1. ARTÍCULOS DE ESTADÍSTICA

EL ANÁLISIS DE SERIES TEMPORALES: SITUACIÓN Y PERSPECTIVAS

Daniel Peña e Ismael Sánchez
Universidad Carlos III de Madrid

1. Introducción

Este trabajo presenta los modelos más utilizados en el análisis de series temporales, sus aplicaciones más importantes y algunas de las lineas abiertas de investigación. Se presentan en primer lugar los análisis descriptivos que se utilizaron extensamente hasta la segunda mitad del siglo XX para representar series temporales. Estos modelos fueron sustituidos por los modelos ARIMA y los modelos en el espacio de los estados, que han alcanzado a finales del siglo XX una posición central en las aplicaciones y pueden considerarse ya como los modelos clásicos de series. En la última parte del siglo XX comenzó el desarrollo de los modelos no lineales y no paramétricos, que han permitido la extensión del análisis de series a situaciones donde los modelos clásicos no resultan apropiados.

Una serie temporal es el resultado de observar una variable a lo largo del tiempo, generalmente en intervalos regulares (cada día, cada mes, cada año, etc). Cuando los valores de la serie oscilan alrededor de un nivel fijo con variabilidad constante a lo largo del tiempo y la dependencia entre las observaciones sólo depende de la distancia que las separa y permanece constante en el tiempo decimos que la serie es estable o estacionaria. En otro caso, decimos que la serie es no estacionaria. Un caso particular importante de una serie no estacionaria aparece cuando el nivel de la serie varía siguiendo un ciclo, como ocurre con las temperaturas mensuales dentro del año, y este ciclo se repite aproximadamente en periodos de tiempo fijos (cada año, cada mes, cada día). A esta clase de series no estacionarias las llamamos series estacionales.

2. Los modelos descriptivos de series temporales

Las primeras series que se analizaron cuantitativamente fueron las de variables climatológicas, como la temperatura y la pluviosidad. Estas series presentan un comportamiento cíclico estacional, como consecuencia de la rotación de la tierra alrededor del sol. Fourier demostró a principios del siglo XIX que toda función periódica puede representarse como suma de funciones sinusoidales de distinta amplitud y frecuencia y los primeros análisis de series climatológicas utilizaron el ajuste de funciones sinusoidales de distinta amplitud y frecuencia. Sea z_t una serie temporal observada en los instantes, $t = 1, ..., T$. Podemos representar exactamente esta serie mediante la descomposición

$$z_t = \mu + \sum_{j=1}^{T/2} A_j \text{sen}(w_j t) + \sum_{j=1}^{T/2-1} B_j \cos(w_j t).$$

donde μ es la media de la serie, A_j y B_j son amplitudes y $w_j = 2\pi j/T$ siendo $j = 1, 2, ..., T/2$. Es fácil comprobar que esta ecuación contiene T parámetros (la media μ más $T/2$ amplitudes A_j y los $T/2 - 1$ amplitudes B_j). En consecuencia, tenemos que encontrar un procedimiento para seleccionar las frecuencias que debemos incluir para explicar la evolución de la serie. Este es el objetivo del periodograma, que es una representación de la contribución de cada frecuencia a la varianza de la serie y puede construirse estimando el modelo anterior. De forma similar al análisis de la varianza podemos descomponer la varianza de la serie en partes debidas a cada uno de los componentes (funciones sinusoidales de distinta frecuencia) y obtener un modelo capaz de generar buenas predicciones seleccionando los componentes más importantes. El periodograma es también útil para detectar ciclos deterministas en la serie. Por ejemplo, en una serie mensual estacional esperamos encontrar un valor alto del periodograma para $f = 1/12$, pero también son previsibles valores altos para $f = j/12$, es decir, $1/6$, $1/4$, $1/3$, que son armónicos del periodo estacional.

