FACULTAD DE CIENCIAS ECONÓMICAS Y DE ADMINISTRACIÓN

IESTA INSTITUTO DE ESTADÍSTICA

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Modelos Bayesianos para series diarias:

Modelado de temperaturas extremas en Uruguay.

Manuel Hernández Banadik

Ignacio Alvarez-Castro

Natalia da Silva

Santiago de Mello

Serie Documentos de Trabajo

№4/21

Abril, 2021

ISSN: 1688-6453

Universidad de la República.
Facultad de Ciencias Económicas y de Administración,
Instituto de Estadística (IESTA)

Montevideo, Uruguay.

Forma de citación sugerida para este documento:

Hernández Banadik, M., Álvarez-Castro, I., da Silva, N., de Mello, S.(2021). *Modelos Bayesianos para series diarias: Modelado de temperaturas extremas en Uruguay* (Serie Documentos de Trabajo; №4/21). Montevideo: Universidad de la República. Facultad de Ciencias Económicas y de Administración, Instituto de Estadística.
https://www.colibri.udelar.edu.uy/jspui/handle/20.500.12008/10518

Modelos Bayesianos para series diarias. Modelado de temperaturas extremas en Uruguay.

Manuel Hernández Banadik [1]
Facultad de Ciencias Económicas y de Administración, Universidad de la República;
Ignacio Alvarez-Castro [2] Natalia da Silva [3]
Instituto de Estadística, Facultad de Ciencias Económicas y de Administración, Universidad de la República;
Santiago de Mello
Departamento de Ciencias de la Atmósfera, Facultad de Ciencias, Universidad de la República

RESUMEN

El estudio de eventos extremos ha tomado una gran relevancia en los últimos años debido principalmente al gran impacto que presentan en la sociedad y la economía de los países, así como en los ecosistemas. Para estudiar estos fenómenos es necesario contar con series temporales lo suficientemente largas y fundamentalmente completas, a un paso temporal de por lo menos un día y de alta calidad. Uruguay cuenta con registros suficientemente largos en ciertos puntos del país, pero se han detectado muchos períodos sin información. La modelización estadística de las series de temperatura observada es el primer paso para obtener bases de datos completas que permitan estudiar los eventos climáticos extremos. Este trabajo presenta métodos estadísticos para modelar series diarias multivariadas con importantes ventanas de datos faltantes, y también métodos de visualización de estas series que permitan explorar la presencia de secuencias de valores extremos. Se trabajó con datos de temperaturas mínimas y máximas en 11 estaciones meteorológicas de Uruguay durante un período de más de 60 años. Los resultados se presentan parcialmente en el documento y de forma completa en una aplicación web accesible en IESTA-INUMET. Este trabajo fue realizado en el marco del proyecto *Modelado de temperaturas extremas en Uruguay*, financiado por el Fondo Sectorial a partir de datos 2017 - ANII (FSDA_1_2017_1_144032).

Palabras clave: Modelos dinámicos Bayesianos, Series de tiempo Bayesianas, Datos faltantes. **CÓDIGOS JEL**: C11, C22
Clasificación MSC2010: 62F15, 62M20

[1] *email:* manuel.hernandez@fcea.edu.uy ORCID: 0000-0002-9046-6423
[2] *email:* ignacio.alvarez@fcea.edu.uy ORCID: 0000-0003-1633-2432
[3] *email:* natalia.dasilva@fcea.edu.uy ORCID: 0000-0002-6031-7451

ABSTRACT

Extreme events have a huge impact on society, economics and ecosystems, for this reason its study has become more relevant in recent years. Understanding extreme events in temperature series are based on high-quality, high-frecuency long-run temporal data. e.g. dayly records over several decades period. Uruguay has enough records for temperature extreme event study but these records are not complete, several periods with low-quality information has been detected. Thereofre, a statistical modelling for imputation is a first step in to generate complete, multivariate, long run datasets to allow study temperature extreme events. This work presents statistical methods robust to long missing data window in dayly multivariate time series data, and vizualisation methods to explore the presence of extreme value short secuences. Daily maximum and minimum temperature from 11 metheorogical stations in Uruguay over a 60 years period are used. Results are partially show in this document and fully available in a web app IESTA-INUMET. The present work was done within the research project *Modelado de temperaturas extremas en Uruguay* founded by the Fondo Sectorial a partir de datos 2017 - ANII (FSDA_1_2017_1_144032).
Key words: Dynamic Bayesian Model, Bayesian Time Series, Missing data

JEL CODES: C11, C22

Mathematics Subject Classification MSC2010: 62F15, 62M20 .

1. Introducción

El estudio de eventos extremos ha tomado una gran relevancia en los últimos años debido principalmente al gran impacto que presentan en la sociedad y la economía de los países, así como en los ecosistemas. Dentro de la región del Sudeste de Sudamérica los principales eventos climáticos extremos que han sido analizados son los relacionados con la temperatura (olas de calor y de frío, heladas, días cálidos, etc) y a la precipitación (sequías, precipitaciones intensas, etc). Muchos de estos estudios son relativamente actuales, debido principalmente a que es necesario contar con series temporales lo suficientemente largas y fundamentalmente completas, a un paso temporal de por lo menos un día y de alta calidad.

Bajo el escenario de cambio climático, es necesario comprender cómo los eventos extremos cambian en frecuencia y/o intensidad, identificando cómo se han comportado en las últimas 3 o 4 décadas. Identificar este tipo de eventos ayudará a comprender su dinámica para luego realizar previsiones a un plazo menor asociado a la variabilidad climática y hasta incluso estudiar el comportamiento a escala sinóptica o de días.

Uruguay cuenta con registros suficientemente largos en ciertos puntos del país, pero se han detectado muchos períodos sin información, tanto para temperaturas extremas diarias como para la variable precipitación (Amiel, 2012; Renom Molina, 2009). En los trabajos de Renom Molina (2009) y De Mello (2013) se identificaron gran cantidad de días sin el dato correspondiente de temperatura con varias periodicidades. Desde períodos prolongados sin datos, llegando a varios años, hasta datos puntuales sin registros, pasando por todos los períodos intermedios. Esto representa un gran inconveniente principalmente para el estudio de eventos extremos, ya que por definición presentan una baja probabilidad de ocurrencia, por lo que no contar con información completa es posible que se pierdan eventos.

