

Metodología para la predicción con modelos SARIMA Bayesianos

Daniel Dala
Departamento de Estadística
Universidad Nacional Autónoma de Honduras
e-mail: daniel.dala@unah.hn

ÍNDICE

I.	Introducción	1
II.	Definición del Problema	1
III.	Preliminares y Notación	1
III-A.	Modelos ARIMA no estacionales	2
III-B.	Modelos ARIMA estacionales	2
III-C.	Modelos SARIMA Bayesianos	2
	Referencias	2

ÍNDICE DE FIGURAS

1.	Metodología Box-Jenkins	1
----	-----------------------------------	---

ÍNDICE DE CUADROS

Metodología para la predicción con modelos SARIMA Bayesianos

Resumen—El resumen no solo hace referencia al trabajo reportado, también sintetiza el trabajo documentado en aproximadamente 200 palabras. Establece el propósito, reporta la información obtenida, provee conclusiones, y recomendaciones. En esencia, resume los puntos principales del estudio de forma adecuada y precisa. Es importante referirse a los resultados principales obtenidos en tiempo pasado al describir el trabajo realizado.

I. INTRODUCCIÓN

En la introducción se describe de forma concisa pero con un poco más de detalle el trabajo realizado.

En uno de los párrafos de esta sección se hace especial énfasis en la contribución realizada como parte del trabajo reportado, considerando como punto de partida el problema central que motivo el proyecto de investigación y las soluciones obtenidas por el autor del documento como resultado del trabajo de investigación realizado.

Se presentan además descripciones breves del resto de las secciones del documento.

II. DEFINICIÓN DEL PROBLEMA

Sea $\{y_1, y_2, \dots, y_n\}$ una serie de tiempo cualquiera, generalmente se busca predecir los valores futuros y_{n+m} con $m = 1, 2, \dots, h$ que son desconocidos, para esto existen diferentes metodologías como modelos de espacio y estado [10], Prophet [11], redes neuronales [12], splines [13], procesos Gaussianos [14] entre otros. Algunos de los modelos más populares son los modelos SARIMA [3], [4] por su fácil interpretación y alta capacidad predictiva, pero su implementación con datos reales es compleja debido a que seleccionar el orden del modelo es una tarea complicada. Por consiguiente, Box y Jenkins (1970) [2] propusieron una metodología para una adecuada estimación de dichos modelos, la cual se basa en cinco etapas iterativas: revisar los datos, seleccionar modelo, estimación de parámetros, diagnóstico, selección de modelo, y predicción. Dicha metodología se ilustra en la Figura 1.

En los últimos años los métodos bayesianos se han vuelto una alternativa muy utilizada para el análisis de datos con muchas aplicaciones en economía, física, química, psicología, entre otras. Su creciente popularidad se debe a su capacidad de incorporar información externa al modelo mediante una distribución a priori, y actualizar las creencias mediante el Teorema de Bayes. Este esquema de inferencia en la práctica es muy complicado, por lo cual en los últimos años se han aproximado los resultados mediante los métodos de Markov Chain Monte Carlo [15]. Estos métodos consisten en generar una cadena de markov cuya distribución estacionaria es la distribución a posteriori del modelo, existen muchos procedimientos para implementar estos métodos, uno de los más

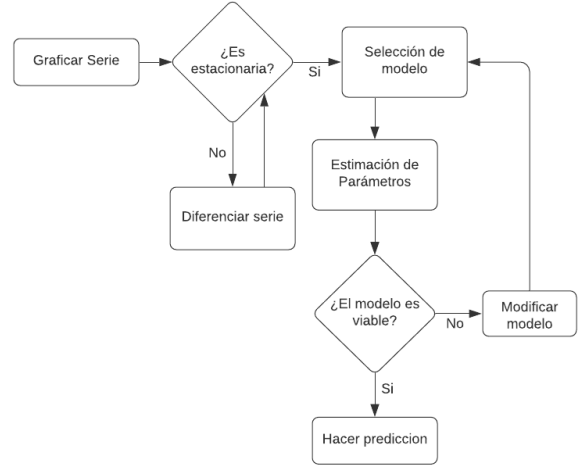


Figura 1. Metodología Box-Jenkins

comunes es el Monte-Carlo Hamiltoniano que por su flexible implementación en el lenguaje Stan ha sido de utilidad en múltiples aplicaciones.