El análisis de Fourier de una serie se extendió en la primera mitad del siglo XX para series que además de estacionalidad cíclica presentan tenden-

cias, de manera que su nivel no es constante en el tiempo, como ocurre en muchas series económicas. Los llamados métodos de descomposición de series suponen que los datos se generan como suma de tres efectos:

$$z_t = \mu_t + S_t + a_t$$

donde μ_t es el nivel de la serie, S_t es el componente estacional y a_t es el componente puramente aleatorio o innovación. Inicialmente, se supuso que el nivel de la serie podría explicarse por una función determinista del tiempo de tipo polinómico (generalmente lineal) mientras que la estacionalidad se representaba mediante ciclos. Sin embargo, es muy poco frecuente que una serie tenga una tendencia determinista y, como puede verse en Peña (1995) estos modelos generan predicciones con estructuras difíciles de justificar. Un avance importante en el análisis de series temporales es la representación de tendencias mediante métodos de suavizado, como los métodos de alisado exponencial, que permiten que la tendencia cambie suavemente en el tiempo. Estos modelos son casos particulares de los modelos ARIMA,que comentamos en la sección siguiente.

3. Modelos ARIMA y modelos en el espacio de los estados

Según el teorema de descomposición de Wold, una serie estacionaria gausiana de media μ puede siempre representarse mediante

$$z_t = \mu + \sum_{j=0}^{\infty} \psi_j a_{t-j}, \qquad (1)$$

donde $\sum_{j=0}^{\infty} \psi_j < \infty$ y a_t es un proceso de variables normales independientes y con la misma distribución $N(0, \sigma^2)$, que llamaremos en adelante innovaciones o proceso de ruido blanco para simplificar. Esta representación se conoce como la forma MA(∞) de un proceso estacionario y no es muy operativa al incluir infinitos parámetros. Puede verse en (1) cómo z_t es causado por las innovaciones pasadas y presentes. Los modelos ARMA son aproximaciones a esta representación general (1) con pocos parámetros. Los modelos MA(q), con $q < \infty$ suponen que sólo los primeros q coefficientes son no nulos. Por ejemplo, un MA(1) de media μ sigue el modelo $z_t = \mu + a_t - \theta a_{t-1}$, donde θ es una constante a determinar. Otro tipo de modelos lineales que son casos particulares de (1) son los modelos autorregresivos (AR), que establecen una dependencia lineal entre los valores presentes y pasados de una serie temporal. Por ejemplo, el modelo más simple supone una dependencia de sólo un periodo y conduce al modelo AR(1) $z_t = c + \phi z_{t-1} + a_t$ donde c y $-1 < \phi < 1$ son constantes a determinar. La generalización de este modelo para cualquier orden de dependencia lleva al proceso AR(p). Si superponemos la estructura AR y MA se obtienen los procesos ARMA. Su forma es:

$$z_t = \phi_1 z_{t-1} + \cdots + \phi_p z_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}. \qquad (2)$$

Existen muchos procedimientos para identificar la estructura de un modelo ARMA para una serie concreta (Peña, 2005) destacando las representaciones gráficas basadas en las funciones de autocorrelación. La funcion de autocorrelación mide la dependencia lineal entre la serie en el instante t y la serie en un momento anterior, t-k. Para series estacionarias las autocorrelaciones sólo dependen del retardo k y deben tender a cero al aumentar k, para que se mantenga la constancia del nivel de la serie a lo largo del tiempo. Box y Jenkins (1976) propusieron un procedimiento eficiente para identificar el tipo de modelo en una serie real, estimar sus parámetros por máxima verosimilitud y comprobar analizando los residuos si el modelo es adecuado. Puesto que el principal objetivo en la modelización ARMA es la predicción de valores futuros la modelización de series temporales debe incluir una evaluación de la capacidad predictiva del modelo seleccionado (West, 1996; Peña y Sánchez, 2005). En muchas ocasiones, además de la predicción puntual es necesario proporcionar intervalos de predicción o estimaciones de la densidad de la predicción (Pascual et al, 2001).

La generalización de estos modelos a procesos no estacionarios lleva a los modelos ARIMA, que suponen que alguna diferencia de la serie sigue un modelo ARMA. Diremos que un proceso es *integrado de orden* $h \geq 0$, y lo representaremos por $I(h)$, cuando al diferenciarlo h veces se obtiene un proceso estacionario. En la práctica, la mayoría de las series no estacionarias que son integradas tienen un orden $h \leq 3$. Los modelos ARIMA han mostrado ser una herramienta muy eficaz para modelar y prever series económicas, demográficas y sociales.