La modelización estadística de las series de temperatura observada es el primer paso para obtener bases de datos completas que permitan estudiar los eventos climáticos extremos. Los modelos Bayesianos dinámicos son una herramienta muy potente, flexible y bien establecida para la modelización de fenómenos temporales que contienen datos faltantes (West y Harrison, 1996). El modelo puede ser descrito en forma general por una primer ecuación que describe el proceso observado condicional a variables latentes o parámetros (no observados) y una segunda ecuación que describe la evolución de las variables latentes. Es decir, la dependencia se modela a través de las variables latentes lo que vuelve estos

modelos muy flexibles.

Hay dos características que hacen atractivos a los modelos Bayesianos dinámicos para los desafíos planteados en este proyecto. En primer lugar, el tratamiento natural de la información faltante. El problema Bayesiano de inferencia consiste en obtener la distribución posterior de las variables latentes que determinan la evolución del proceso condicional a los datos observados. Esto se realiza secuencialmente con la información que se tiene hasta cada momento de tiempo. En cada paso, la información pasada se incorpora en la distribución previa y es combinada con la información actual. Cuando el dato del momento actual no se encuentra disponible simplemente la distribución posterior es igual a la distribución previa de interés, ya que no hay información nueva (Prado y West, 2010). En segundo lugar, es relativamente simple descomponer la serie temporal en varios componentes y con variables latentes separadas para cada uno de ellos. Dentro de la familia de modelos que forman los modelos Bayesianos dinámicos, la estrategia más simple consiste en trabajar cada estación meteorológica por separado y establecer una dinámica lineal en cada serie mediante el uso de modelos dinámicos lineales. Esto puede extenderse para incorporar la información espacial (del resto de las estaciones) en modelos jerárquicos y también es posible incorporar evoluciones no lineales de los datos (West y Harrison, 1996; Prado y West, 2010).

El resto del documento se organiza de la siguiente manera. En la próxima sección, se presenta el conjunto de datos y una breve descripción de las características relevantes para este trabajo. Luego, en la sección 2 se presenta un desarrollo de los modelos lineales dinámicos y su estimación Bayesiana siguiendo el tratamiento presentado en Petris *et al.* (2009), acompañada con un ejemplo de implementación usando la biblioteca `dlm` (Petris, 2010) en base a datos simulados. La sección 2 concluye con el modelo DLM concreto utilizado para modelar la series diarias multivariadas con períodos sin información. Las dos secciones finales del trabajo muestran los resultados y comentarios de cierre.

Todos los cálculos y estimaciones presentadas en el trabajo se realizan en `R` (R Core Team, 2020), usando `dlm` para los modelos, `ggplot2` (Wickham, 2016) para los gráficos y `shiny` (Chang *et al.*, 2020) para la presentación dinámica de los resultados.

2. Modelos lineales dinámicos

Los modelos lineales dinámicos (DLM) son una familia de modelos que constituye una subclase de los llamados modelos de espacio de estado. En DLM se considera una serie de tiempo, Y_t, como una observación ruidosa de un proceso dinámico no observable θ_t. Estos modelos se definen a través de dos ecuaciones básicas: una *ecuación de evolución* y una *ecuación de observación*. La estructura general de un DLM se muestra en (1), Y_t es un vector m-dimensional y θ_t un vector p-dimensional, luego F_t G_t son dos matrices **conocidas** de dimensiones $m \times p$ y $p \times p$ respectivamente que caracterizan al modelo. Las perturbaciones v_t y w_t son son secuencias de vectores Gaussianos independientes, mientras que V_t, W_t serán los parámetros que caractericen el modelo. Adicionalmente, se agrega una previa Normal para el momento inicial, es decir $\theta_0 \sim N(m_0, C_0)$.

$$\begin{array}{ll} Y_t = F_t\theta_t + v_t & v_t \sim \mathcal{N}(0, V_t) \\ \theta_t = G_t\theta_{t-1} + w_t & w_t \sim \mathcal{N}(0, W_t) \end{array} \tag{1}$$

La ecuación de evolución responde a la dinámica del proceso estocástico: define cómo se vincula el estado del proceso en tiempo t con su estado en tiempo $t+1$. Por otro lado, la ecuación de observación, explicita cómo es observado el proceso en tiempo t dado el estado en el que se encuentra a tiempo t. En ambas ecuaciones, se puede introducir una componente aleatorio.

2.1. Estructura de un DLM

La estructura de dependencia entre las observaciones es un aspecto clave para entender como funcionan los DLM. Al igual que la clase más general de modelos espacio-estado en los DLM se basa en dos supuestos fundamentales. En primer lugar se asume que el proceso inobservable o proceso de estado θ_t constituye una cadena de Markov. En tanto, las observaciones Y_t son condicionalmente independientes dado θ_t y se asume que Y_t depende únicamente de θ_t.

La Figura 1 representa un diagrama de la estructura de dependencia en DLM, donde los arcos indican dependencia estadística. Se puede ver como θ_t, toda la dependencia entre

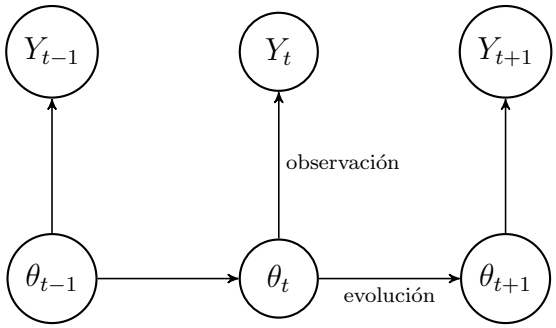

Figura 1: Diagrama de evolución y observación del proceso

observaciones pasa a través del proceso latente. A partir de esta estructura, se puede considerar que las observaciones representan una medida ruidosa del proceso de estado.

La principal consecuencia de la estructura de dependencia es que la distribución conjunta de todas las variables de interés queda determinada por la distribución previa del estado inicial y las densidades condicionales $p(\theta_t|\theta_{t-1})$, $p(y_t|\theta_t)$, de forma:

$$p(\theta_{0:t}, y_{1:t}) = p(\theta_0) \prod_{j=1}^{t} p(\theta_j|\theta_{j-1}) p(y_j|\theta_j)$$

donde la notación $z_{1:t}$ se utiliza para referirse a las observaciones $z_1, z_2, \ldots, z_t$.

