PREDICCIÓN DE CASOS DE LEPTOSPIROSIS: MODELO PARAMÉTRICO VERSUS MODELO SEMIPARAMÉTRICO

Llop, María José[1]

[1] *CEVARCAM, Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral, Argentina*
Director/a: Müller, Gabriela
Codirector/a: López, María Soledad

Área: Ciencias Biológicas

Palabras claves: Leptospirosis, Predicción, Modelos.

INTRODUCCIÓN

Los eventos hidrometeorológicos extremos contribuyen a la propagación de enfermedades infecciosas como dengue, leptospirosis y meningitis. Dichos eventos cada vez más frecuentes e intensos como las precipitaciones extremas, inundaciones, entre otros, provocan impactos más complejos y de largo plazo en la salud de las personas. Particularmente, en las provincias de Santa Fe y Entre Ríos las características geográficas y climáticas favorecen el hábitat de la bacteria leptospira, que genera la enfermedad de leptospirosis. Estas provincias están rodeadas de ríos, y la exposición de las personas durante la inundación se considera uno de los principales factores de riesgo para la leptospirosis. Por lo tanto, es de suma importancia la estimación a futuro de la incidencia de la enfermedad para que los sistemas de salud puedan contener la epidemia, cortando o retardando la cadena de transmisión.
En este trabajo se propone estudiar métodos estadísticos de predicción, aplicarlos a datos epidemiológicos reales de leptospirosis y comparar su desempeño para concluir cual es el más conveniente de utilizar a la hora de predecir la incidencia de dicha enfermedad.

OBJETIVOS

- Estudiar métodos estadísticos predictivos paramétricos y semiparamétricos.
- Aplicar los métodos a datos reales para predecir la incidencia de leptospirosis.
- Evaluar y comparar el desempeño predictivo de los métodos.

METODOLOGÍA

Fuente de datos

En (López M.S., Müller G.V., Lovino M.A., Gómez A.A., Sione E.F.) y (López M.S., Müller G.V., Sione W., 2016) los autores detectaron las variables hidrometeorológicas que más influyen en lo brotes de leptospirosis. Con esa fundamentación, las variables de interés en este trabajo son: cantidad de caso registrados de leptospirosis, nivel hidrométrico máximo del río, precipitación media e

Título del proyecto: "Incidencia del cambio climático en la costa del Río Uruguay de la provincia de Entre Ríos y su impacto en la salud: estrategias de capacitación y empoderamiento de la población local"
Instrumento: PIO
Año convocatoria: 2019
Organismo financiador: CONICET
Directora: Gabriela Müller

índices oceánicos de El Niño y La Niña (ONI). Se cuenta con mediciones mensuales de dichas variables para cuatro ciudades ribereñas de las provincias de Santa Fe y Entre Ríos. Los datos corresponden a diez años consecutivos, desde enero de 2009 hasta diciembre de 2018, lo que significa un total de 120 observaciones para cada una de las variables. El origen de los datos es el siguiente: la información epidemiológica es brindada por la Dirección Provincial de Promoción y Prevención de la Salud - Ministerio de Salud de Santa Fe; los datos hidrometeorológicos de precipitación media son proporcionados por el Servicio Meteorológico Nacional (SMN) de Argentina; los datos de nivel hidrométrico máximo son proporcionados por el Instituto Nacional del Agua de Argentina (INA) y Prefectura Naval Argentina (PNA). Por último, los Índices Oceánicos de El Niño y La Niña (ONI) se obtienen a través de la página de web https://ggweather.com/enso/oni.htm.

