores propios correspondientes al grupo de series de la primera estrella guía, se muestran como ejemplo en el siguiente gráfico (figura 9):

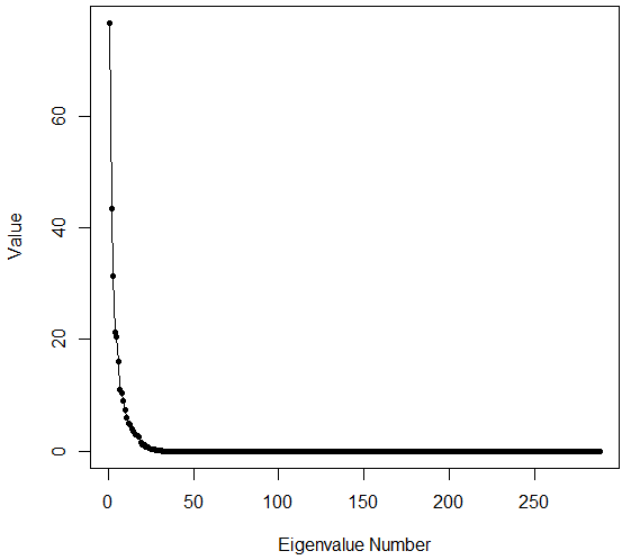



Figura 9.

Para el resto de grupos de series, las gráficas de valores propios son similares. Debido al elevado número de valores propios próximos a uno, se ha optado por escoger sólo los estrictamente mayores que 1, con la finalidad de no complicar innecesariamente el modelo y su correspondiente tiempo de computación. De ésta forma se obtienen 17 valores propios por encima de 1 para el grupo con mayor número de valores propios cumpliendo ésta condición, que se corresponde con el tercer grupo de series. Así, se tomarán 17 factores a considerar para explicar las características de los modelos TSFA de las series.

En las dos gráficas siguientes, se observan los resultados del ajuste del modelo TSFA, en concreto, en cada gráfica se representan las puntuaciones en el tiempo de cada factor, correspondiente a cada uno de los 5 conjuntos de series considerado.

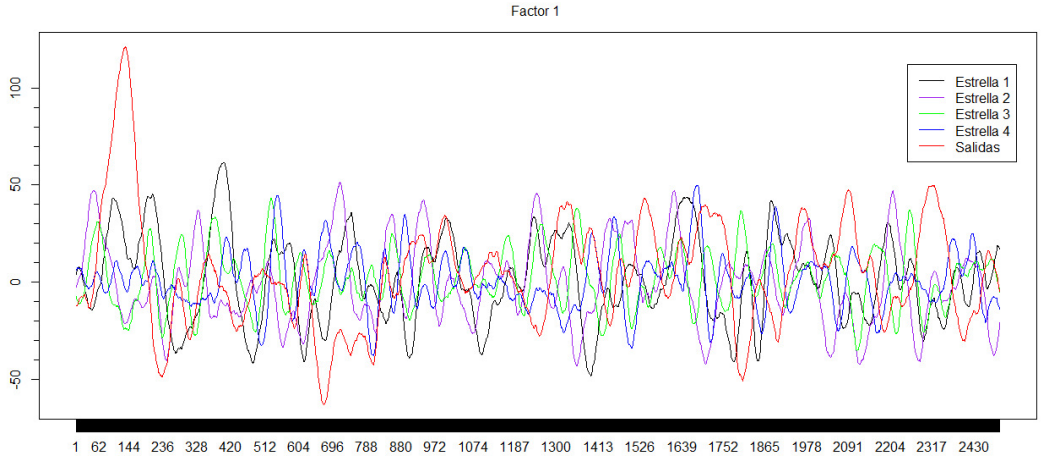



Figura 10.

El rasgo más llamativo al comparar las variaciones estructurales es la diferencia importante que sucede al comienzo del factor 1. Para explicar dicha discrepancia, se debe tener en cuenta la siguiente explicación física:

En el modelado del frente de onda, se utiliza la descomposición del mismo en combinaciones de los polinomios conocidos como polinomios de Zernike. Debido a que dicha descomposición está ponderada, en la mayor parte de los frentes de onda el polinomio que más influye es el primero, correspondiente al tipo de aberración atmosférica denominada tilt (es decir, la componente que modela la variación en la inclinación).

Si los modelos TSFA detectan en el primer factor la componente más relevante o explicativa, se corresponderá con éste modo, el tilt, debiéndose el pico de discrepancia a que habitualmente éste modo es el más complicado de corregir (de hecho, en la práctica, se puede utilizar un sistema reconstructor para únicamente ese modo y otro sistema reconstructor para los demás) y es el modo que más errores genera al intentar corregirlo, de ahí la discrepancia

únicamente en las salidas.

Más aún, que el pico se encuentre sólo al principio, puede deberse a que el sistema de redes neuronales es suficientemente robusto en el tiempo como para ir corrigiendo el tilt con más exactitud cuanta más información se dispone, cometiendo errores más grandes al principio.

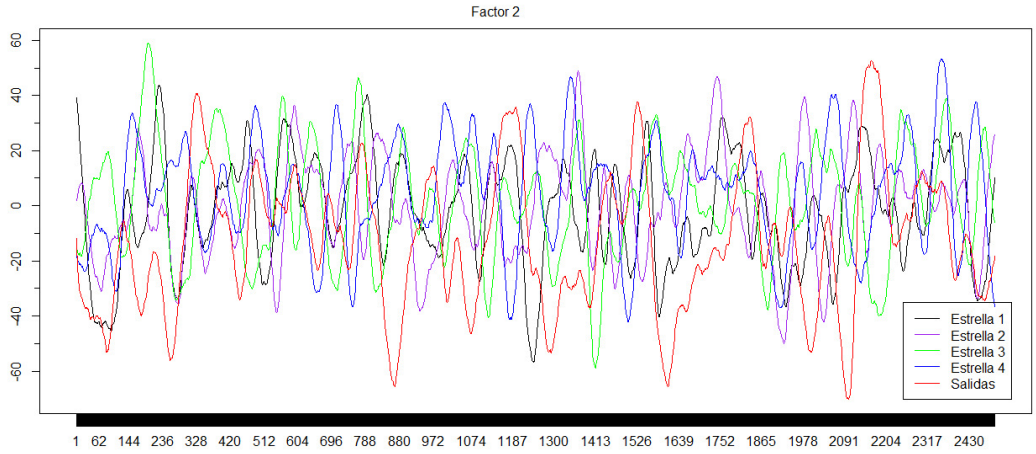



Figura 11.

En el segundo factor, no hay diferencias relevantes en la evolución en el tiempo, manteniéndose dichos valores en los mismos rangos para los cinco grupos de series.

Éste comportamiento, de diferencias no relevantes en la evolución de las puntuaciones del factor y que además son de similar magnitud entre los factores de los 4 conjuntos de entrada y el de salida, se mantiene a lo largo del resto de factores, lo cual se puede comprobar en las gráficas del resto de los factores, añadidas en el anexo I.

Los resultados de las comparaciones de la evolución temporal de los factores

(en los cuatro grupos de estrellas guía y en las salidas) sugieren que la propiedad que se quiere comprobar, que en las salidas se mantiene la estructura temporal que se observa en las entradas, puede ser correcta.

## 3.3. Modelos VARMA

Los modelos VARMA (vector auto-regressive-moving average), son la extensión de los modelos básicos de series temporales presentados en el capítulo 2. En éste caso, son modelos no para una única serie temporal sino para un vector de series temporales, que están relacionadas entre sí.

Formalmente, el proceso $k$-dimensional $\{X_t\}$, donde $X_t' = (X_{t1}, \dots X_{tk})$, está generado por un esquema estacionario e invertible VARMA(p,q) si satisface

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t - \theta_1 Z_{t-1} - \cdots - \theta_q Z_{t-q}, \qquad (3.3)$$

donde $\phi_i$ y $\theta_j$ para $i = 1, \dots, p$ y $j = 1, \dots, q$ , son matrices de coeficientes y $\{Z_t\} = (Z_{t1}, \dots, Z_{tk})$ es el vector de procesos de ruido blanco.

Equivalentemente, se puede expresar como

$$\phi(B)X_t = \theta(B)Z_t.$$

Las propiedades referentes a éste tipo de modelos y la deducción formal de los mismos pueden encontrarse en referencias como [Luceño, 1994].

## 3.4. Análisis de los datos utilizando modelado VARMA.

### 3.4.1. Paquete 'MTS'

El paquete MTS (Multivariate Time Series) permite la estimación de modelos para series temporales multivariantes. En particular, ajusta modelos vectoriales AR, vectoriales MA, vectoriales ARMA. El paquete permite el cálculo de las matrices de autocovarianza y correlación de modelos dados, o realizar el ajuste de modelos permitiendo incluso que haya datos incompletos.

Las funciones de éste paquete que se han utilizado para el análisis de los datos son las siguientes:

- *VARMA()*, que realiza una estimación máximo verosímil del modelo VARMA indicado.

  Como argumentos de entrada, requiere la matriz de datos correspondiente a los valores en el tiempo del vector de series temporales, aceptando objetos con estructura de modelo factorial. Requiere la especificación del orden AR y el orden MA.

  Como argumentos de salida, la función proporciona la estimación de los coeficientes del modelo con sus respectivos errores estándar (almacenados en secoef), la matriz de residuos y criterios de información del modelo ajustado, como el AIC.

- *VARMAcov()*, función que computa las matrices de autocovarianza y correlación de un modelo dado. Como argumentos de entrada se intro-

ducen los parámetros del modelo, por ejemplo, los que proporciona la orden *VARMA()* (matriz de coeficientes VAR, matriz de coeficientes VMA, etc.) .

### 3.4.2. Resultados de los modelos VARMA

El objetivo que se pretende obtener con éste análisis es comprobar si los modelos VARMA tienen la misma estructura y son equivalentes para el conjunto de salida y para los cuatro de entrada, lo que indicará la similitud de la estructura temporal de los 5 vectores de factores.

Uno de los criterios más habituales para escoger los órdenes del modelo VARMA más adecuado es el criterio de información de Akaike, aunque en éste caso, se plantean cinco modelos del tipo VAR(1), dado que a pesar de estar considerando vectores de dimensión pequeña en comparación con los datos originales sin procesar (vectores de 17 series frente a vectores de 288 series) la complejidad del modelo es suficientemente grande como para que no sea posible en la práctica plantear modelos más complejos.

También se ha buscado, con el objeto de facilitar su comparación, que todos los modelos tuvieran el mismo orden. Además, aunque el modelo sea sencillo, permitirá extraer conclusiones acerca del comportamiento temporal de cada uno de los conjuntos de datos que se estudian.

Dada la gran cantidad de coeficientes, es imposible hacer una comparación individual numérica, por tanto se plantea una comparación gráfica. En las siguientes gráficas se representan los valores de las matrices de coeficientes obtenidos para cada modelo VAR(1).