La familia de DLM es amplia y flexible, muchos modelos estadísticos para datos temporales pueden verse como casos particulares de esta familia o de la clase más general de modelos espacio-estado. En particular los modelos ARMA con ruido Gaussiano son equivalentes a un DLM cuando los parámetros F_t, G_t, V_t, W_t son constantes en el tiempo. Por ejemplo, el proceso AR(2): $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + e_t$, con $e_t \sim N(0, \sigma^2)$ independientes, puede ser visto como un DLM si definimos $\theta_t = \begin{pmatrix} \theta_{1,t} \\ \theta_{2,t} \end{pmatrix}$ y luego,

$$y_t = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \theta_t = \theta_{1,t}$$

$$\theta_t = \begin{pmatrix} \theta_{1,t} \\ \theta_{2,t} \end{pmatrix} = \begin{bmatrix} \phi_1 & 1 \\ \phi_2 & 0 \end{bmatrix} \begin{pmatrix} \theta_{1,t-1} \\ \theta_{2,t-1} \end{pmatrix} + \begin{bmatrix} e_t \\ 0 \end{bmatrix}$$

para ver la equivalencia entre ambas representaciones, debemos operar en la segunda ecuación, que representa la evolución del proceso latente y verificar que se verifica que

$\theta_{1,t} = \phi_1\theta_{1,t-1} + \phi_2\theta_{1,t-2} + e_t$. Detalles sobre la correspondencia entre modelos ARIMA(p, d, q) y DLM están tratados en Petris *et al.* (2009) (sección 3.2.5).

Aparte de DLM, la familia de modelos espacio-estado admite modelos no lineales y el uso de otras distribuciones no gaussianas. Por ejemplo, modelos para las distribuciones de la familia exponencial pueden ser especificados para tener DLM *generalizados*, por otro lado en casos que el proceso de estados tenga una distribución discreta el modelo es referido como modelos ocultos de Markov, también modelos de volatilidad estocástica pueden ser vistos como espacio-estado. El estudio de DLM entonces, es relevante ya que permite introducirnos al trabajo con una familia de modelos que es muy amplia y puede aplicarse en una gran variedad de problemas concretos.

2.2. Estimación del proceso de estados

Si se tienen las observaciones $y_1, \ldots, y_T$, y se asume un modelo como el descrito en (1), hay varias distribuciones posteriores que puede ser de interés. En relación al proceso no observable se distinguen tres distribuciones relevantes condicional en los datos observados: filtrada, suavizada y de predicción.

La **distribución filtrada** (filtering distribution en inglés) consiste en la distribución del estado no observable θ_t dada la información observada hasta ese mismo momento, es decir $p(\theta_t|y_{1:t}$, se tiene una distribución para cada momento t. La posterior filtrada es de principal interés en aplicaciones secuenciales donde es necesario realizar una nueva estimación para cada observación nueva. Si se trabaja en el caso en que operan las restricciones de normalidad y linealidad, es posible tener soluciones cerradas para $p(\theta_t|y_{1:t})$ mediante el Filtro de Kalman. El filtro de Kalman puede verse como la resolución recursiva de un problema de estimación Bayesiana en el modelo lineal Gaussiano, distribución posterior de interés es $p(\theta_t|y_{1:t})$, que se obtiene como

$$p(\theta_t|y_{1:t}) = \frac{p(y_t|\theta_t)p(\theta_t|y_{1:t-1})}{p(y_t|y_{1:t})}$$

En primer lugar, suponemos conocida la distribución filtrada del paso anterior, $p(\theta_{t-1}|y_{1:t-1}) = N(m_{t-1}, C_{t-1})$. Luego, como la ecuación de evolución del proceso latente implica que $p(\theta_t|\theta_{t-1} = N(G_t\theta_{t-1}, W_t)$ se puede obtener la distribución previa $p(\theta_t|y_{1:t-1}$ como en (2). Es simple mostrar que el resultado es que la previa para el estado θ_t es normal,

$\theta_t|y_{1:t-1} \sim N(a_t, R_t)$, donde $a_t = G_t m_{t-1}$ y $R_t = G_t C_{t-1} G_t^\top + W_t$.

$$p(\theta_t|y_{1:t-1}) \int p(\theta_t|\theta_{t-1}) p(\theta_{t-1}|y_{1:t-1}) d\theta_{t-1} \tag{2}$$

Por otro lado, la verosimilitud $p(y_t|\theta_t)$ es también normal, esto surge directamente de (1) en que se explicita que $y_t|\theta_t \sim N(F_t\theta_t, V_t)$. Con estos ingredientes, modelo de datos y previa gaussianas, podemos obtener la posterior de interés aplicando resultados conocidos de modelo de regresión lineal conjugado para obtener el resultado del filtro de Kalman como

$$\begin{gathered}\theta_t|y_{1:t} \sim N(m_t, C_t) \\ m_t = C_t(R_t^{-1} a_t + F_t^\top V_t^{-1} y_t) \\ C_t = (R_t^{-1} + F_t^\top V_t^{-1} F_t)^{-1}\end{gathered}$$

Utilizando la previa normal para el momento inicial (θ_0) se procede iterativamente para obtener la distribución filtrada para todos los estados latentes θ_1 a θ_T.

Una segunda distribución posterior que puede tener relevancia en aplicaciones es la **distribución suavizada** (*smoothing distribution* en inglés) del proceso de estados. La distribución suavizada consiste en $p(\theta_t|y_{1:T}$, es similar a la posterior filtrada con la diferencia que el condicional se realiza sobre todos los datos observados disponibles. La distribución suavizada es relevante para aplicaciones donde estamos interesados en un estudio retrospectivo y se desea analizar todo el proceso de estados latente. Finalmente, en caso que estemos interesados en predecir un valor futuro del estado se utilizan distribuciones de predicción $p(\theta_{t+k}|y_{1:t}$. Para obtener las distribuciones posteriores de suavizado y predicción se utilizan recursiones similares a las descritas arriba para el filtro de Kalman.

Para finalizar con la descripción de cómo estimar modelos del tipo (1), se comentan dos aspectos adicionales: la estimación de los hiperparámetros y el tratamiento de datos faltantes. La aplicación del filtro de Kalman y su adaptación para obtener posteriores de suavizado y predicción como fue descrito arriba supone **conocidas** las matrices de varianza V_t y W_t. En la práctica, obviamente es necesario estimarlas. Hay varias opciones para hacer esta estimación, se puede incorporar una distribución previa y realizar una estimación Bayesiana o maximizar la verosimilitud marginal integrando el efecto de θ_t. Ambas alternativas pueden ser muy costosas computacionalmente cuando hay muchos datos. El otro aspecto que es importante destacar es que los modelos DLM incorporan *naturalmente* el tratamiento de datos faltantes. Por ejemplo, en el caso de la distribución filtrada, $\theta_t|y_{1:t} \sim N(m_t, C_t)$, supongamos que el dato y_t no se encuentra disponible, entonces no hay información nueva en la que condicionar, $p(\theta_t|y_{1:t}) = p(\theta_t|y_{1:t-1})$, es decir la distribución de θ_t antes de observar el dato y_t y su posterior son iguales.