Una generalización de los modelos ARIMA es permitir un orden de integración $0 < h < 1$. Es decir, el orden de integración h es fraccionado (modelos ARFIMA). Estos modelos fueron introducidos

por Granger y Joyeux (1980) y tienen la propiedad de memoria larga, ya que la dependencia entre las observaciones se mantiene durante largos periodos. Estos modelos son útiles en series que observamos con alta frecuencia (cada minuto, cada hora) y han sido útiles para explicar series físicas, climatológicas y financieras.

La inclusión de variables exógenas en el modelo lleva a los denominados modelos de función de transferencia o modelos ARMAX, que son la extensión de los modelos de regresión múltiple a series temporales. Un ejemplo sencillo de función de transferencia sería $z_t = \phi z_{t-1} + \alpha_0 x_t + \alpha_1 x_{t-1} + a_t$. La identificación de una función de transferencia es compleja cuando hay más de una variable exógena (véase Peña, 2005). Asimismo, existe la generalización de los modelos ARIMA univariantes al caso de vectores; es decir, al caso en que z_t es un vector de series temporales (modelos VARIMA) (Lutkepohl, 1993). Todos los procedimientos de análisis multivariante pueden extenderse al caso de las series temporales. En concreto, el análisis factorial de series temporales (Peña y Poncela, 2006) es una linea muy activa de investigación.

Una alternativa a los modelos ARIMA son los modelos en el espacio de los estados (Durbin y Koopman, 2001). Estos modelos provienen del mundo físico y concretamente de la ingeniería aeroespacial. En la versión más simple que vamos a presentar aquí se supone que la serie se ha generado como función de un vector de variables de estado o parámetros, α_t, de dimensión $p \times 1$ que no se observa

$$z_t = \mathbf{H}_t \alpha_t + \epsilon_t, \tag{3}$$

donde $\mathbf{H}_t$ es un vector $1 \times p$ que suponemos conocido para todo t y ϵ_t es un proceso de ruido blanco. Por otro lado los parámetros α_t evolucionan en el tiempo mediante la llamada ecuación de estado

$$\alpha_t = \mathbf{\Omega}_t \alpha_{t-1} + \mathbf{u}_t \tag{4}$$

donde $\mathbf{\Omega}_t$ es una matriz conocida de dimensión $p \times p$ y $\mathbf{u}_t$ otro proceso de ruido blanco, independiente del anterior.

Es fácil comprobrar que los modelos ARMA y los modelos en el espacio de los estados son equivalentes y que un modelo ARMA puede considerarse como una forma reducida de la forma de estado. Para ver esta relación dado un modelo ARMA(p,q), definamos $z_t = (1, 0, .., 0)\alpha_t$ donde $\alpha_t = (\alpha_{1,t}, \alpha_{2,t}, ..., \alpha_{m,t})'$ y $m = \text{máx}(p, q+1)$. La ecuación de estado es:

$$\begin{bmatrix} \alpha_{1,t} \\ \alpha_{2,t} \\ \vdots \\ \alpha_{m,t} \end{bmatrix} = \begin{bmatrix} \phi_1 & 1 & ... & 0 \\ \phi_2 & 0 & ... & 0 \\ \vdots & 0 & \vdots & 1 \\ \phi_m & 0 & ... & 0 \end{bmatrix} \begin{bmatrix} \alpha_{1,t-1} \\ \alpha_{2,t-1} \\ \vdots \\ \alpha_{m,t-1} \end{bmatrix} + \begin{bmatrix} 1 \\ -\theta_1 \\ \vdots \\ -\theta_m \end{bmatrix} a_t \tag{5}$$

y es fácil comprobar que sustituyendo sucesivamente en las variables de estado se obtiene la representación del proceso ARMA. Los modelos en espacio de estados han sido muy utilizados en el análisis bayesiano de series temporales (West y Harrison, 1989) y tienen la ventaja de que su generalización multivariante es muy simple y con menos riesgos de sobreparametrización que los modelos VARIMA.

4. Modelos no lineales

Un modelo no lineal puede considerarse como una generalización de la descomposición de Wold (1) al caso en que los coeficientes ψ_j cambian con el tiempo. Al igual que sucede con (1), esta formulación tiene limitado interés práctico, por lo que la modelización de series no lineales suele circunscribirse a casos que tengan aplicaciones reales.