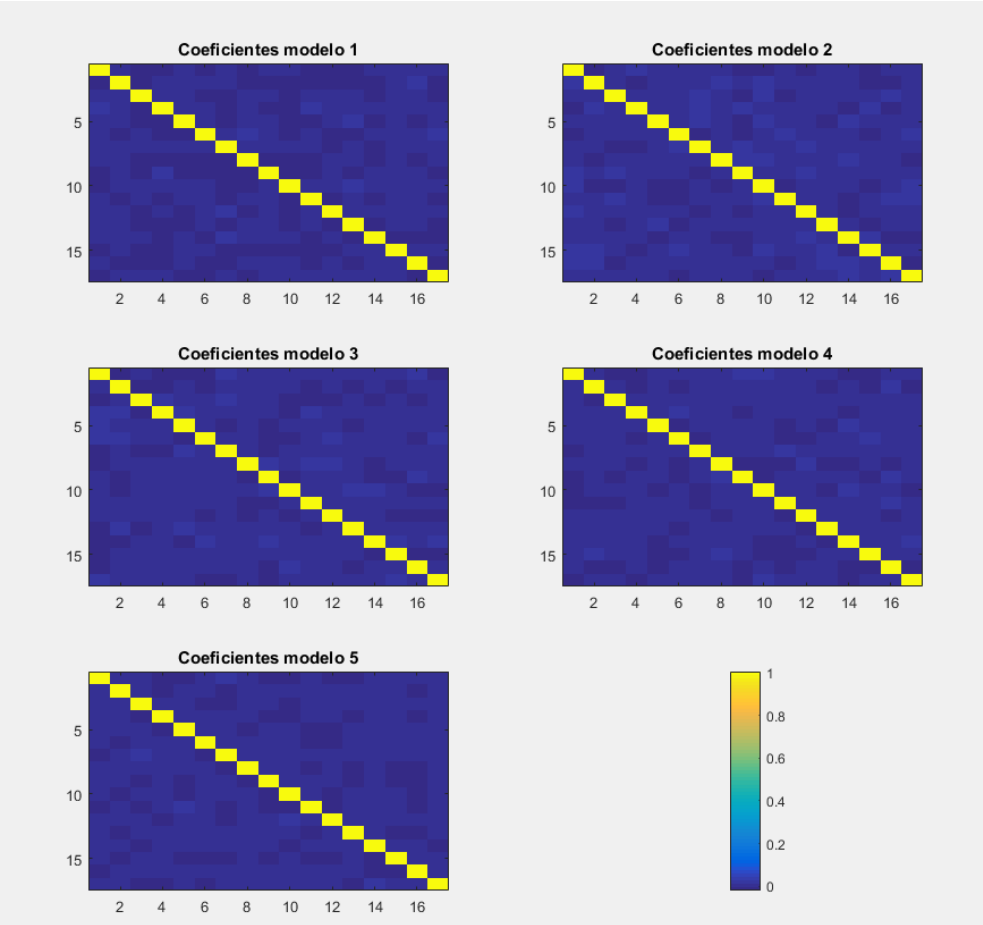



Figura 12.

Como se puede observar en los gráficos de los coeficientes de la Figura 12, la dependencia principal es la dependencia serial con el retardo de orden 1 del propio factor y tienen mucha menor importancia los retardos de todos lo demas factores. Es decir, el primer elemento de la diagonal es, en la ecuación que define el primer factor, el coeficiente del retardo 1 del primer factor, el segundo elemento de la diagonal es, en la ecuación que define el segundo factor, el coeficiente del retardo 1 del segundo factor y así sucesivamente.

Aunque ésta es una estructura muy habitual en datos reales, la similitud de

los coeficientes en los 5 modelos VAR apoya la hipótesis de la similitud en la estructura temporal de los 5 conjuntos.

Para completar la comparación de los coeficientes de los 5 modelos se analiza cuáles de esos coeficientes son significativamente distintos de 0, utilizando test t de Student individuales para contrastar la hipótesis nula $\beta_j = 0$ frente a la alternativa $\beta_j \neq 0$. La utilización del test devuelve los siguientes porcentajes de coeficientes significativamente distintos de cero por cada modelo:

| Datos | Estrella 1 | Estrella 2 | Estrella 3 | Estrella 4 | Salidas |
|---|---|---|---|---|---|
| Porcentajes | 76.797 % | 79.738 % | 78.104 % | 70.588 % | 73.856 % |

Tabla 1.

Para facilitar la comparación a partir de las matrices de p-valores obtenidos de la realización de los test, se transforman en matrices binarias que toman el valor 0 si el coeficiente no es significativamente distinto de 0 y 1 si sí lo es. En la Figura 13 se muestran gráficamente estas matrices binarias, representando en negro los valores 0 y en blanco los valores 1.

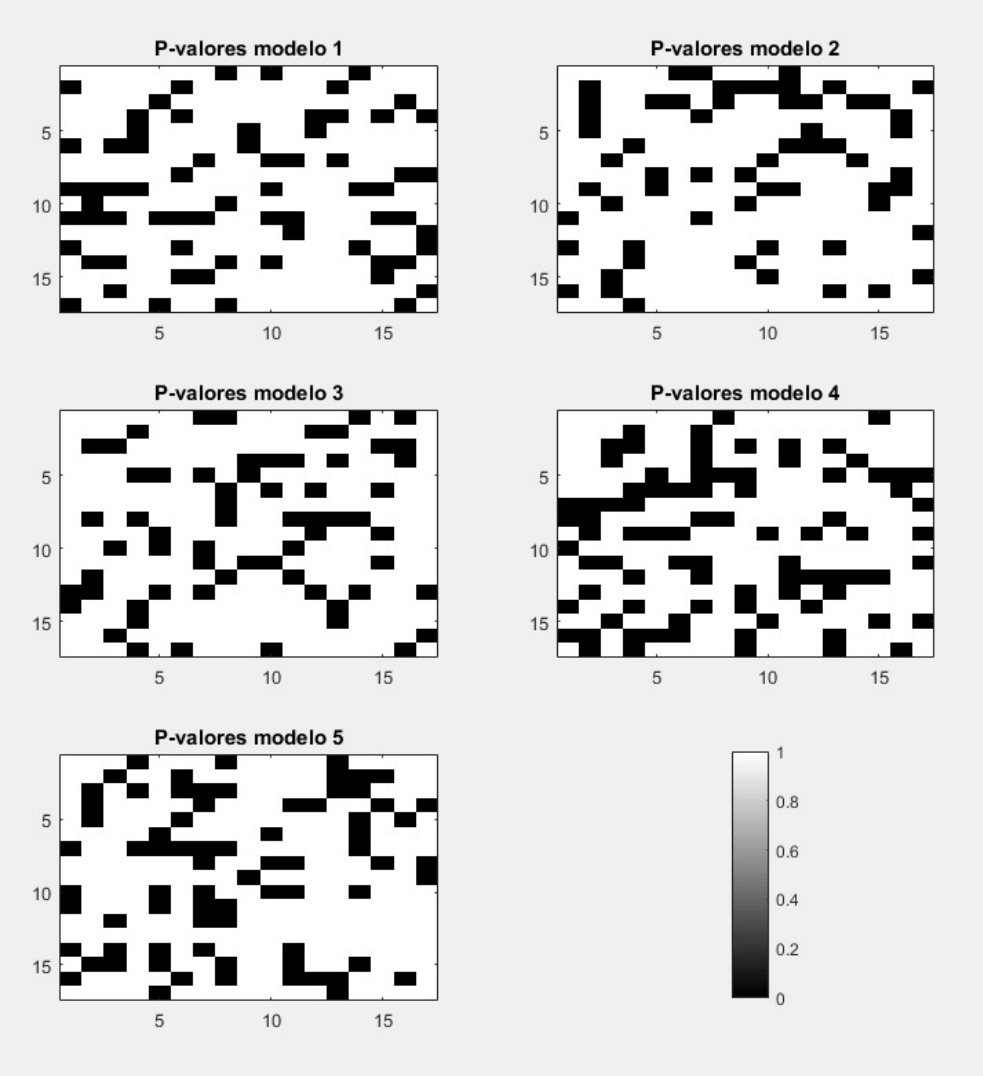



Figura 13.

Como se puede comprobar, la mayor parte de los coeficientes son significativamente distintos de cero, incluyendo la mayoría de los coeficientes para los que hay mayor dependencia.

Para completar la comparación de los coeficientes de los 5 modelos se analizan las diferencias entre los coeficientes del modelo VAR(1) de salida y los mismos coeficientes en cada uno de los 4 modelos VAR(1) de entrada. Éstas diferencias se resumen gráficamente en la Figura 14.

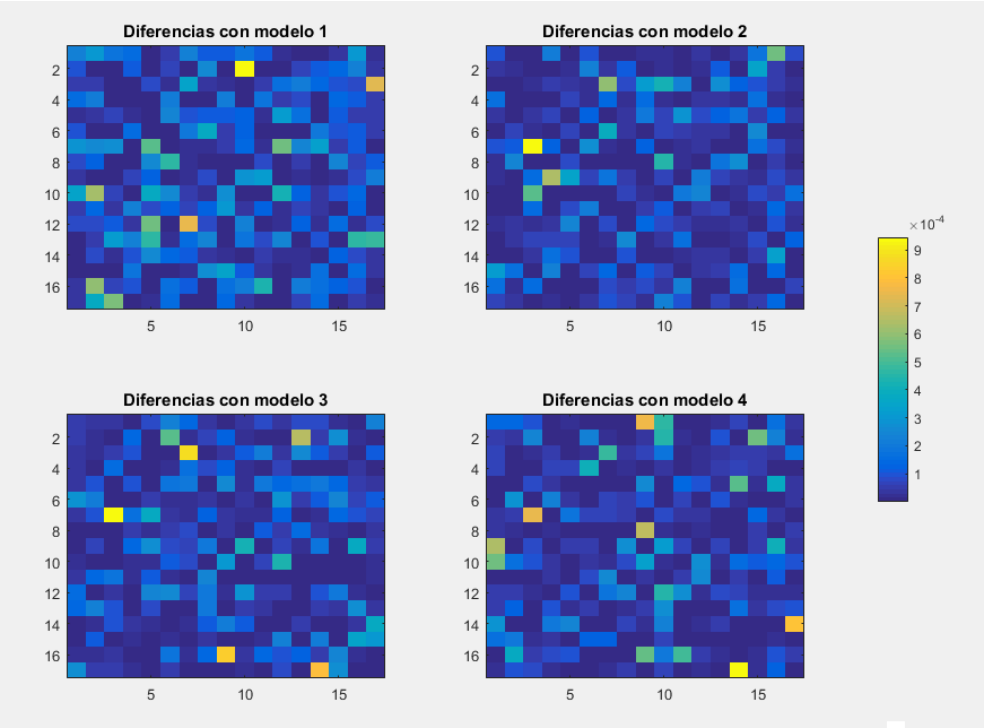



Figura 14.

Ésta representación gráfica permite comprobar fácilmente que la mayor parte de los valores de ésta matriz de diferencias son pequeños, y entre los más grandes sólo es destacable uno para la primera comparación, con valor de $7{,}254 \cdot 10^{-4}$, en la segunda comparación, el valor más destacable es $13{,}437 \cdot 10^{-4}$. El valor más alto de la tercera comparación se encuentra en $10{,}934 \cdot 10^{-4}$ y por último, el máximo de las diferencias de la cuarta comparación es encontrado en el valor $9{,}447 \cdot 10^{-4}$.