2.3. Un ejemplo simulado

En lo que sigue se ilustra la estimación de DLM mediante la biblioteca `dlm` que permite trabajar con toda la familia de modelos lineales dinámicos.

Se utilizan datos simulados de un proceso simple, la ecuación (3) muestra el modelo usado para simular datos, en este caso G_t y F_t se definieron constantes con valor 1. El proceso de estado $\{\theta_t\}$ describe un paseo al azar, mientras que podemos considerar a $\{y_t\}$ como una observación ruidosa de $\{\theta_t\}$. El objetivo será recuperar los valores de los parámetros de varianzas y el proceso de estado.

$$\begin{aligned} Y_t &= \theta_t + v_t \qquad & v_t &\sim \mathcal{N}(0, 0.1^2) \\ \theta_t &= \theta_{t-1} + w_t \qquad & w_t &\sim \mathcal{N}(0, 0.3^2) \end{aligned} \tag{3}$$

El siguiente código de `R` realiza la simulación de un conjunto de datos con el modelo descrito en (3) y su estimación con el Filtro del Kalman.

```
# Obtenemos estados y datos simulados
set.seed(1237); N <- 100

dd <- tibble(t = 1:100, #seq(0,1, length.out = N),
             w = rnorm(N,0,.1),
             v = rnorm(N,0,.3) ) %>%
  mutate( theta = cumsum(w), y = ts(theta + v) )

# Función para estimar varianzas por Max.Ver
parMLE <- function(par) dlmModPoly(1, dV = par[1], dW = par[2])
modPoly <- dlmMLE(dd$y,parm = c(1,1),build = parMLE)

# Kalman filter y Kalman smoother
mod_filter <- dlmFilter(dd$y, parMLE(modPoly$par))
mod_smooth <- dlmSmooth(mod_filter)
```

En primer lugar, se debe estimar los parámetros desconocidos del modelo con la función `dlmMLE()`, que aproximará una estimación óptima, con un algoritmo iterativo. Para ello,

primero debemos crear una función auxiliar (`parMLE()`), cuyo argumento sea un vector de valores iniciales propuesto para los parámetros que se desea estimar, en nuestro caso desconocemos V_t y W_t, la varianza del ruido de evolución y la varianza del ruido de observación. El objeto `modPoly` contiene el modelo con los parámetros estimados.

Una vez estimado el modelo, podemos aplicar el filtro de Kalman mediante (`dlmFilter()`) para obtener las distribuciones posteriores de filtro y predicción a un paso, así como la función (`dlmSmooth()`) para obtener la distribución de suavizado de Kalman. Notemos que dado que todas estas distribuciones posteriores son gaussianas, alcanza con tener las estimaciones puntuales de esperanza y varianza posterior para cada momento del tiempo. En particular, las series de tiempo `m`, `a` y `f` contienen la estimación del estado, la predicción del estado a un paso y la predicción de las observaciones a un paso, respectivamente. Adicionalmente, la función `dlmSmooth()` aplica el suavizado de Kalman y nos devuelve la estimación suavizada del proceso de estados contenida en la serie `s`.

La Figura 2 muestra los resultados de la estimación modelo, en cada panel se muestran los datos observados (serie gris), el verdadero valor del estado θ_t (serie negra) y una una estimación del modelo (serie roja). Los tres tipos de estimación comentados previamente corresponden a cada panel de la figura, en el panel superior se muestra la estimación filtrada que utiliza datos hasta el momento en que se quiere estimar, es decir $m_t = \mathbb{E}(\theta_t|y_1, \ldots, y_t)$. El panel intermedio muestra la predicción a un paso del proceso de observación, $f_t = \mathbb{E}(y_t|y_1, \ldots, y_{t-1})$, y finalmente, el en el panel inferior se muestra la estimación suavizada del estados, utilizando toda la información disponible para obtenerla, es decir $s_t = \mathbb{E}(\theta_t|y_1, \ldots, y_N)$

Anteriormente, comentamos que la estimación Bayesiana de filtros y suavizado de Kalman incorpora naturalmente la presencia de valores faltantes en la serie. En el proceso iterativo de actualización de m_t y C_t para cada nuevo dato que se considera, un valor faltante simplemente implica que no es necesaria realizar una actualización, es decir $m_t = m_{t-1}$ cuando el dato y_t no esta presente. En lo que sigue mostramos como el código para realizar la estimación del proceso latente es el mismo en caso que la serie presente valores faltantes.

```
# Simular valores faltantes
falta = sample(1:N, 20)
dd$yna = dd$y
dd$yna[falta] <- NA

# Ajustar modelo: estimacion por MV, filtro y suavizado de Kalman
modPoly.na <- dlmMLE(dd$yna, parm = c(1,1),build = parMLE)
```

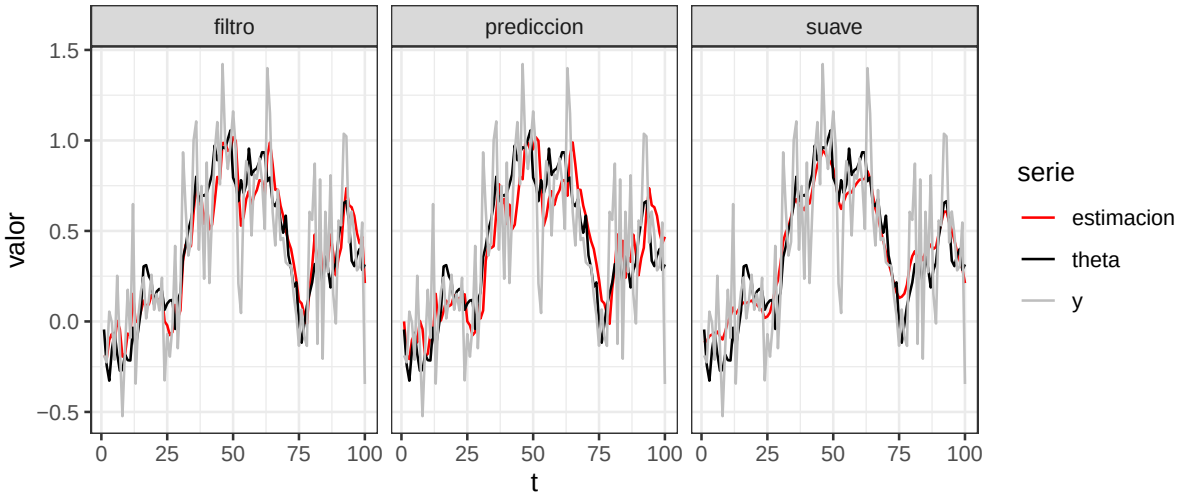

Figura 2: Ejemplo de un proceso de espacio de estado y su estimación.

```
filter.na <- dlmFilter(dd$yna, parMLE(modPoly.na$par))
smooth.na <- dlmSmooth(mod_filter)
```