4.1. Modelos ARCH/GARCH

En un modelo lineal, la varianza de z_t condicionada a las observaciones anteriores es constante. Este supuesto, sin embargo, no se cumple en muchas series reales, siendo ejemplos especialmente relevantes muchas series económico-financieras. En un modelo ARCH, sin embargo, la varianza condicionada tiene dependencia. Un modelo ARCH se define como $z_t = \sigma_t a_t$ donde

$$E(z_t^2 | z_{t-1}, z_{t-2}, ...) = \sigma_t^2 = c_0 + c_1 z_{t-1}^2 + \cdots + c_p z_{t-p}^2, \tag{6}$$

con c_i constantes a determinar. Estos modelos fueron introducidos por Engel (1982) para la modelización de de la volatilidad en series financieras. La práctica ha demostrado que muchas series financieras requieren de valores de p muy altos. Para evitar este efecto se han propuesto los llamados modelos GARCH, que tienen la forma

$$\sigma_t^2 = c_0 + \sum_{i=1}^{p} c_i z_{t-i}^2 + \sum_{j=1}^{q} b_j \sigma_{t-j}^2, \tag{7}$$

de esta forma es posible conseguir modelos con menos parámetros que los ARCH. Los modelos GARCH tienen una estructura que recuerda a los modelos ARMA, compartiendo con ellos muchas de sus propiedades.

4.2. Modelos de regímenes cambiantes

Otro tipo de modelos que ha interesado a los analistas son los llamados modelos de regímenes cambiantes o modelos por umbrales (Tong, 1990). Estos modelos son lineales por tramos. El paso de un modelo a otro puede venir regido por los valores de cierta variable. El ejemplo más sencillo serían los denominados modelos SETAR (self-exciting threshold autoregressions). Un ejemplo muy sencillo de modelo SETAR sería utilizando un modelo AR(1) con sólo dos regímenes:

$$z_t = \begin{cases} \phi_1 z_{t-1} + a_t; & z_{t-d} < r, \\ \phi_2 z_{t-1} + a_t & z_{t-d} \geq r. \end{cases}$$

La dificultad de la modelización estriba en la identificación de los valores de d y r, siendo un área de investigación abierta. Existen muchas variantes de modelos por umbrales, como aquellos que permiten transiciones suaves (modelos STAR).

4.3. Modelos de parámetros cambiantes

Siguiendo la generalización de la descomposición de Wold, una forma de modelizar la evolución no lineal de una serie temporal es mediante la modelización ARMA con parámetros que cambien con el tiempo. Este tipo de modelos recibe también el nombre de modelos con coeficientes funcionales. Por ejemplo, un modelo AR(1) con parámetros cambiantes sería

$$z_t = \phi_{0,t} + \phi_{1,t} z_{t-1} + a_t,$$

donde la complejidad de la modelización depende de la estructura temporal que queramos asumir para $(\phi_{0,t}, \phi_{1,t})$. Una forma sencilla de estimar la secuencia de valores de $(\phi_{0,t}, \phi_{1,t})$ es mediante la estimación recursiva utilizando mínimos cuadrados adaptativos (Sánchez, 2006). Este tipo de estimación resulta adecuada cuando $(\phi_{0,t}, \phi_{1,t})$ tiene una evolución suave con el tiempo. En caso contrario, otras formas de estimación serían más apropiadas.

En los casos en los que la evolución de $(\phi_{0,t}, \phi_{1,t})$ pueda relacionarse con otra variable, como por ejemplo una variable exógena, o retardos de z_t, puede estimarse de forma no paramétrica de manera análoga a los modelos no paramétricos de regresión, como los basados en un estimador Kernel (ver, por ejemplo, Vilar-Fernandez y Cao, 2007, y las referencias que contiene), o a los modelos aditivos generalizados (Chen y Tsay, 1993). De esta forma, muchos métodos estadísticos no paramétricos desarrollados para los modelos de regresión son aplicables al caso dinámico. Surgen, sin embargo, numerosos problemas relacionados con la dependencia de las observaciones. Uno de ellos es la selección del parámetro de suavizado para los estimadores Kernel. En el caso de observaciones independientes, el método de cross-validation es una herramienta habitual para seleccionar dicho parámetro. En el caso de observaciones dependientes, el método de cross-validation ya no es directamente aplicable, y los métodos existentes para su adaptación a series temporales requieren eliminar muchas observaciones.