Aunque éstos valores son pequeños en términos absolutos, para tener una referencia se calculan las diferencias entre los coeficientes de dos modelos de series de entrada; por ejemplo en la Figura 15 se muestran las diferencias entre los coeficientes de los modelos de factores de entrada de la primera estrella guía y la segunda.

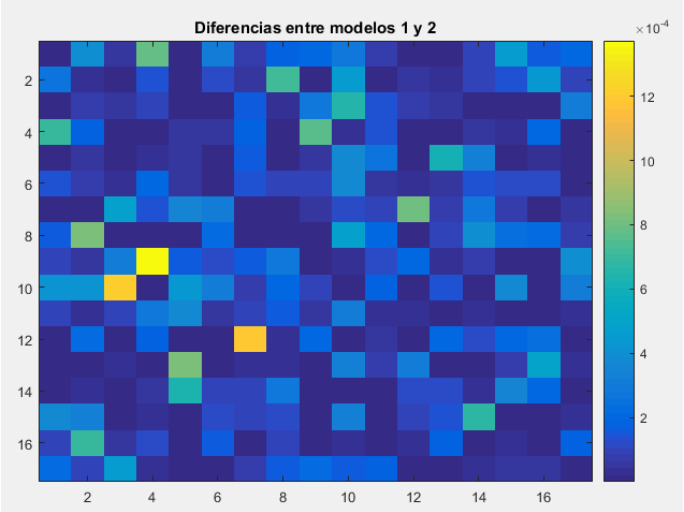



Figura 15.

El orden de magnitud de las diferencias entre modelos de entrada es igual que para las comparaciones con el modelo de salida, ya que la mayor diferencia observada entre los 2 conjuntos de entrada es 0.0014.

En resumen, los resultados indican que los coeficientes de los 5 modelos VAR(1) ajustados son similares aunque la significatividad de los mismos muestra diferencias en los distintos modelos. Sin embargo, las diferencias del patrón de significatividad en el modelo de salida no son mayores que las existentes en los patrones de los 4 modelos de entrada.

Como consecuencia, se puede concluir que no hay evidencias de que la estructura temporal de los cuatro conjuntos de factores correspondientes a cada conjunto de entrada sea diferente a la estructura temporal de los factores del conjunto de salidas.

# Capítulo 4

# Clustering de series temporales

Los métodos de clustering son técnicas que se utilizan para la resolución de problemas de reconocimiento y agrupación. Éstos problemas aparecen con frecuencia en áreas como la informática, economía, sociología y salud.

La finalidad de las técnicas de clustering es identificar estructuras en conjuntos de datos organizando objetivamente en grupos homogéneos en los que se busca minimizar la similitud entre distintos grupos y maximizar la similitud intra grupos.

En particular, las técnicas de clustering adaptadas a datos con estructura de serie temporal, permiten reconocer patrones en las estructuras de evolución temporal de dichas series. Por ésta razón, parecen la herramienta más adecuada para el análisis de las series de datos obtenidas del sistema de redes neuronales, que nos permitirán comparar estructuras y realizar agrupaciones según sus similitudes.

En éste capítulo se hace una presentación de las diversas ramas y formas de afrontar los problemas de agrupación y algunas de las técnicas más representativas.

## 4.1. Técnicas de clustering.

Tradicionalmente y en su mayoría, las técnicas de clustering han sido desarrolladas para datos estáticos, es decir, para conjuntos de datos que no varían con el tiempo, o en los que la variación en el tiempo resulta irrelevante. Éstas técnicas se pueden clasificar en cinco categorías, como aparece en [Han & Kamber, 2001]:

- Métodos de partición.

  Dado un conjunto de $n$ tuplas de datos, un método de partición es aquel que contruye $k$, con $k \leq n$ particiones de datos, donde cada partición representa un cluster que contine al menos un objeto. Si cada objeto pertenece exactamente a un cluster, las particiones se llamarán definidas (*"crisp"*); para el caso en el que los objetos tengan permitido pertenecer a diversos clusters, las particiones se llamarán borrosas (*"fuzzy"*).

  Para las particiones definidas, dos de los algoritmos de clustering más utilizados son *k-means*, en el cual cada cluster está representado por el valor medio de los objetos en el clúster y el algoritmo *k-medoids* donde cada cluster está representado por el objeto localizado más cerca del centro del cluster. Las particiones fuzzy tienen su análogo a éstos algoritmos en los conocidos como *fuzzy c-means* y *fuzzy c-medoids*.

- Métodos jerárquicos.

  Los métodos jerárquicos agrupan los datos en un árbol de clusters, la creación de éstos árboles se lleva a cabo de dos formas; con algoritmos aglomerativos y divisivos. Por una parte los aglomerativos consisten en colocar cada objeto en su propio cluster e ir fusionando clusters según ciertas condiciones especificadas (que se encuentran a cierta distancia

entre ellos, que se alcanza un número de clusters deseado, etc. ), por otra parte los divisivos hacen exactamente lo contrario.

- Métodos basados en la densidad.

  La idea general en los métodos basados en densidad es aumentar el tamaño de los clusters mientras que la densidad (número de objetos o datos puntales) en el entorno del cluster sea mayor que una cota dada.

- Métodos basados en mallados.

  Éstos métodos fraccionan el espacio donde se distribuyen los datos en un número finito de celdas, formando una estructura de mallado en la cual se desarrollan todas las operaciones sobre la información estadística extraída de los datos que se encuentran en cada celda.

- Métodos basados en modelos.

  Para éstos métodos, cada cluster asume un modelo posible que puedan tomar los datos y se intenta ajustar los datos a los modelos asumidos en cada clúster. Para éste tipo de métodos, habitualmente se emplea análisis Bayesiano o redes neuronales.

Para aplicar cualquiera de éstos métodos es necesario fijar una distancia o medida de similitud tanto entre individuos como entre clusters. La medida de distancia más habitual es la distancia euclídea o de Maharaj, y si se opta por una medida de similitud, el coeficiente de correlación.

## 4.2. Técnicas de clustering de series temporales

Al igual que para datos estáticos, para hacer clustering de series temporales es necesario un algoritmo o procedimiento de clustering, cuya elección depende del tipo de datos disponibles y de la intención en su aplicación. Se han desarrollado algoritmos que ofrecen soluciones atendiendo al orden temporal entre los datos que caracteriza a las series temporales y algunas particularidades de las mismas, como puede ser que no sean de la misma longitud, que no estén idénticamente distribuidas, etc.

La mayor parte de los algoritmos para series temporales tienen como base métodos ya existentes de clustering para conjuntos de datos estáticos, y la mayor diferencia reside en modificar la forma de definir las medidas de similitud o distancias entre los datos o los clusters, para que sean apropiadas para mantener la estructura de serie temporal.

Hay tres formas de enfocar los métodos de clustering de series temporales dependiendo de si se estudian directamente los datos sin procesar (raw data), indirectamente con aproximaciones basadas en las características de los datos, y con aproximaciones basadas en modelos.

En la aplicación de un análisis cluster se deben determinar tres aspectos: el algoritmo de clustering o agrupación, la medida de distancia o similitud y la evaluación de los grupos formados. A continuación se detallan las alternativas en cada uno de ellos.

### 4.2.1. Algoritmos de clustering.

En ésta sección se presentan algunos algoritmos básicos de los más conocidos y utilizados en general, tanto en clustering como en clustering de series tem-

porales, k-means y self-organizing maps, y otros algoritmos específicos para series temporales, como el algoritmo SOTA. Hay que mencionar que éstos son únicamente una pequeña muestra de todos los disponibles, algunos también muy conocidos como puede ser k-medoids y el algortimo de p-valores de Maharaj no se presentan aquí.

- *k-means.*

  Éstos algoritmos se basan en la minimización de una función objetivo, la cual se escoge para que sea la distancia total de un patrón respecto del centro del cluster. Es un proceso iterativo que se inicia de forma aleatoria, es decir, se colocan arbitrariamente elementos dentro de los clusters o se escogen centros de clusters aleatoriamente.

  La distribución de objetos en los clusters y la actualización de los clusters son los dos pasos principales de los algoritmos k-means; se alterna entre éstos dos pasos hasta encontrar la agrupación que minimice la función objetivo. Dadas las series $\{x_k|k = 1, \ldots, n\}$, los algoritmos k-means determinan $c$ centros de cluster, $V = \{v_i|i = 1, \ldots, c\}$ , minimizando la función objetivo para el cluster $U$, dada por

  $$min \;\; J(U, V) = \sum_{i=1}^{c} \sum_{k=1}^{n} ||x_k - v_i||^2 \tag{4.1}$$

  La norma de la ecuación anterior es la asociada a la distancia euclídea, aunque se puedan utilizar las asociadas a otras distancias.

  El procedimiento para hallar la solución iterativa utilizando éste tipo de algoritmos sigue el siguiente esquema general:

  1. Se escoge un $c$ tal que $2 \leq c \leq n$ y un $\varepsilon$ pequeño con el fin de determinar un criterio de parada. Se establece un contador $l = 0$ y los centros de cluster $V^{(0)}$ de forma arbitraria.

2. Se distribuyen $x_k$ en los clusters de forma que para todo $k$ el cluster $U^{(l)}$ sea tal que $J$ se minimice. Ésto se consigue reasignando $x_k$ a un nuevo cluster, que esté más cerca de él.
3. Se revisa la variación en los centros de cluster $V^{(l)}$, calculados con la nueva distribución de observaciones $x_k$ (suma de las observaciones entre cardinalidad del cluster). Se aplica el criterio de parada si la variación en los centros es más pequeña que $\varepsilon$; en otro caso se incrementa $l$ y se repiten los pasos 2) y 3).

Cuando se utiliza para series temporales, éste grupo de algoritmos trabaja mejor con series temporales de la misma longitud, dado que no está definido el concepto de centro de cluster cuando el cluster contiene series de distinta longitud.

- *Self-organizing maps.*

  Los mapas de auto-organización son un tipo de red neuronal entrenada con un proceso iterativo no supervisado o auto-organizante. Después de introducir un patrón en la red, se calcula la distancia euclídea entre el vector introducido y el vector de pesos, la neurona con la menor distancia se marca como $t$. En cada iteración $l$ del algoritmo, dependiendo de que cada neurona $i$ se encuentre en un entorno espacial $N_t(l)$ de $t$, su peso se actualiza de acuerdo a la siguiente expresión:

$$w_i(l+1) = \begin{cases} w_i(l) + \alpha(l)[x(l) - w_i(l)] & \text{si } i \in N_t(l), \\ w_i(l) & \text{si } i \notin N_t(l). \end{cases} \tag{4.2}$$

  Tanto el tamaño del entorno $N_t$ y el tamaño del coeficiente $\alpha$ de adaptabilidad de los pesos disminuyen monótonamente con las iteraciones según el criterio de corrección de los pesos escogido. En los algoritmos SOM (self-organizing maps) se pretende, mediante éste entrenamiento de la red, que la distribución de las neuronas representen la distribu-

ción de los datos, es decir, se espera que la topología de los datos se vea reflejada en la red.

Al igual que ocurre con los algoritmos del tipo k-means y fuzzy c-means, SOM no funciona bien con series de distintas longitudes debido a la dificultad al escoger el tamaño de la estructura de la red.