La figura 3 es análoga a la anterior pero cuando la serie tiene valores faltantes. Únicamente se muestra el suavizado de Kalman. En este ejemplo sencillo, se ve como el modelo logra reconstruir el proceso latente en presencia de un 20 % de valores faltantes.

3. Series diarias de temperatura

El modelo que se describe en 2 es útil para modelar series de tiempo multivariadas con observaciones faltantes que pueden aparecer en forma consecutiva. Los datos específicos utilizados en el trabajo son series de temperatura en Uruguay.

La temperatura diaria, mínima y máxima es relevada en once estaciones meteorológicas del país, la Figura 4 muestra la ubicación geográfica de las mismas. Se cuenta con información de la temperatura máxima y mínima diaria registradas en 11 estaciones meteorológicas uruguayas para el período 1951 a 2014. Definimos $y_t^n = (y_{1t}^n, \ldots, y_{Dt}^n)^\top$ como el vector de temperaturas mínimas observadas en $D = 11$ estaciones meteorológicas en el día t,

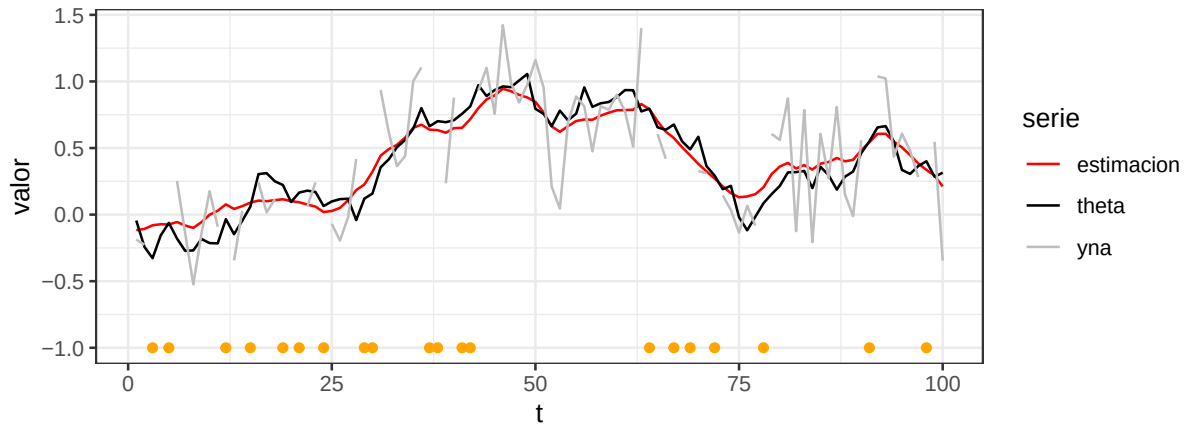

Figura 3: Estimación de proceso de estados latentes en presencia de valores faltantes.

Figura 4: Ubicación de las estaciones meteorológicas

mientras que y_t^x se refiere a las temperaturas máxima.

Debido a que uno de los problemas relevantes en este trabajo tiene que ver con la falta de información en las series, se presenta una breve descripción de los patrones de datos faltantes que aparecen en los datos. La Figura 5 muestra el porcentaje de datos faltantes por estación representado con una barra horizontal y el color indica si se refiere a la temperatura máxima o mínima, las estaciones están ordenadas de mayor a menor proporción de días faltantes. Observamos que hay 4 estaciones (Artigas, Melo, Prado, Rivera) en las que los días sin dato superan el 10 %, y en Artigas algo más que el 20 %, mientras que en el resto de las estaciones la cantidad de datos faltantes no supera el 5 % de el periodo. A su vez, no se ven grandes discrepancias entre el total de datos faltantes en máximas y mínimas para cada estación.

Un problema adicional tiene que ver con los días con falta de dato *consecutivos*. A modo de ejemplo, se considera la estación Paysandú que presenta apenas 72 días sin información sobre la temperatura mínima concentra esos muestra una ventana datos faltantes de 40 días consecutivos. La Figura 6 presenta la serie de temperatura mínima en la estación Paysandú en todo el año 1962, se observa que desde el 1 de Abril al 10 de Mayo de 1962 no existe datos sobre la temperatura.