Análisis

Los análisis estadísticos para la predicción de la incidencia de leptospirosis son considerablemente escasos. Uno de los métodos paramétricos que se utiliza generalmente para la predicción con series temporales de este tipo es el modelo $\mathrm{ARIMA(p,q,d)}$ (modelo autorregresivo integrado de medias móviles de orden p, q y d) que fue propuesto por Box y Jenkins en 1976. En la notación, p indica qué tan atrás en el tiempo se va usando valores previos de la variable de interés. De manera similar, q se refiere a cuántos retrasos se usan en el término de error y d indica cuántas veces se toma la diferencia de la variable dependiente. Si la variable de interés tiene un comportamiento estacional, se crea una nueva variable estacional que refleje esta variación, que es el valor actual de la variable dependiente menos el valor de un período estacional anterior. Luego se pueden aplicar diferencias y retrasos a esta variable e incluir estos términos en el modelo.

$$\phi(B)\nabla^d y_t = \theta(B) z_t \qquad (1)$$

En la ecuación (1) y_t es la serie de tiempo; z_t es el error aleatorio; $\phi(B)$ y $\theta(B)$ son los polinomios autorregresivo y de medias móviles de grados p y q, respectivamente, en el operador de retardos B que verifica $y_{t-k} = B^k y_t$ y $\nabla = (1 - \mathrm{B})$.

Si bien este método parece ser eficiente, solo utiliza la serie temporal a analizar y no propone la utilización de covariables que pueden influir en la respuesta y mejorar la precisión de la predicción. Por ese motivo, en este trabajo se propuso ajustar a los datos el modelo semiparamétrico SFPLR, en inglés “Semi-Functional Partial Linear Regression”, expresado en la ecuación (2), que fue propuesto por Aneiros-Pérez y Vieu en 2008. El mismo permite incorporar covariables hidrometeorológicas que afecten a la respuesta. Aunque este modelo no suele utilizarse para datos de estas características, se lo aplicó y se evaluó su desempeño para compararlo con el modelo ARIMA, que es comúnmente utilizado para este tipo de datos.

Para ajustar el modelo SFPLR es necesario, en principio, preparar los datos. Bajo la suposición de que la cantidad total de observaciones N de cada variable se puede escribir como $N = n\tau$, particionamos las observaciones de cada variable en n períodos de longitud τ, y consideramos esos períodos como datos funcionales.

Así, contamos con n datos funcionales observados para cada variable.

$$G(Y_{i+1}) = \sum_{j=1}^{p} X_{ij}\beta_j + m(Y_i) + \varepsilon_i \quad , i = 1 \ldots n \qquad (2)$$

En la ecuación (2) β_j son parámetros a ser estimados; m es una función a ser estimada; ε_i son errores aleatorios de media 0 e iid; X_{ij} son las variables independientes numéricas; Y_i es la

variable respuesta funcional y $G(Y_{i+1})$ es una característica numérica del período $i+1$, como por ejemplo la incidencia de leptospirosis en un mes determinado.
Los estimadores de $\boldsymbol{\beta}$ y m están dados en las ecuaciones (3) y (4).

$$\widehat{\boldsymbol{\beta}} = \left(\widetilde{\boldsymbol{X}}_h^T \widetilde{\boldsymbol{X}}_h\right)^{-1} \widetilde{\boldsymbol{X}}_h^T \widetilde{\boldsymbol{Y}}_h \tag{3}$$

$$\widehat{m}_h(t) = \sum_{i=1}^{n} w_{n,h}(t, Y_i)\left(Y_i - \boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}_h}\right) \tag{4}$$

donde h es el parámetro de suavizado, elegido mediante un proceso de validación cruzada. $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)^T$ es la matriz de covariables de dimensión $n \times p$ con $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T$, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ y para cualquier matriz $\boldsymbol{A}$ de $n \times p$ $\widetilde{\boldsymbol{A}}_h = (\boldsymbol{I} - \boldsymbol{W}_h)\boldsymbol{A}$, donde $\boldsymbol{W}_h = (w_{n,h}(Y_i, Y_j))_{ij}$ con $w_{n,h}(.,.)$ una función de pesos, en este caso utilizamos pesos del tipo Nadaraya-Watson.
Una vez implementados los métodos se realizaron predicciones de la serie temporal para un año consecutivo (2018) y posteriormente se compararon los resultados con los datos reales de ese año. El criterio utilizado para comparar los modelos estadísticos fue, por un lado de carácter exploratorio, es decir, se observó las gráficas para ver de que manera se ajustó la predicción a la tendencia de los datos reales, y por otro lado se calculó el error cuadrático medio (ECM) como se muestra en la ecuación (5) y se los comparó. Para ello se utilizó una muestra de entrenamiento, que contó con todos los datos de las covariables hidrometeorológicas y de la variable epidemiológica, excepto los correspondientes al último año, y otra de testeo que tuvo solamente los casos de leptospirosis en el último año observado. Así, se realizaron las predicciones del último año y luego se las comparó con los datos reales del mismo.
La implementación de los métodos anteriormente mencionados se realizó con el software estadístico R (R Core Team 2017).