- *Algoritmo SOTA.*
  El algoritmo Self-Organizing Tree Algorithm (SOTA) es un híbrido entre un algoritmo tipo SOM, una red neuronal no supervisada, y un algoritmo jerárquico divisivo con estructura de árbol binario. Se presenta en la referencia [J. Herrero et al., 2001]. Debido a la rapidez de computación de la red es recomendado para tratar con un gran número de datos.

  El algortimo escoge el cluster $C$ ( que se corresponde con las neuronas de la red, denominadas células en éste algoritmo) con el valor más alto de diversidad, utilizando la distancia euclídea o la asociada al coeficiente de correlación de Pearson y le añade dos neuronas. Para ello, se siguen los siguientes pasos, para cada iteración $l$:

  - Se computan las distancias $d_{ij}$ entre las entradas $j$ y las neuronas $i$.
  - Se selecciona la neurona de salida $i$ con menor distancia $d_{ij}$.
  - Se actualiza la neurona como

    $$C_i(l+1) = C_i(l) + \eta_{l,i} \cdot (d_{ij} - C_i(l)) \tag{4.3}$$

    donde $\eta$ es un peso para el tiempo de convergencia del algoritmo, y se escoge por el usuario.

- La diversidad se define como

$$R_i = \frac{\sum_{j=1}^{J} d_{ij}}{J} \tag{4.4}$$

siendo $J$ el número de series que han sido asignadas a la neurona $i$ por la red. En cada iteración, el tamaño de la red aumenta añadiendo dos neuronas en estructura de árbol binario a la topología de la red.

El algoritmo finaliza cuando se cumpla un criterio de parada, como por ejemplo llegar a un cierto tamaño en la red o una cota de diversidad, y la agrupación de clusters se corresponde con la topología de la red entrenada, al igual que ocurre en los algoritmos SOM.

### 4.2.2. Medidas de similitud y distancia

Las técnicas de clusting requieren de una forma de comparar los datos, para ello se utilizan medidas de distancia o bien medidas de similitud entre los individuos del conjunto de datos, con el fin de tener un valor que permita comparar y agrupar los individuos al custer más adecuado.

Al igual que en el apartado anterior, sólo se presentan algunas de las distancias más utilizadas, y otras más especificas para clustering de series temporales. Entre las más comunes están la distancia euclídea, la cuadrática media y la de Minkowski. Como medida de similitud, la más utilizada es el coeficiente de correlación de Pearson. Entre las medidas específicas de series temporales se presenta el índice de disimilitud basado en la correlación cruzada.

- *Distancia euclídea, distancia cuadrática media y distancia de Minkowski.*

  Sean $x_i = (x_{i1}, \ldots, x_{iP})$ y $v_j = (v_{j1}, \ldots, v_{jP})$ vectores P-dimensionales. La distancia euclídea viene dada por

  $$d_E(x_i, v_j) = \sqrt{\sum_{k=1}^{P} (x_{ik} - v_{jk})^2} \tag{4.5}$$

  La distancia cuadrática media,

  $$d_{rms}(x_i, v_j) = d_E(x_i, v_j)/n. \tag{4.6}$$

  La distancia de Minkowski es una generalización de la distancia euclídea y se define como

  $$d_M(x_i, v_j) = \left( \sum_{k=1}^{P} (x_{ik} - v_{jk})^q \right)^{1/q}. \tag{4.7}$$

- *Coeficiente de correlación de Pearson.*

  Es la única medida de similitud. El coeficiente de correlación de Pearson

entre $x_i$ y $v_j$ se define como

$$\rho^2(x_i, v_j) = \frac{\sum_{k=1}^{P}(x_{ik} - \mu_{x_i})(v_{jk} - \mu_{v_i})}{S_{x_i} S_{v_j}}, \tag{4.8}$$

donde $\mu_{x_i}$ y $S_{x_i}$ son, respectivamente

$$\mu_{x_i} = \frac{1}{P}\sum_{k=1}^{P} x_{ik} \quad \text{y} \quad S_{x_i} = \sqrt{\sum_{k=1}^{P}(x_{ik} - \mu_{x_i})^2} \tag{4.9}$$

- *Índice de disimilitud basado en la función de correlación cruzada entre dos series temporales.*

  Sea $\rho^2_{i,j}(\tau)$ la correlación cruzada entre dos series temporales $x_i$ y $v_j$ con retardo $\tau$. Un índice de disimilitud se puede definir de la forma

  $$d_{i,j} = \sqrt{(1 - \rho^2_{i,j}(0)) \Big/ \sum_{\tau=1}^{max} \rho^2_{i,j}(\tau)}, \tag{4.10}$$

  en la que $max$ denota el valor máximo del retardo que se considera adecuado. De éste índice se sigue el correspondiente índice de similitud definido como

  $$s_{i,j} = exp(-d_{i,j}). \tag{4.11}$$

### 4.2.3. Evaluación de los resultados de clustering

Después de aplicar un método de clustering, surge la necesidad de comprobar la validez de la partición obtenida. Para ello se utilizan diferentes criterios, los cuales se dividen en dos grandes grupos, dependiendo de si se conoce a priori la partición correcta o no. El número de clusters que se han de crear se conoce en el primer grupo pero se suele desconocer para el segundo.

- *Criterios basados en partición conocida.*

  Sea $G$ el conjunto de clusters de la partición conocida y $C$ el conjunto de clusters que se quieren evaluar, obtenidos a partir del algoritmo o método de clustering. La medida de similitud se define como :

  $$Sim(G, C) = \frac{1}{k}\sum_{i=1}^{k} \max_{1 \le j \le k} Sim(G_i, C_j), \tag{4.12}$$

donde $G_i, C_j$ son los clusters de la partición conocida y los clusters a evaluar, respectivamente. Además,

$$Sim(G_i, C_j) = \frac{2|G_i \cap C_j|}{|G_i| + |C_j|}. \tag{4.13}$$

Siendo $|\cdot|$ la cardinalidad de los elementos de un conjunto. Éste tipo de criterios se utiliza para validar teóricamente algoritmos de clustering, o para comprobar la idoneidad de los mismos frente a otros.

- *Criterios basados en partición desconocida.*

  Éste tipo de criterios no requieren la información de la partición correcta a priori. Son criterios que proporcionan un valor que permite comparar diversos métodos y números de clusters, discerniendo el más eficiente, pero no permiten confirmar la validez o no de la partición obtenida. Se pueden distinguir dos casos, si se conoce a priori el número de clusters y si se desconoce.

*Criterio de Kosmelj.*

Sea $P_k$ el conjunto de todos los clusterings $C$ que particionan un conjunto de series temporales multivariables en un número $k$ específico de clusters. El mejor clustering , $C^*$, de entre todos los posibles viene dado según el siguiente criterio:

$$P(C^*) = \min_{C_j \in C \in P_k} \sum_{j=1}^{k} p(C_j), \tag{4.14}$$

donde

$$p(C_j) = \sum_{t=1}^{T} \alpha_t(C_j) p_t(C_j) \tag{4.15}$$

donde $\alpha$ es un peso que se escoge tal que disminuya cuando aumenta $t$, y además, para $X$ e $Y$ elementos distintos del cluster $C_j$,

$$p_t(C_j) = \frac{1}{2\omega(C_j)} \sum_{X,Y \in C_j} \omega(X)\omega(Y) d_t(X, Y). \tag{4.16}$$

En la ecuación anterior, $\omega(X)$ representa el peso de $X$,
$\omega(C_j) = \sum_{X \in C_j} \omega(X)$ representa el peso del cluster $C_j$, y $d_t(X, Y)$ es la medida de disimilitud entre $X$ e $Y$ en el instante $t$. El número de clusters $k$ más adecuado será aquel con menor $P(C^*)$.

Para determinar el número de clusters $k$, se pueden utilizar criterios como la maximización de Baragona, que consiste en maximizar la función

$$\sum_{\omega=1}^{k} \sum_{i,j \in C_\omega, i \neq j} s_{i,j} \tag{4.17}$$

donde $s_{i,j}$ es el índice de similitud definido en la ecuación 4.11.

*Índice de Dunn.*
Se basa en la relación de la distancia más pequeña entre observaciones de distintos clusters y la distancia más grande de observaciones del mismo cluster. Se define como:

$$D(C) = \frac{\min\limits_{C_k, C_l \in C, C_k \neq C_l} \left( \min\limits_{i \in C_k, j \in C_l} dist(i,j) \right)}{\max\limits_{C_m \in C} diam(C_m)}, \tag{4.18}$$

donde $diam(C_m)$ es la distancia máxima entre observaciones en el cluster $C_m$. El índice de Dunn toma valores entre cero e infinito y se busca maximizarlo.

Los criterios disponibles ofrecen distintas formas de medir la conectividad, es decir, que observaciones cercanas estén clasificadas en clusters cercanos, la compacidad, es decir, la homogeneidad de los valores de las series de cada clúster, y por último a la separación entre clusters (por ejemplo midiendo la distancia entre los centroides de cada cluster.)

## 4.3. Análisis de los datos utilizando técnicas de clustering

### 4.3.1. Paquetes 'TSclust' y 'clValid'

*Paquete 'TSclust'.*

El paquete TSclust implementa una gran cantidad de medidas de similitud y distancias, incluyendo medidas específicas para algoritmos basados en modelos, en datos sin procesar o medidas para las carácterísticas de los datos.

La función principal del paquete es la de presentar la posibilidad de computar las distintas distancias, pero no permite realizar un proceso de clustering completo, a excepción del algoritmo jerárquico de valores propios de Maharaj con la orden *pvalues.clust()*.

El paquete tiene implementado un criterio de evaluación de partición conocida, con la orden de *cluster.evaluation()*. Éste criterio hace una comparación directa del clustering conocido con el clustering que realiza un algoritmo; de ésta forma, se puede utilizar para validar teóricamente el correcto funcionamiento de algoritmos de clustering.

Algunas de las medidas que implementa éste paquete y que han sido consideradas en éste trabajo son:

- Distancia euclídea, para datos sin procesar, se puede utilizar con el comando *diss.EUCL()*.
- Medida de disimilitud basada en el coeficiente de correlación de dos series, con el comando *diss.COR()*.
- Distancia de Maharaj, medida de disimilitud de dos series basada en

los paramétros de un ajuste ARIMA de dichas series, con el comando *diss.AR.MAH()*.

*Paquete 'clValid'.*

El paquete clValid es un paquete que implementa una serie de criterios de validación de clustering de series temporales. Éste paquete, a través del paquete 'cluster', permite computar varios algoritmos de clustering. Los criterios que se presentan son aquellos que tienen en cuenta las medidas internas de los clusters, tales como la conectividad, compacidad y separaciones de las particiones cluster creadas.

La función *clValid()* permite, simultáneamente, aplicar diversos algoritmos de agrupación, calcular medidas de validación y seleccionar el número de clusters para determinar el método y el número de clústers óptimo para un conjunto de datos. Por otra parte, también se pueden computar los criterios por separado utilizando órdenes como *Dunn()* o *connectivity()*.

Se puede hacer el clustering de las series temporales utilizando los algoritmos disponibles: k-means, DIANA, PAM, CLARA, FANNY, SOM y SOTA. Las descripciones de cada uno de ellos se pueden encontrar en [Brock et al., 2008].

De la salida de los algoritmos de clustering se pueden obtener resúmenes numéricos del clustering formado, así como vectores en el cual cada componente se corresponde con una de las series consideradas para el clustering, cuyo valor es el cluster en el que se ha clasificado a dicha serie.