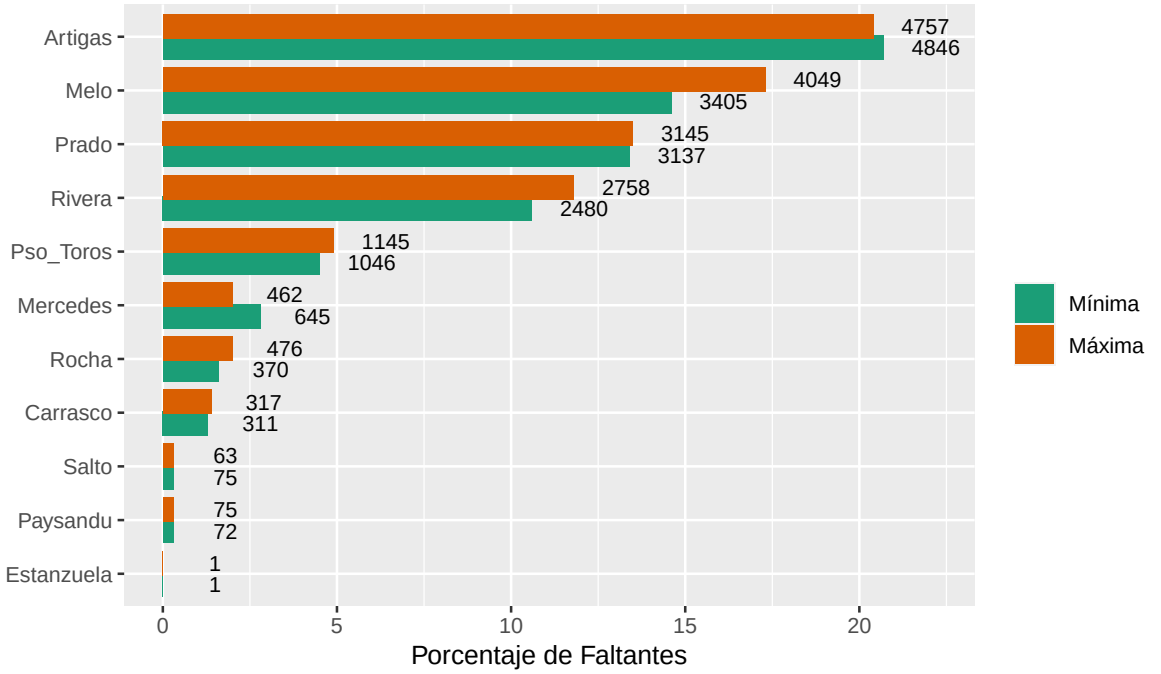

Figura 5: Porcentaje de Faltantes en cada estación meteorológica durante todo el período considerado.

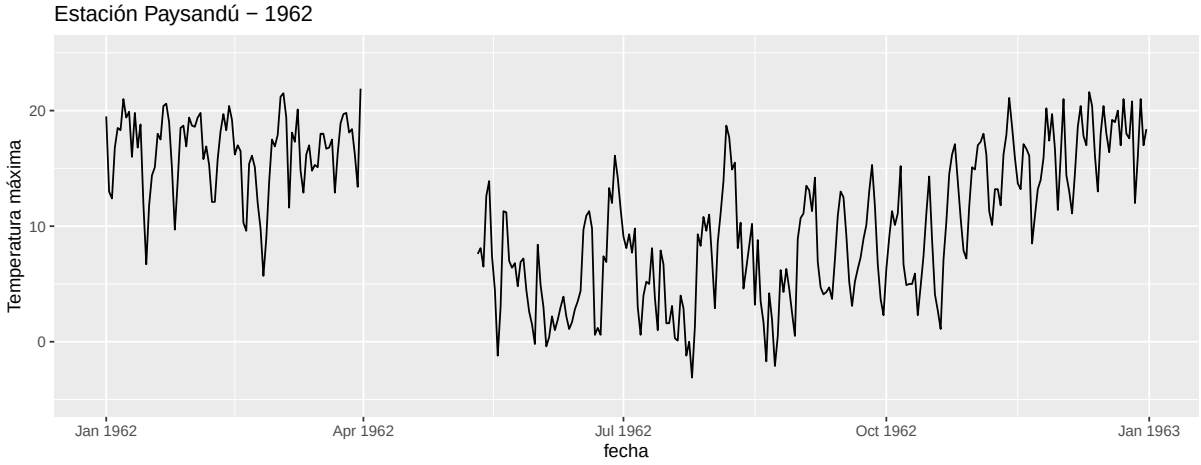

Figura 6: Serie de temperatura mínima en Paysandú en 1962

Bloques de datos faltantes consecutivos son difíciles de imputar debido que una de las principales fuentes de información para los modelos de series temporales es precisamente la dependencia estocástica de los datos, en ese caso períodos prolongados de datos pueden romper la estructura temporal y reducir sensiblemente la calidad de la imputación.

4. Series completas y su visualización

En lo que sigue describimos el DLM específico que utilizamos para modelar las series diarias de temperaturas y obtener series completas para todo el periodo. Modelaremos la serie de mínimas y máximas de forma independiente, utilizando un modelo con igual estructura para ambos casos.

La ecuación 4 presenta el modelo utilizado en el caso de la serie de temperatura mínima diaria.

$$\begin{array}{rcll} y_t^n & = & \theta_t + v_t & v_t \sim \mathcal{N}_D(0, \Sigma) \\ \theta_t & = & \theta_{t-1} + w_t & w_t \sim \mathcal{N}_D(0, W) \end{array} \tag{4}$$

donde $\theta_t = (\theta_{t1}, \ldots, \theta_{tD})^\top$, Σ es una matriz diagonal y W es una matriz de varianzas y covarianzas con todas sus entradas (potencialmente) distintas de cero.

Atendiendo a la ecuación de evolución de la señal latente de temperatura, se puede notar que el modelo anterior toma en cuenta la dependencia temporal de la serie de temperatura respecto del día anterior así como la dependencia espacial entre estaciones ya que el componente w_t es modelado en forma multivariada y no se toman restricciones para las entradas de su matriz de covarianzas. Una debilidad de esta modelización es que no considera el carácter extremo de los datos observados (mínimas o máximas), sin embargo, es discutible que esto sea necesario ya que los datos que se pretenden imputar no son extremos respecto de los datos observados.

El modelo descrito en 4, implica que la temperatura registrada se concibe compuesta de una señal de interés, θ_t, y un ruido aleatorio que es independiente temporalmente y entre estaciones. De esta forma, una **serie completa** de la señal de temperatura para todo el período se puede construir en base a estimaciones puntuales para θ_t, y finalmente la serie de estimaciones puntuales diarias se puede tomar como base para calcular olas de extremos.

La estimación del modelo se realiza aplicando las técnicas descritas las subsecciones previas. Condicional en los valores de las matrices Σ y W, la estimación del proceso θ_t se basa en la aplicación del suavizador de Kalman para obtener la distribución posterior $f(\theta_t|y_1^n \dots y_T^n)$ y luego tomar su esperanza en cada momento para obtener una estimación puntual.

La estimación de las matrices Σ y W se realiza en dos pasos. En primer lugar se estiman por máxima verosimilitud modelos auxiliares con una estructura similar a 4 pero con datos de cada estación individual y_{td}^n, es decir, tomando en cuenta únicamente la dependencia temporal de la serie. A partir de este modelo auxiliar se obtiene valores estimados para la varianza de el ruido en la ecuación de observaciones $(\Sigma)_{dd}$. El segundo paso para obtener una estimación de W, consiste en:

$$(W)_{ij} = \frac{1}{T-1} \sum_t (\hat{\theta}_{ti} - \bar{\theta}_i)(\hat{\theta}_{tj} - \bar{\theta}_j)$$

donde $\hat{\theta}_{td}$ representa la estimación puntual de la señal de temperatura en el día t para la estación d. Es decir, la estimación de W se obtiene como la covrainza muestral de las series completas modeladas *individualmente*.