$$ECM = \frac{\sum_{i=1}^{12}(y_i - \hat{y}_i)^2}{12} \tag{5}$$

donde y_i es el valor real de la serie temporal y $\hat{y}_i$ es el valor estimado.

CONCLUSIONES

Observando la Figura 1 se puede concluir que el modelo SFPLR capta la tendencia de los datos, mostrando brotes en los primeros meses de año. En cambio, el modelo ARIMA tiene un comportamiento más constante y no muestra los meses en donde podría haber brotes. Queda comprobado por el ECM de la Tabla 1 que el modelo SFPLR ajusta mejor.
En la Figura 2 se observa que el modelo SFPLR predice la mayor cantidad de brotes entre abril y julio, tendencia que se observa también en los datos reales. El modelo ARIMA vuelve a tener un comportamiento constante, por lo tanto el modelo SFPLR ajusta mejor y además su ECM es menor.
En la Figura 3 se observa que aunque el modelo ARIMA tiene un comportamiento constante ajusta mejor a los datos reales y tiene menor ECM, el modelo SFPLR predice brotes en algunos meses de la primer mitad del año, como sucedió en los casos anteriores. En la Figura 4 se observa que el modelo SFPLR predice la mayor cantidad de brotes en la primera mitad del año, que es aproximadamente la tendencia que tienen los datos, sin embargo presenta muchas oscilaciones en los siguientes meses, lo que incrementa el ECM.
Si bien en algunos casos el modelo SFPLR tiene un ECM mayor que el ARIMA, capta mejor la tendencia de los datos y predice los períodos donde puede haber brotes, en cambio el modelo ARIMA tiene un comportamiento constante que no aporta información sobre los posibles brotes. Por lo tanto, es más conveniente utilizar el modelo SFPLR que incluye las variables hidrometeorológicas, ya que las mismas aportan información sobre la variable de interés.