### 4.3.2. Resultados del análisis cluster

La diferencia entre un sistema en el que se utilice una única red neuronal, y el sistema propuesto con el que se han obtenido los datos es la variación con el tiempo, y es necesario comprobar si la estructura temporal que tienen las entradas se ve reflejada en la estructura temporal que tienen las salidas, o por el contrario no está corrigiendo las variaciones de la atmósfera.

El objetivo de ésta sección es realizar un análisis de clustering para comprobar si el algoritmo agrupa en los mismos clusters las series de entrada y las series de salida. En concreto, se van a unir todas las series (de los conjuntos de entrada y de salida) y se va a buscar la mejor partición en grupos. Si el algoritmo integra series de salida en clusters en los que hay series de entrada indicará que tienen la misma estructura, dado que en ningún momento se le indica al algoritmo que sean series distintas o la distribución u orden de las series. Si las series de salida no estuvieran relacionadas de ninguna forma con las de entrada el algoritmo debería agruparlas en clusters separados de los clusters formados por las series de entradas.

Dentro de la amplia variedad de algoritmos y distancias que están disponibles, la mejor elección depende fundamentalmente de el tipo de datos y del objetivo del análisis. En éste caso se ha elegido el algoritmo SOTA ya que permite el manejo de un gran número de series temporales. Aunque no es el algoritmo más robusto para cantidades pequeñas de series, es más adecuado para el volumen de datos considerado. Como medida de similitud se ha elegido el coeficiente de correlación.

*Elección del número de clusters.*

Se ha seleccionado como criterio de validación el índice de Dunn para saber cuál es la elección más adecuada de número de clústers. Se muestran los valores del índice para distintos números de clusters en la tabla 2:

| Nº clusters SOTA | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| Indice Dunn | 0.1363 | 0.1430 | 0.1477 | 0.1477 | 0.1478 |

| Nº clusters SOTA | 7 | 8 | 9 | 10 |
|---|---|---|---|---|
| Indice Dunn | 0.1478 | 0.1533 | 0.1594 | 0.1594 |

Tabla 2.

La elección más acertada parece ser de 9 o 10 clusters en adelante (el valor proporcionado por el índice de Dunn se mantiene para números de clusters mayores que 10), y se escoge por simplicidad, los datos obtenidos para 9 clusters.

*Partición en clusters.*

La aplicación del algoritmo SOTA con k=9 clusters da lugar a un partición con los siguientes tamaños:

| Cluster | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Tamaño | 147 | 157 | 208 | 158 | 154 | 195 | 225 | 102 | 94 |

Tabla 3.

La salida completa del algoritmo se muestra en el Anexo II.

Debido a la complejidad para examinar la salida numérica de dicha partición, los resultados se resumen de forma gráfica en la Figura 16. En esta gráfica

se representan mediante puntos cada una de las 1440 series de tiempo que resultan de la unión de las series de los 5 conjuntos de 288 series cada uno.

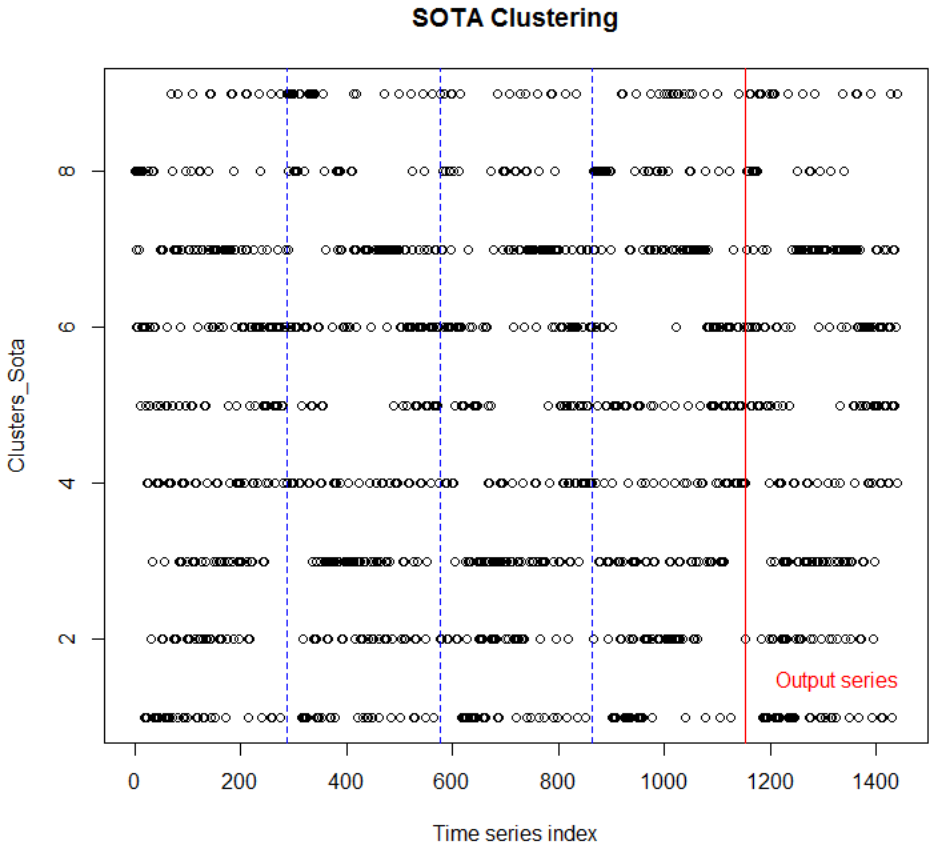



Figura 16.

En el eje X se muestra un índice de dichas series, siendo las 288 primeras las correspondientes a la primera estrella guía, las 288 siguientes las correspondientes a la segunda estrella y así sucesivamente hasta las 288 últimas que corresponden a las series del conjunto de salida (la separación de las series de cada conjunto se ha señalado con una línea vertical discontinua).

Además, las series en cada uno de los cinco conjuntos están ordenadas según su posición, de forma que la primera serie de cada uno de los conjuntos corresponde a la señal proporcionada por el mismo fotodetector, de igual forma la segunda, tercera, etc. En el eje Y del gráfico se indica a cuál de los 9 clusters definidos está asignada cada serie en la partición final.

Como se argumentaba anteriormente, si las series de salida tienen la misma estructura temporal que las de entrada cabe esperar que las series que ocupan la misma posición en los 5 conjuntos, se agrupen en el mismo cluster. Es decir, que las primeras series de cada uno de los cinco conjuntos deberían pertenecer al mismo cluster, de igual forma las segundas, etc.

En efecto, los resultados de la Figura 16 se pueden observar claramente la tendecia que tiene el algoritmo a repetir patrones de "densidad" a la hora de clasificar las series, ya que en cada uno de los clusters formados, se observa un patrón muy parecido en los 5 conjuntos de series. Esta similitud indica que los datos de entrada correspondientes a cada fotodetector, siguen el mismo patrón temporal y tienden a agruparse en los mismos grupos y lo mismo sucede con las series de salida.

La comparación gráfica se complementa con una comparación numérica. El procedimiento que se va a realizar proporcionará una medida que caracteriza la similitud que hay entre las estructuras temporales de cada conjunto de

entrada y el conjunto de las salidas.

Para calcular ésta medida se comparan las series que comparten la misma posición en los dos conjuntos de series que se comparan. Si dichas series están asignadas al mismo cluster aportan un 1 a la medida y si están asignadas a clusters de distancia máxima un 0.

Éstos valores se promedian dando lugar a una medida de similitud global entre los dos conjuntos, que se expresa en tanto por ciento. Los resultados obtenidos se muestran en la tabla 4.

| Estrella guía | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Similitud ( %) | 83.07 | 83.33 | 83.63 | 80.90 |

Tabla 4.

El resultado obtenido en media es de un 82.73 % de similitud entre las estructuras de clustering, muy similar en los tres primeros conjuntos de entrada y ligeramente más baja en el cuarto.

Todos los resultados, gráficos y numéricos, apoyan la hipótesis de que la estructura temporal de las series de los conjuntos de entrada se mantiene en el conjunto de salida.

# Capítulo 5

# Conclusiones y perspectivas de futuro

## 5.1. Conclusiones

El objetivo del análisis es comprobar la validez de un sistema de redes neuronales en constante re-entrenamiento. Para ello se busca comprobar si el sistema mantiene la estructura temporal que tienen las series de entradas y ésta se mantiene en las series de salida.

Para analizar ésta propiedad se han ensayado dos procedimientos. El primero se basa en reducir la dimensión de los datos mediante técnicas TSFA y representar la estructura temporal de los mismos mediante modelos VARMA. El segundo se basa en la aplicación de técnicas de clustering al conjunto de todas las series (sin distinguir series de entrada y de salida) y comprobar si las series correspondientes a las mismas posiciones en los cinco conjuntos se agrupan en los mismos clusters.

Las principales conclusiones respecto a la metodología son las siguientes.

- No se ha podido aplicar modelos de series temporales multivariantes directamente debido al elevado número de series que se deben analizar. El principal problema de éstas técnicas es la poca capacidad de las mismas para tratar con un conjunto grande de datos, exigiendo tiempos de computación impracticables.

- La aplicación de técnicas TSFA ha sido útil para reducir la dimensión de cada conjunto de series aunque todavía el número de factores considerado, 17 en cada conjunto, es alto.

- Las técnicas de clustering han resultado más útiles como método de comparación de la estructura temporal de los distintos conjuntos de series que la combinación de los métodos TSFA y los modelos VARMA, ya que este procedimiento presenta dificultades debido al tiempo de computación y además conllevan una pérdida de información derivada de la reducción de la dimensión mediante la represetación por factores.

Respecto a los resultados, las principales conclusiones son,

- Los modelos ajustados a los 5 conjuntos, presentan la misma estructura y permite hacer una comparación entre sus coeficientes, en la cual no se han encontrado diferencias significativas relevantes.

- En las técnicas de clustering se ha seleccionado como algoritmo de clasificación el algoritmo SOTA y como medida de similitud el coeficiente de correlación, dando buenos resultados en el análisis de datos temporales.

Se ha obtenido como resultado un conjunto de 9 clusters en los que las series que ocupan las mismas posiciones en los cinco conjuntos tienden a agruparse en los mismos clusters, pudiendo concluirse que las estructuras temporales de las entradas se mantienen a través del sistema de redes neuronales y se ven reflejadas en las salidas.

Los resultados apoyan la hipótesis que se buscaba comprobar de que el sistema de redes neuronales con re-entrenamiento funciona de forma coherente y conserva la estructura temporal de las series de entrada.

Aunque los resultados obtenidos en el trabajo desarrollado sirven para justificar un caso concreto de red neuronal aplicado a datos concretos, y no se puede generalizar, el procedimiento propuesto puede ser la base de un método de validación para comprobar si se mantiene la estructura temporal.

## 5.2. Perspectivas de futuro

Los resultados obtenidos sugieren que las redes neuronales modelan coherentemente procesos físicos en los que la dependencia temporal es una propiedad clave. Ésta comprobación empírica anima a buscar sistemas más complejos de redes que modelen éste fenomeno manteniendo la dependencia temporal con la intencion de mejorar el proceso de aprendizaje y disminuir el tiempo de entrenamiento.