De la aplicación del modelo descrito en (4) se obtiene una serie sin datos faltantes para cada una de las 11 estaciones meteorológicas de interés, esta serie completa, $\left[\hat{\theta}_t, \;\; t = 1 \dots T\right]$, consiste en una estimación de la *señal de temperatura*

La Figura 7 muestra una comparación de la serie observada en la estación Paysandú en 1962 con una ventana de 40 días consecutivos sin información de la temperatura mínima y la señal estimada por el modelo. Se observa que en los días en que el dato esta presente el modelo se ajusta muy bien a la temperatura observada, y a su vez en el período sin información logra reconstruir una evolución consistente con el patrón general de la serie.

La presentación de todas las 11 series completas con distintas visualizaciones y comparaciones a lo largo de todo el período se realiza mediante una aplicación web basada en la biblioteca 'r shiny' disponible en IESTA-INUMET. El sitio cuenta con 3 pestañas básicas:

Describe: Presenta una descripción de la estructura y funcionalidades de la aplicación y de los métodos estadísticos utilizados para obtener las series de temperatura.

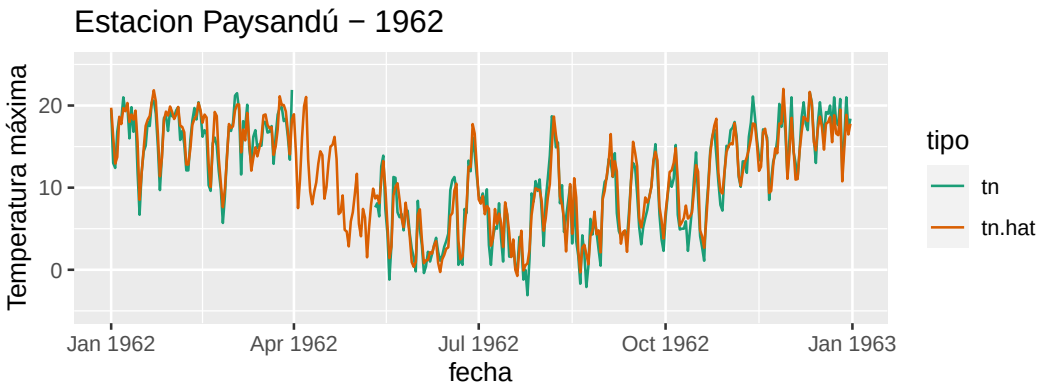

Figura 7: Serie de temperatura minima en Paysandu en 1962

Olas de extremos: Presenta gráficos para visualizar olas de temperatura, un ejemplo de esta figura se presenta en la Figura 8. Se puede seleccionar la estación y el año a graficar.

Series Imputadas: Presenta gráficos comparando las series observadas con la estimación de señal, similar a la Figura 7 recién descrita. Se puede seleccionar la estación, el año de inicio y la cantidad de años a visualizar.

Esta sección se cierra con la definición de ola de temperatura extrema que adopta en el trabajo. Existen diversas formas de caracterizar una racha de frío, que responden a distintas aplicaciones de su estudio. Las diversas definiciones acuerdan en la necesidad de establecer un umbral de bajas temperaturas (puede ser absoluto o relativo) y en delimitar una ventana de tiempo durante la cual, la temperatura observada debe mantenerse por debajo del umbral definido. En el presente trabajo, se define una **ola de frío** como un período de tiempo mayor o igual a 3 días, en los cuales las temperaturas mínimas y máximas son inferiores a los respectivos percentiles 10 esperados para tales días. Definimos el percentil 10 de mínima para el día t, como $p^n_{10_t} := \text{ínf}\{y : \mathbb{P}(Y^n_t \leq y) \geq 0.1\}$. Análogamente definimos el percentil 10 de máxima para el día t como $p^x_{10_t} := \text{ínf}\{y : \mathbb{P}(Y^x_t \leq y) \geq 0.1\}$. Podemos decir entonces que una sucesión de días $t_1, \ldots, t_k$ constituyen una ola de frío de largo k si, siendo $k \geq 3$, se cumple simultáneamente que:

$$\begin{cases} y^n_{t_i} < p^n_{10_i} \\ y^x_{t_i} < p^x_{10_i} \end{cases} \quad \text{para } i = 1, \ldots, k$$

Análogamente, definimos una **ola de calor** a un período de tiempo mayor o igual a 3 días,

en los cuales las temperaturas mínimas y máximas superan sus respectivos percentiles 90 esperados para tales días.

Visualizar fenómenos de olas de extremos no es trivial, aplicando la definición anterior debemos tener en cuenta varios factores. En primer lugar la condición recae simultáneamente sobre la temperatura mínima y máxima por lo que debemos prestar atención a ambas series. Al mismo tiempo debemos comparar la temperatura de cada día en particular con los percentiles para ese día, o más en general, con las temperaturas esperables. Finalmente, los días deben ser consecutivos para formar una ola de extremos.

La figura 8 muestra la distribución de temperatura para cada día del año, sobre la cual están graficados los valores que tomó en el año 2012. El gráfico tiene tres capas básicas: en primer lugar calculamos un histograma bidimensional en que los intervalos son hexágonos en el plano de día y temperatura, (la intensidad del color representa la frecuencia). Sobre el histograma se dibuja las series de temperatura observada para cada día en 2012 y finalmente se divide en dos paneles con la serie de mínima (superior) y máxima (inferior) por separado.

El período de días del 5 al 9 de junio de 2012 (señalado en rojo en la Figura 8) la temperatura observada se acerca a la cola inferior de la distribución para ese día. Eso indica que ese período fue bastante más frío de lo esperado, podemos considerar que durante esos días estuvimos ante la presencia de una *ola de frío*.

5. Comentarios Finales

Este trabajo presenta métodos estadísticos para modelar series diarias multivariadas con importantes ventanas de datos faltantes, y también métodos de visualización de estas series que permitan explorar la presencia de secuencias de valores extremos. Se trabajó con datos de temperaturas mínimas y máximas en 11 estaciones meteorológicas de Uruguay durante un período de más de 60 años.