Figura 1: Santa Fe-2018

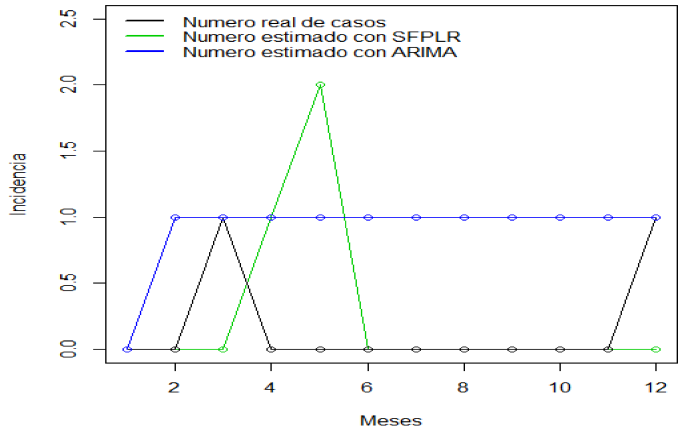

Figura 2: Paraná-2018

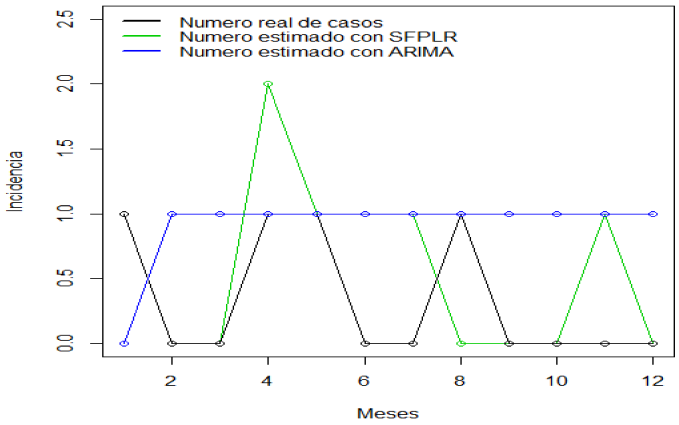

Figura 4: Rosario-2018

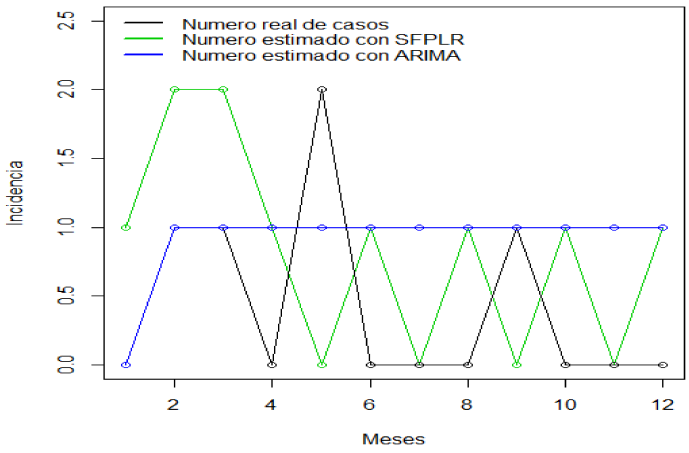

Figura 3: Gualeguaychú-2018

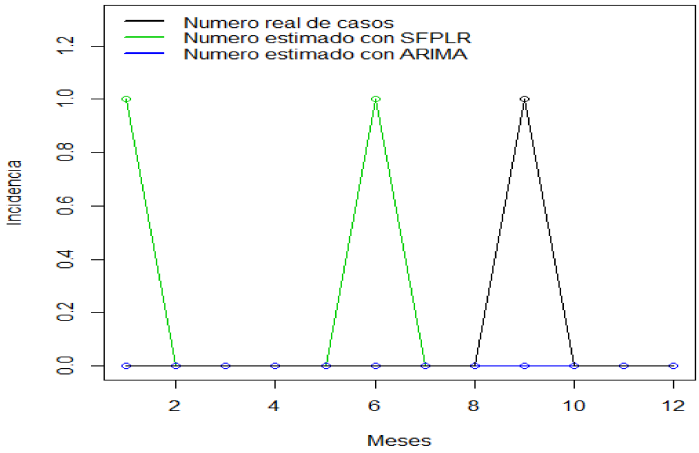

Tabla 1: Error Cuadrático Medio

	ECM ARIMA	***ECM SFPLR***
Santa Fe	*0.833*	*0.583*
Paraná	*0.75*	*0.416*
Rosario	*0.666*	*1.083*
Gualeguaychú	*0.083*	*0.25*

BIBLIOGRAFÍA BÁSICA

Aneiros-Pérez, G., Vieu, P., (2008). *Nonparametric time series prediction: a semi-functional partial linear modelling*. Journal Multivariate Anal, 99, 834-857.

Box, G., Jenkins, G., Reinsel, G., (1994). *Time Series Analysis: Forecasting and Control.* 3th ed. Canada: Prentiuce Hall Canada.

López M.S., Müller G.V., Lovino M.A., Gómez A.A., Sione E.F.. *Spatio-temporal analysis of leptospirosis incidence and its relationship with hydroclimatic indicators in northeastern Argentina*. En revisión. Science of the Total Environment.

López M.S., Müller G.V., Sione W., (2016). *Análisis de los casos de leptospirosis en el Noreste de Argentina y su relación con los eventos ENSO.* Libro de resúmenes de la XXVIII Reunión Científica de la Asociación Argentina de Geofísicos y Geodestas (AAGG 2017): Tercer Simposio sobre Inversión y Procesamiento de Señales en Exploración Sísmica (IPSES 17).