Además puede resultar interesante realizar un estudio análogo para comprobar que ocurre en las matrices de pesos de estas redes que son reentrenadas constantemente para poder modelar una situación física cambiante, por ejemplo, los clusters comunes en las entradas y en las series de la matriz de pesos

indicaría como influye las variaciones de cada entrada en el modelo de la red neuronal para éste tipo de procesos físicos.

# Bibliografía


1. G.Brock, V. Pihur, Susmita Datta, Somnath Datta. *"clValid: An R Package for Cluster Validation."* Journal of Statistical Software, Volume 25, Issue 4, November 2008.

2. P. J. Brockwell, R. A. Davis. *"Time Series: Theory and Methods."* Springer, 1991.

3. F. J. de Cos Juez, F. Sánchez Lasheras, N. Roqueñí, J. Osborn. *"An ANN-Based Smart Tomographic Reconstructor in a Dynamic Environment."* Sensors, Volume 12, Pages 8895-8911, 2012.

4. M. Forni, M. Hallin, M. Lippi, L. Reichlin *"The Generalized Dynamic Factor Model One-Sided Estimation And Forecasting."* Journal of the American Statistical Association,Volume 100, No. 471, Pages 830-840, 2005.

5. P.D. Gilbert, E. Meijer. *" Time Series Factor Analysis with an Application to Measuring Money."* University of Groningen. SOM. 2005

6. J. Ham , M. Kamber. *" Data Mining : Concepts and Techniques."* Morgan Kaufmann, San Francisco, Pages 346-389, 2001.

7. J. Herrero, A. Valencia,J. Dopazo.*"A hierarchical unsupervised growing neural network for clustering gene expression patterns."* Bioinformatics, Volume 17, pages 126-136, 2005.

8. A. Luceño, *" A fast algorithm for the exact likelihood of stationary and partially nonstationary vector autoregressive-moving average processes"* Biometrika, Volume 81,3 , pages 555, 1994.

9. P. Montero, J.A. Vilar. *" TSclust: An R Package for Time Series Clustering."* Journal of Statistical Software, Volume 62, Issue 1, November 2014.

10. J. Osborn, F. J. De Cos Juez, D. Guzman, T. Butterley, R. Myers, A. Guesalaga, J. Laine. *"Using artificial neural networks for open-loop tomography."* Optics Express, Volume 20, Pages 2420-2434, 30 January 2012.

11. J. Osborn, D. Guzman, F. J. de Cos Juez, A. G. Basden, T. J. Morris, E. Gendron, T. Butterley, R. M. Myers, A. Guesalaga, F. Sanchez Lasheras, M. Gomez Victoria, M. L. Sanchez Rodríguez, D. Gratadour, G. Rousset. *"Open-loop tomography with artificial neural networks on CANARY: on-sky results."* Monthly Notices of the Royal Astronomical Society, Volume 441, Pages 2508-2514, 2014.

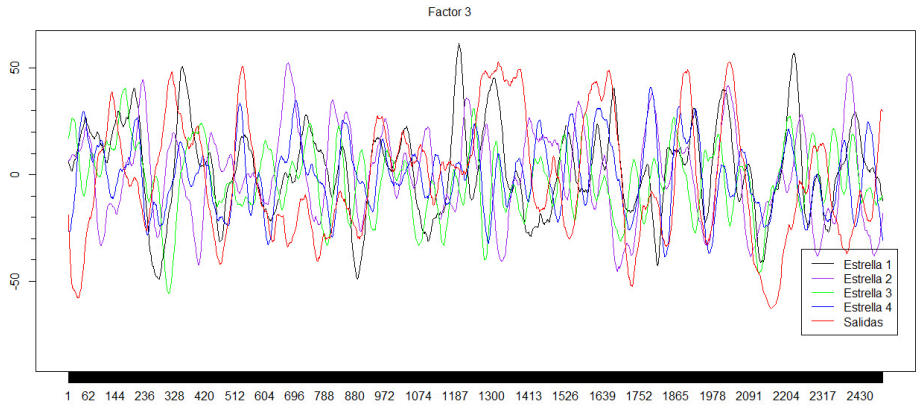
12. J.A. Vilar, A.M Alonso, J.M Vilar. *"Non-Linear Time Series Clustering Based on Non-Parametric Forecast Densities."* Computational Statistics & Data Analysis, Volume 54(11), Pages 2850-2865, 2010.

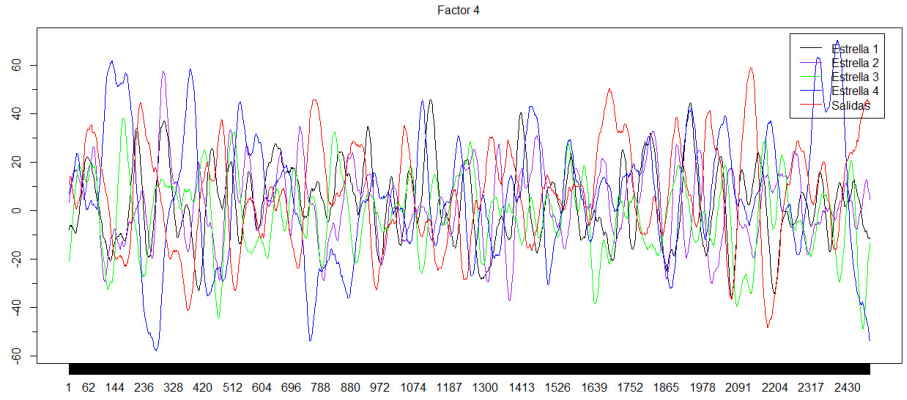
13. T. Warren Liao. *"Clustering of time series data - a survey."* Pattern Recognition 38, Pages 1857 - 1876, 2005.

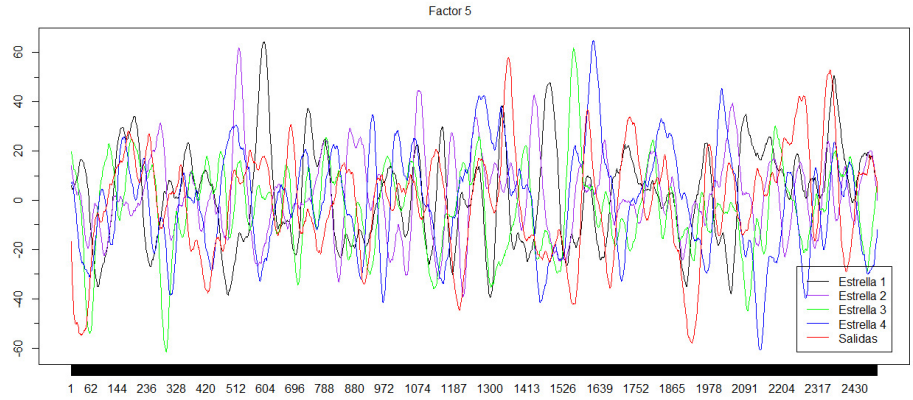
14. Y. Xiong, D. Yeung. *"Mixtures of ARMA models for model-based time series clustering."* Proceedings of the IEEE International Conference on Data Mining, Maebaghi City, Japan, 9-12 December 2002.

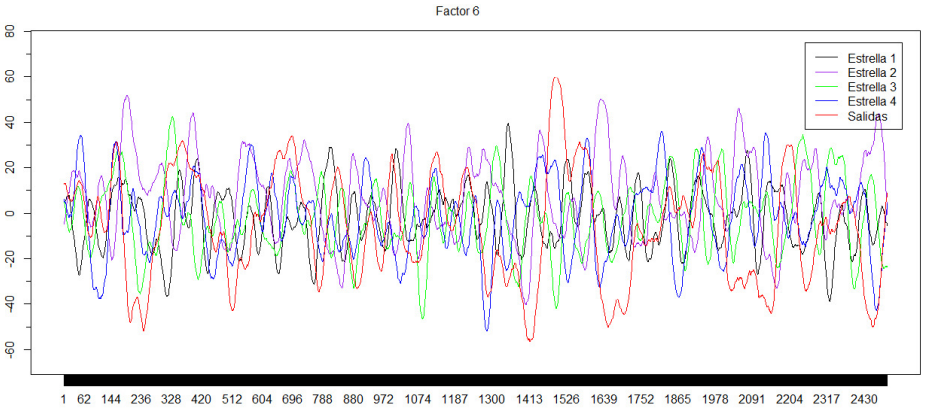
# Anexo I. Factores del modelo TSFA.

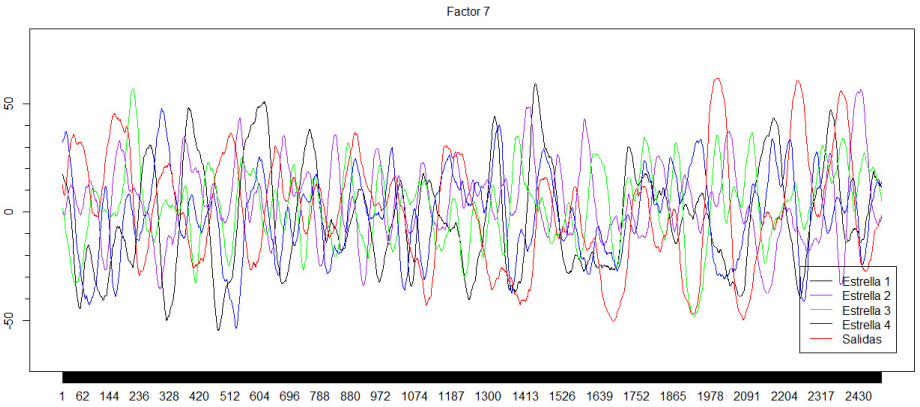
Las gráficas correspondientes a los valores en el tiempo de los factores, para los factores 1 y 2, están recogidas en la sección 3.2.2. Resultados de modelos TSFA, el resto de factores, hasta los 17 modelados, se pueden observar a continuación:

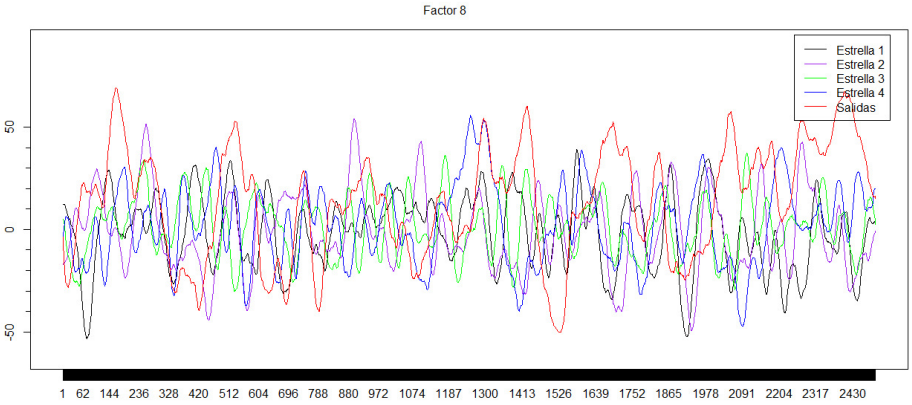


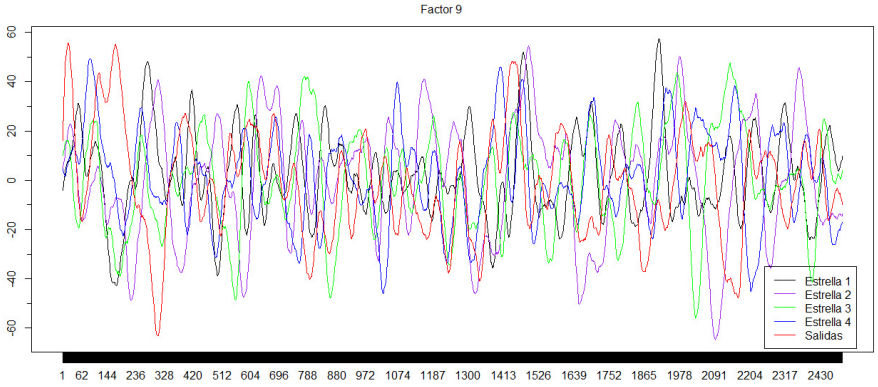

Factor 4
Estrella 1
Estrella 2
Estrella 3
Estrella 4
Salidas
60
40
20
0
-20
-40
-60
1 62 144 236 328 420 512 604 696 788 880 972 1074 1187 1300 1413 1526 1639 1752 1865 1978 2091 2204 2317 2430