5. Áreas de investigación futuras

La evolución de la Estadística ha estado siempre marcada por las aplicaciones que desde otros campos se ha hecho de ella. Es por tanto razonable pedecir que el futuro de las series temporales avanzará en paralelo a los avances en otros campos. Es fácil ver cómo muchas líneas de investigación, como ocurre con los modelos no lineales basados en datos funcionales, van de la mano de las mayores facilidades computacionales. De esta forma, métodos estadísticos no lineales que hace unos años tenían un interés minoritario se convertirán en herramientas habituales de muchos analistas. Las oportunidades que se presentan entonces para el desarrollo de los modelos no lineales son claras. Por otra parte, la mayor facilidad para obtener series de alta frecuencia, como ocure en el campo de las telecomunicaciones, internet, o finanzas, seguirán demandando el desarrollo de investigaciones en procesos de larga memoria y modelos sobre la volatilidad condicionada. Asímismo, el mayor empleo de grandes bases de datos tanto por parte de instituciones públicas como privadas, seguirá impulsando el campo de las series multivariantes así como de técnicas de reducción de la dimensión. Un campo en sus inicios es el de modelos vectoriales no lineales, donde es esparable avances importantes en el futuro. Los problemas de clasificación y discriminación de series

temporales son también objeto de numeros trabajos recientes. Un factor de desarrollo importante para este área serán las aplicaciones en Internet. Para la ampliación de estas ideas y un punto de vista complementario sobre el desarrollo de otras lineas de investigación remitimos al lector al trabajo de revisión del campo de las series temporales de Tsay (2000).

Referencias

[1] Box, G.E.P. y Jenkins, G.M. (1976). *Time Series Analysis: Forecasting and Control*, San Francisco: Holden-Day.

[2] Chen, R. y Tsay, R. (1993). Functional Coefficient Autoregressive Models, *Journal of the American Statistical Association*, **88**, 298-308.

[3] Durbin, J. y Koopman, S.J. (2001). *Time Series Analysis by State Space Methods*. Oxford University Press. New York.

[4] Engel, R.F. (1982). Autoregressive conditional heteroscedasticity with estimates of the varianza of U.K. inflation, *Econometrica*, **50**, 987-1008.

[5] Granger, C.W.J. y Joyeux, F. (1980). An introduction to long-memory time series models and fractional differencing, *Journal of Time Series Analysis*, **1**, 15-29.

[6] Lutkepohl, H. (1993). *Introduction to Multiple Time. Series Analysis*. Springer Verlag, Berlin.

[7] Pascual, L., Romo, J, y Ruíz, E. (2001). Effects of Parameter Estimation on Prediction Densities: A Bootstrap Approach, *International Journal of Forecasting*, **17**, 83-103.

[8] Peña, D. (1995). Forecasting Growth with Time Series, *Journal of Forecasting,* **14**, 97-105.

[9] Peña, D. (2005): *Analisis de series temporales*. Alianza Universidad.

[10] Peña, D. y Poncela (2006). Nonstationary Dynamic Factor Analysis, *Journal of Statistical Planning and Inference*, **136**, 1237-1257.

[11] Peña, D. y Sánchez, I. (2005). Multifold Predictive Validation in ARMAX Time Series Models, *Journal of the American Statistical Association*, **100**, 135-146.

[12] Sánchez, I. (2006). Recursive estimation of dynamic models using Cook's distance, with application to wind energy forecast, *Technometrics*, **48**, 61-73.

[13] Tong, H. (1990). *Non-linear time series: a dynamical system approach*. Clarendon Press. Oxford.

[14] Tsay, R.S. (2000). Time Series and Forecasting: Brief History and Future Research, *Journal of the American Statistical Association*, **95**,638-43.

[15] Vilar-Fernández, J.M. y Cao, R. (2007). Nonparametric forecasting in time series. A comparative study, *Communications in Statistics: Simulation and Computation*, **36**, 311-334.

[16] West, K.D. (1996). Asymptotic Inference about Predictive Ability, *Econometrica*, **68**, 1084-1097.

[17] West, M. y Harrison, P.J. (1989). *Bayesian Forecasting and Dynamic Models*. Springer-Verlag. New York.