Cabe destacar que la nueva metodología, si bien es cierto es análoga al método Box-Jenkins, cada componente usará herramientas detalladas del *Bayesian Workflow* propuesto en [5] el cual maneja bastante bien la estimación de parámetros de modelos y la incertidumbre en los datos.

Por último, una vez establecido el nuevo método se realizarán tres diferentes pruebas en tres conjuntos de datos que miden el IPC en Honduras de 1980 al 2018, la tasa de cambio de divisas entre Alemania y Reino Unido de 1984 a 1991 y la afluencia de turistas en Australia de 1995 al 2015, cada uno de estos conjuntos se encuentran en el paquete *bayesforecast* [9]. Finalmente con los resultados de dichas pruebas se demostrará la funcionalidad del nuevo método propuesto.

III. PRELIMINARES Y NOTACIÓN

En este estudio denotaremos las series de tiempo como $\{y_1, y_2, \dots, y_n\}$. Diremos que un proceso $\{x_i\}$ con $i \in \mathbb{Z}$ tiene la propiedad de estacionariedad fuerte si:

$$F_X(x_{t_1}, x_{t_2}, \dots, x_{t_n}) = F_X(x_{t_1+\tau}, x_{t_2+\tau}, \dots, x_{t_n+\tau})$$

para $t \in \mathbb{Z}$ con $n \in \mathbb{N}$ y cualquier $\tau \in \mathbb{Z}$, esto es, si para cualquier colección finita del proceso su distribución conjunta se mantiene constante en el tiempo. Por otro lado diremos que el proceso $\{x_i\}$ tiene la propiedad de estacionariedad débil si:

$$\mu(t) = \mu, \quad \sigma^2(t) = \sigma^2$$

esto es, que el proceso tiene una media y varianza constante a través del tiempo.

Diremos que una serie $\{y_t\}$ presenta estacionalidad si muestra alguna tendencia u oscilaciones periódicas constantes sobre la media del proceso, en este sentido para estabilizar la media de la serie y reducir la estacionalidad se aplica una transformación en los datos llamada diferenciación denotada por el operador diferencia:

$$\nabla^d y_t = y_t - y_{t-1}$$

Diremos que una serie es diferenciada si al aplicar el operador diferencia se vuelve estacionaria, de manera similar la serie es de diferencias estacionales si al aplicar la diferenciación estacional:

$$\nabla_s^D y_t = y_t - y_{t-s}$$

donde s es el periodo estacional, se vuelve estacionaria. También diremos que una serie de tiempo es un ruido blanco si es i.i.d. y normalmente distribuida con media cero y varianza constante.

III-A. Modelos ARIMA no estacionales

En el análisis de series de tiempo si combinamos los modelos autorregresivos, modelos de medias móviles y la diferenciación obtenemos los modelos ARIMA (Autoregressive Integrated Moving Average) no estacionales denotado por:

$$\text{ARIMA}(p, d, q)$$

en donde p, d y q son los órdenes del modelo autorregresivo, diferenciación y modelo de medias móviles respectivamente, de manera explícita se expresa como:

$$\nabla^d y_t = c + \sum_{i=1}^p \phi_i \nabla^d y_{t-i} + \sum_{j=1}^q \theta_j \varepsilon_{t-j} + \varepsilon_t$$

donde ε_t es ruido blanco.

III-B. Modelos ARIMA estacionales

Los modelos ARIMA pueden modelar datos con estacionalidad esto se puede lograr agregando parámetros estacionales al modelo y lo denotaremos como:

$$\text{SARIMA}(p, d, q) \times (P, D, Q)$$

SARIMA (Multiplicative seasonal autoregressive integrated moving average) de manera explícita:

$$Z_t = c + \sum_{i=1}^p \phi_i Z_{t-i} + \sum_{j=1}^q \theta_j \varepsilon_{t-j} + \sum_{k=1}^P \Phi_k Z_{t-km} + \sum_{w=1}^Q \Theta_w \varepsilon_{t-m} + \varepsilon_t$$

$$Z_t = \nabla_m^D \nabla^d y_t,$$

donde ε_t es un ruido blanco, los parámetros (p, d, q) y (P, D, Q) representan el orden de la parte no estacional y la parte estacional del modelo respectivamente y m es el periodo de oscilación de la media en los datos.