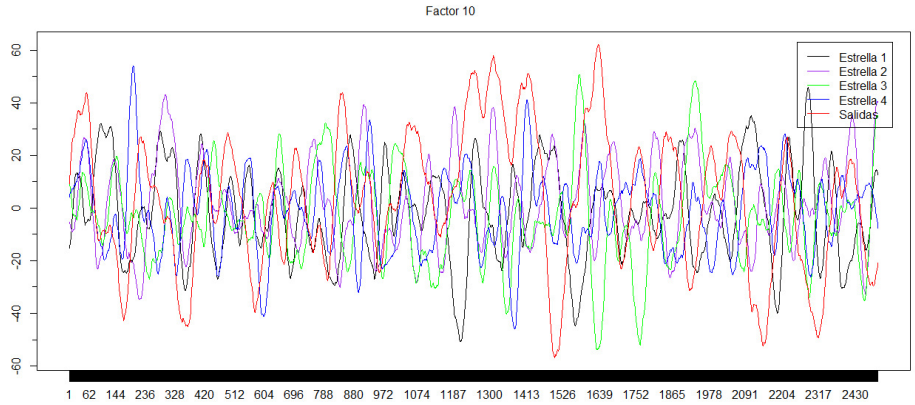

Factor 5
Estrella 1
Estrella 2
Estrella 3
Estrella 4
Salidas
60
40
20
0
-20
-40
-60
1 62 144 236 328 420 512 604 696 788 880 972 1074 1187 1300 1413 1526 1639 1752 1865 1978 2091 2204 2317 2430

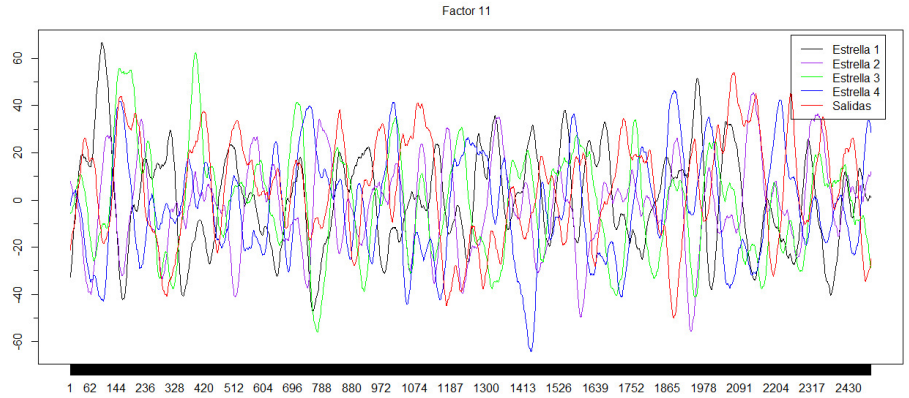

Factor 6
Estrella 1
Estrella 2
Estrella 3
Estrella 4
Salidas
80
60
40
20
0
-20
-40
-60
1 62 144 236 328 420 512 604 696 788 880 972 1074 1187 1300 1413 1526 1639 1752 1865 1978 2091 2204 2317 2430

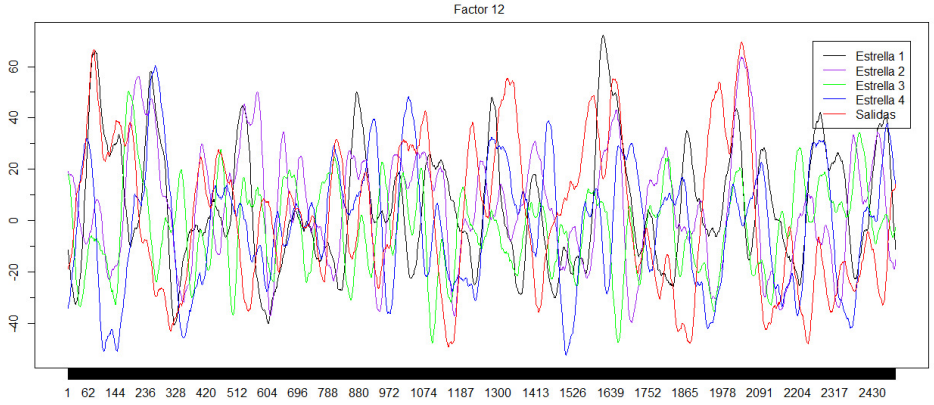

Factor 7
50
0
-50
1 62 144 236 328 420 512 604 696 788 880 972 1074 1187 1300 1413 1526 1639 1752 1865 1978 2091 2204 2317 2430
Estrella 1
Estrella 2
Estrella 3
Estrella 4
Salidas

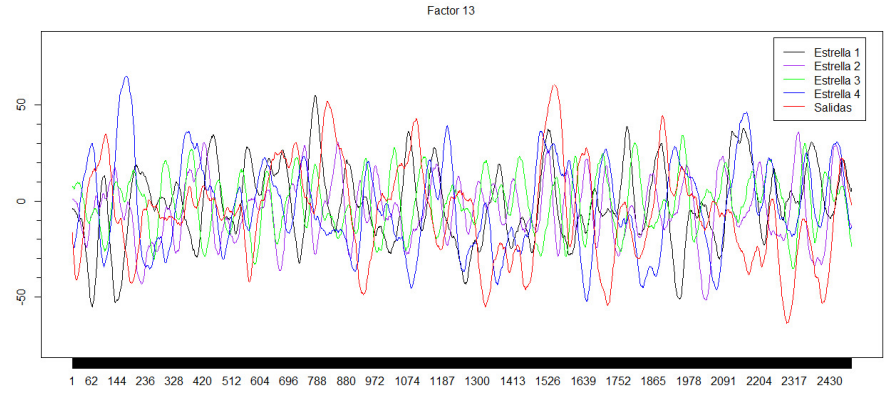

Factor 8
50
0
-50
1 62 144 236 328 420 512 604 696 788 880 972 1074 1187 1300 1413 1526 1639 1752 1865 1978 2091 2204 2317 2430
Estrella 1
Estrella 2
Estrella 3
Estrella 4
Salidas

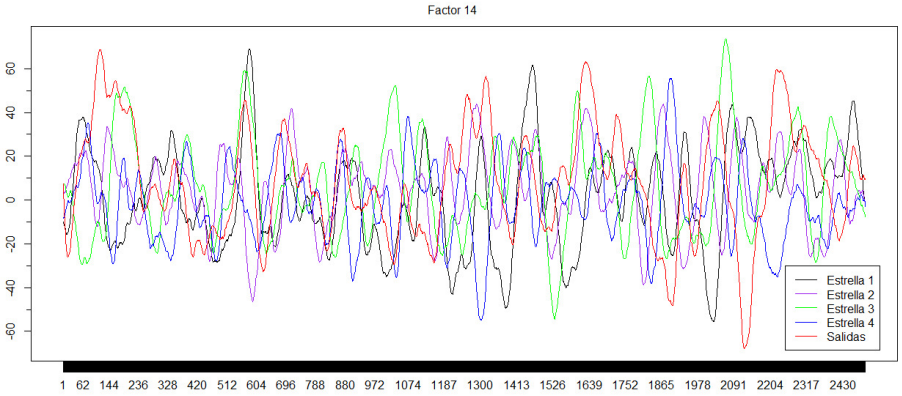

Factor 9
60
40
20
0
-20
-40
-60
1 62 144 236 328 420 512 604 696 788 880 972 1074 1187 1300 1413 1526 1639 1752 1865 1978 2091 2204 2317 2430
Estrella 1
Estrella 2
Estrella 3
Estrella 4
Salidas

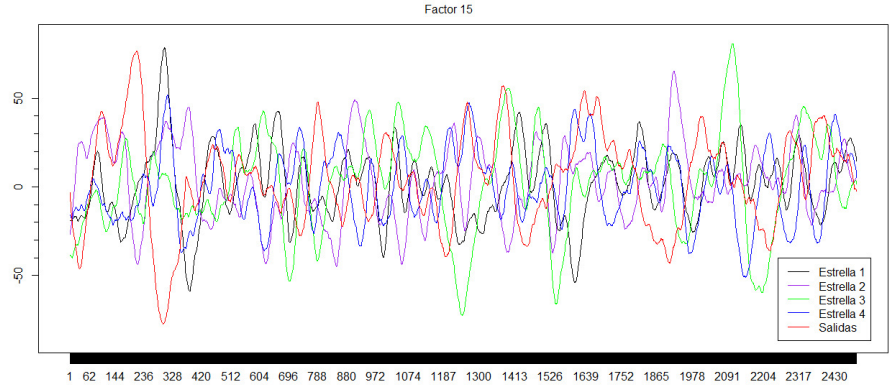

Factor 10
Estrella 1
Estrella 2
Estrella 3
Estrella 4
Salidas
60
40
20
0
-20
-40
-60
1 62 144 236 328 420 512 604 696 788 880 972 1074 1187 1300 1413 1526 1639 1752 1865 1978 2091 2204 2317 2430

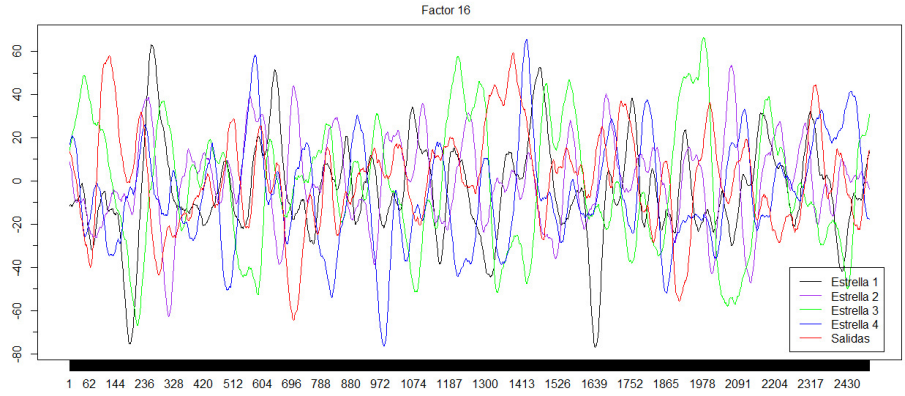

Factor 11
Estrella 1
Estrella 2
Estrella 3
Estrella 4
Salidas
60
40
20
0
-20
-40
-60
1 62 144 236 328 420 512 604 696 788 880 972 1074 1187 1300 1413 1526 1639 1752 1865 1978 2091 2204 2317 2430