La modelización de estas series se realiza en base a un modelo Bayesiano de tipo DLM que incorpora la fuerte dependencia temporal y espacial de los datos. La estructura de dependencia juega un papel central en la imputación de los datos faltantes, especialmente cuando se producen ventanas de días consecutivos sin información. Los resultados del

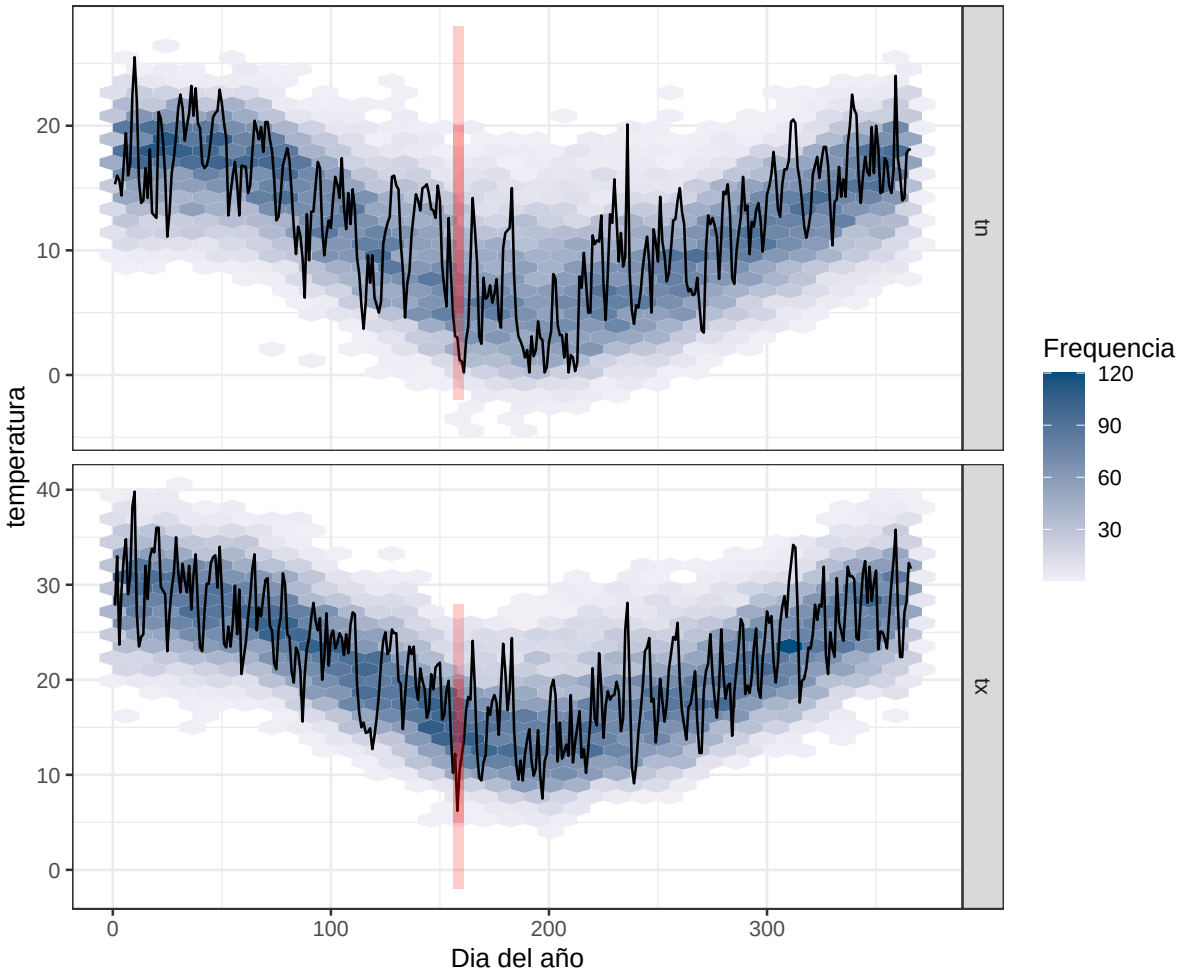

Figura 8: Gráfico de la distribución de temperatura diaria. En negro se representa la serie de mínima y máxima de 2012 en la estación Estanzuela

modelo y las visualizaciones de olas de extremos quedan disponibles en una aplicación web con base en el servidor de IESTA.

En relación a los modelos para obtener series completas, es necesario continuar trabajando para determinar con mayor precisión la calidad de la imputación en ventanas de datos faltantes largas. Esto puede llevarse adelante mediante simulaciones periodos faltantes artificiales para poder comparar con la imputación. Por otro lado, si bien la visualización presentada en el trabajo es un importante primer paso para analizar las olas de extremos con una visión de largo plazo, es necesario avanzar mucho más en esta dirección. En particular, es importante caracterizar las olas de extremos en términos de intensidad y duración, y estudiar como estas características evolucionaron en las últimas décadas. Finalmente, es muy importante estudiar el efecto de la estimación de la señal de temperatura sobre la detección y características de las olas de extremos.

Referencias Bibliográficas

Amiel, J. (2012). *Estudio de diferentes metodologías estadísticas para el control de calidad de bases de datos diarios de precipitación en Uruguay.* Trabajo final de grado. Lic. en Estadística. Fac. De Ciencias Económicas y de Administración. UdelaR.

Chang, W., Cheng, J., Allaire, J., Xie, Y., y McPherson, J. (2020). *shiny: Web Application Framework for R.* R package version 1.5.0.

De Mello, S. (2013). *Estudio climatológico y regionalización de heladas meteorológicas en Uruguay.* Tesis de Licenciatura en Ciencias de la Atmósfera. Facultad de Ciencias.

Petris, G. (2010). An R package for dynamic linear models. *Journal of Statistical Software*, 36(12):1–16.

Petris, G., Petrone, S., y Campagnoli, P. (2009). Dynamic linear models. En *Dynamic Linear Models with R*, pp. 31–84. Springer.

Prado, R. y West, M. (2010). *Time series: modeling, computation, and inference.* CRC Press.

R Core Team (2020). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Renom Molina, M. (2009). *Temperaturas extremas en Uruguay. Análisis de la variabilidad temporal de baja frecuencia y su relación con la circulación de gran escala.* Tesis Doctoral, Universidad de Buenos Aires. Facultad de Ciencias Exactas y Naturales.

West, M. y Harrison, J. (1996). Bayesian forecasting. *Encyclopedia of Statistical Sciences.*

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York.

Instituto de Estadística

Serie Documentos de Trabajo

Gonzalo Ramirez 1926, Piso 1, Oficina 23 - C.P. 11200 -
Montevideo, Uruguay
Teléfono: (598) 2410 2564
https://iesta.fcea.udelar.edu.uy/
Área Publicaciones

Abril, 2021

№4/21