III-C. Modelos SARIMA Bayesianos

En base a la definición previa de un modelo ARIMA estacional podemos definir un Modelo SARIMA Bayesiano como:

$$\text{Modelo} \sim \text{SARIMA}(p, d, q) \times (P, D, Q)_m$$

$$\phi_i \sim \text{priori}_{\phi_i}, \quad i = 1, \dots, p$$

$$\theta_j \sim \text{priori}_{\theta_j}, \quad j = 1, \dots, q$$

$$\Phi_k \sim \text{priori}_{\Phi_k}, \quad k = 1, \dots, P$$

$$\Theta_w \sim \text{priori}_{\Theta_w}, \quad w = 1, \dots, Q$$

$$\mu_0 \sim \text{priori}_{\mu_0}$$

$$\sigma_0 \sim \text{priori}_{\sigma_0}$$

REFERENCIAS

- [1] Hyndman, R.J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2. Accessed on February 17, 2021.
- [2] Rob J. Hyndman (2002). *Box-Jenkins modelling*. In: Informed Student Guide to Management Science, ed., Hans Daellenbach and Robert Flood, Thomson: London. <https://robjhyndman.com/papers/BoxJenkins.pdf>
- [3] Rob J Hyndman (2002). *ARIMA processes*. In: Informed Student Guide to Management Science, ed., Hans Daellenbach and Robert Flood, Thomson: London. <https://robjhyndman.com/papers/ARIMA.pdf>
- [4] PennState Eberly College of Science. *Seasonal ARIMA Models*, <https://online.stat.psu.edu/stat510/lesson/4/4.1>
- [5] Andrew Gelman, Aki Vehtari, Daniel Simpson, Charles C. Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, Martin Modrák (2020). *Bayesian Workflow.*, <https://arxiv.org/abs/2011.01808>
- [6] Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). *Stan: A Probabilistic Programming Language*. Journal of Statistical Software, 76(1), 1 - 32. doi: <http://dx.doi.org/10.18637/jss.v076.i01>
- [7] Hyndman, R., & Khandakar, Y. (2008). *Automatic Time Series Forecasting: The forecast Package for R*. Journal of Statistical Software, 27(3), 1 - 22. doi: <http://dx.doi.org/10.18637/jss.v027.i03>
- [8] Robert H. Shumway, David S. Stoffer (2017). *Time Series Analysis and Its Applications* Fourth Edition.
- [9] Matamoros A., Cruz C., Dala A., Hyndman R., O'Hara-Wild M. (2021). *bayesforecast: Bayesian Time Series Modeling with Stan* version 1.0.1, <https://CRAN.R-project.org/package=bayesforecast>
- [10] J. Durbin & S. J. Koopman (2012). *Time Series Analysis by State Space Methods*, Second Edition.
- [11] Facebook's Core Data Science team (2017). *Prophet: Automatic Forecasting Procedure*, <https://github.com/facebook/prophet>
- [12] Bogdan Oancea and Ștefan Cristian Ciucu (2014). *Time series forecasting using neural networks*, <https://arxiv.org/abs/1401.1333>
- [13] Lotrič, Uroš and Dobnikar, Andrej (1991). *Artificial Neural Nets and Genetic Algorithms*, pag 121-126 Springer Vienna.
- [14] Roberts S., Osborne M., Ebdon M., Reece S., Gibson N. and Aigrain S. (2013). *Gaussian processes for time-series modelling*, Phil. Trans. R. Soc. A.3712011055020110550 <http://doi.org/10.1098/rsta.2011.0550>
- [15] Joshua S. Speagle (2020). *A Conceptual Introduction to Markov Chain Monte Carlo Methods*, <https://arxiv.org/abs/1909.12313>