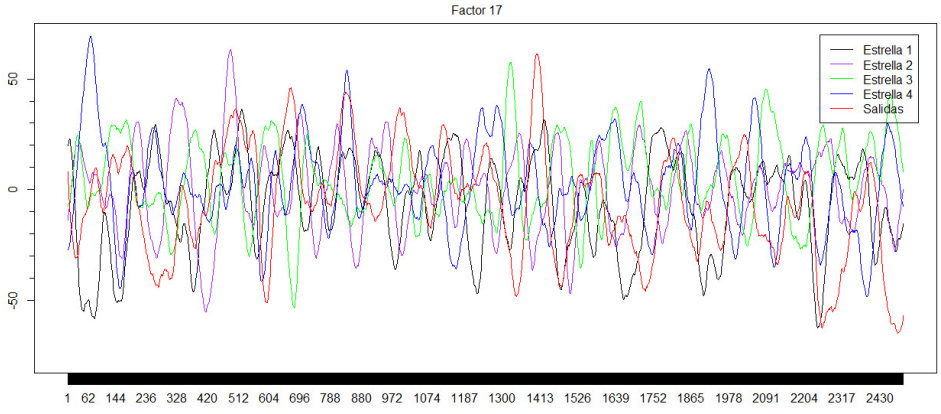

Factor 12
Estrella 1
Estrella 2
Estrella 3
Estrella 4
Salidas
60
40
20
0
-20
-40
1 62 144 236 328 420 512 604 696 788 880 972 1074 1187 1300 1413 1526 1639 1752 1865 1978 2091 2204 2317 2430


Factor 13
Estrella 1
Estrella 2
Estrella 3
Estrella 4
Salidas
50
0
-50
1 62 144 236 328 420 512 604 696 788 880 972 1074 1187 1300 1413 1526 1639 1752 1865 1978 2091 2204 2317 2430


Factor 14
Estrella 1
Estrella 2
Estrella 3
Estrella 4
Salidas
60
40
20
0
-20
-40
-60
1 62 144 236 328 420 512 604 696 788 880 972 1074 1187 1300 1413 1526 1639 1752 1865 1978 2091 2204 2317 2430


Factor 15
Estrella 1
Estrella 2
Estrella 3
Estrella 4
Salidas
50
0
-50
1 62 144 236 328 420 512 604 696 788 880 972 1074 1187 1300 1413 1526 1639 1752 1865 1978 2091 2204 2317 2430


Factor 16
60
40
20
0
-20
-40
-60
-80
1 62 144 236 328 420 512 604 696 788 880 972 1074 1187 1300 1413 1526 1639 1752 1865 1978 2091 2204 2317 2430
Estrella 1
Estrella 2
Estrella 3
Estrella 4
Salidas


Factor 17
50
0
-50
1 62 144 236 328 420 512 604 696 788 880 972 1074 1187 1300 1413 1526 1639 1752 1865 1978 2091 2204 2317 2430
Estrella 1
Estrella 2
Estrella 3
Estrella 4
Salidas

# Anexo II. Datos proporcionados por el algoritmo SOTA.

Clusters:

ID Size Diversity

1 147 4.161186

2 157 4.178428

3 208 4.174285

4 158 4.234844

5 154 4.114185

6 195 4.200422

7 225 4.137497

8 102 4.004849

9 94 3.946852

Array con la clasificación de clusters para cada serie:

(1) 8 7 8 6 8 6 8 7 8 5 8 5 8 6 8 6 8 6 1 6 1 5 1 4 8 4 8 5 6 2 2 2 8 3 8 6 1

(38) 6 1 5 1 4 1 4 1 4 1 5 7 2 7 2 7 2 3 1 3 5 1 6 1 5 1 4 1 4 1 4 9 5 8 1 7 2

(75) 7 2 7 2 7 9 7 1 3 5 3 6 3 7 3 4 1 4 1 4 1 5 8 1 2 3 7 2 7 2 7 8 5 9 5 3 7

(112) 3 2 4 3 4 1 6 1 7 8 2 8 2 2 3 7 2 7 1 5 1 5 3 2 4 2 6 8 7 9 2 9 1 2 7 6 6

(149) 7 3 7 1 7 3 4 4 7 6 3 2 3 2 2 3 2 7 6 7 7 3 7 1 7 2 7 3 7 5 7 4 7 9 7 9 3

(186) 8 2 2 6 7 4 5 4 3 4 3 4 2 4 3 7 3 6 6 6 4 6 4 7 9 7 9 3 1 2 2 6 5 6 3 4 3

(223) 4 7 4 5 6 5 6 6 6 4 6 9 6 8 6 8 7 1 6 3 5 3 5 5 5 4 5 7 4 6 6 6 6 9 6 1 6

(260) 1 6 5 5 5 5 4 5 6 5 7 6 6 6 9 6 1 5 5 5 4 5 6 6 7 6 6 4 9 8 7 9 6 9 6 9 4

(297) 9 4 9 4 8 6 8 6 8 6 8 6 9 4 9 4 9 5 1 5 1 2 8 6 1 6 1 6 1 6 9 4 9 4 9 5 9

(334) 5 9 3 9 2 9 2 9 2 1 3 1 6 1 6 1 4 3 4 1 5 9 5 8 3 8 3 7 2 3 2 3 2 1 3 1 3

(371) 1 6 3 6 3 4 3 4 1 4 8 7 8 3 8 3 8 4 7 2 7 2 6 2 6 3 3 3 3 6 3 4 3 4 3 6 8

(408) 3 8 3 8 9 7 3 7 2 6 9 7 1 3 3 3 4 3 2 3 2 2 2 2 1 7 3 7 3 7 1 7 1 7 2 4 4

(445) 3 6 3 2 1 2 3 3 7 7 4 3 4 3 7 1 7 2 7 3 7 4 7 4 7 9 7 2 3 2 3 2 6 7 7 3 4

(482) 3 7 1 7 2 7 5 7 4 7 4 7 4 7 4 7 9 7 1 6 1 6 2 5 3 5 3 7 2 7 2 6 5 6 4 6 4

(519) 6 4 6 9 6 8 7 1 3 1 5 2 5 2 5 3 6 7 6 6 6 4 6 4 6 9 6 8 7 1 5 2 5 3 5 7 5

(556) 6 6 4 6 4 6 9 6 1 5 7 5 7 5 6 5 6 4 4 6 9 2 7 2 7 8 7 9 6 9 4 8 4 2 6 2 6

(593) 8 6 8 7 9 6 9 4 8 4 8 5 3 2 2 6 2 6 8 6 8 6 9 6 1 5 1 5 1 5 1 3 1 2 1 2 3

(630) 7 3 6 3 3 1 6 1 6 1 5 1 5 1 5 1 5 3 3 3 2 3 2 3 2 3 2 1 6 1 3 1 6 1 6 3 5

(667) 3 4 3 4 8 5 2 3 2 7 3 2 3 2 3 2 3 9 1 3 1 3 3 4 3 4 3 4 8 7 8 3 8 3 2 3 3

(704) 2 7 2 7 9 3 3 3 4 3 6 2 7 8 2 8 1 2 3 2 3 7 2 7 9 3 3 2 4 2 6 2 2 9 8 8 1

(741) 7 7 7 3 7 3 7 1 7 3 7 3 7 4 7 4 4 6 3 9 3 8 1 1 2 7 7 3 7 3 7 1 7 3 7 3 7

(778) 7 7 5 7 4 7 4 7 9 6 9 6 1 3 8 6 7 7 2 7 1 7 3 7 3 6 5 6 5 6 4 6 4 5 9 7 9

(815) 5 1 6 2 4 3 4 3 6 5 6 7 6 5 6 4 6 4 6 9 6 1 5 3 5 7 5 7 4 7 4 5 6 4 4 4 4

(852) 1 5 7 5 7 5 6 4 6 6 4 6 4 2 7 8 7 8 6 8 4 8 5 8 3 8 3 8 7 8 6 8 6 8 4 8 5

(889) 8 3 8 5 2 3 8 3 8 7 8 6 1 4 1 5 1 5 1 5 1 5 3 5 3 3 3 2 2 2 9 3 9 3 1 5 1

(926) 5 1 5 1 4 1 4 1 7 2 1 7 2 3 2 3 2 3 8 3 9 3 1 1 5 1 5 1 5 1 4 1 7 1 7 8 2

(963) 2 4 3 2 3 2 7 8 7 8 5 9 7 1 2 5 3 4 3 7 8 7 8 2 8 9 2 2 2 2 7 8 7 8 7 9 4

(1000) 5 2 7 9 2 9 2 8 9 2 3 2 3 7 2 7 9 7 9 4 5 7 6 2 2 9 2 9 3 2 3 7 7 7 2 7 9

(1037) 4 9 7 1 7 4 7 5 7 9 7 8 3 8 3 3 9 7 2 7 2 2 7 3 4 2 7 3 7 5 7 5 7 4 7 4 7

(1074) 9 7 8 7 1 6 3 6 7 6 3 6 3 5 3 5 3 6 5 6 5 6 4 6 4 6 9 6 8 4 1 6 3 6 3 5 3

(1111) 5 3 5 4 4 5 6 4 6 4 6 8 6 1 6 5 5 5 5 7 5 4 4 4 4 4 6 4 6 9 5 5 5 5 5 4 4

(1148) 4 4 4 6 4 2 7 8 6 8 6 9 6 9 5 9 5 8 6 8 7 8 6 8 6 8 6 8 5 9 5 9 5 9 2 9 7

(1185) 1 7 1 6 1 6 1 7 1 5 1 5 9 4 2 5 3 2 9 2 9 2 9 3 1 6 1 6 1 5 1 4 1 4 1 4 2

(1222) 5 2 3 3 2 3 2 6 2 3 9 1 3 1 5 1 6 1 7 1 4 1 4 1 4 2 7 2 3 8 3 7 2 7 2 7 2

(1259) 7 9 7 3 3 3 3 7 3 4 3 4 3 7 8 1 8 1 2 3 7 2 7 2 7 9 7 3 3 4 3 6 3 7 1 8 2

(1296) 1 2 3 7 3 7 3 7 1 7 3 7 4 4 6 3 2 1 8 1 1 2 7 7 3 4 3 7 1 7 2 7 3 7 5 7 4

(1333) 7 6 3 9 3 8 3 2 3 7 6 7 7 3 7 1 7 2 7 3 7 3 7 5 7 5 7 4 7 9 7 9 6 1 6 2 5

(1370) 7 6 2 5 3 5 3 6 5 6 5 6 5 6 4 6 4 6 9 6 9 4 1 6 2 5 3 5 7 5 7 5 7 6 5 6 4

(1407) 6 4 6 1 6 1 5 7 5 5 5 7 5 7 4 6 6 4 4 9 6 1 5 5 5 7 5 7 5 6 6 6 4 9