

# Modelo de predicción de alturas del agua en el Río de la Plata

Tesis presentada para optar al título de Magister en Explotación de Datos y  
Descubrimiento del Conocimiento

Lic. María Inés Frías  
Director: Dr. Rafael Grimson

29 de abril de 2022



Universidad de Buenos Aires  
Facultad de Ciencias Sociales y Naturales

# Modelo de predicción de alturas del agua en el Río de la Plata

La costa de Argentina comienza en el estuario del Río de la Plata, ubicado cerca de los 35°S sobre la costa Atlántica de América del Sur. Después del río Amazonas, el Río de la Plata, junto a sus principales afluentes (los ríos Paraná y Uruguay), es una de las principales vías fluviales del hemisferio sur, así como la segunda mayor cuenca de América Latina. Las dos ciudades más importantes de su costa son Buenos Aires, Argentina (AR) y Montevideo, Uruguay (UR).

En el Río de la Plata el nivel del agua está afectado por dos fenómenos. Por un lado, el efecto astronómico de las mareas, que genera una fluctuación semi-diurna, predecible sencillamente, que se acopla a fluctuaciones generadas por efectos atmosféricos, generando un efecto complejo y de difícil predicción. Por otro, los efectos causados por fuertes vientos que inciden en el nivel del agua sobre la costa de Buenos Aires y sus alrededores.

En particular, el fenómeno conocido como sudestada está asociado con fuertes vientos del sudeste sobre el estuario del Plata que en ocasiones incide en la costa de Buenos Aires y sus alrededores, causando inundaciones y daños por millones de dólares.

El objetivo del presente trabajo es generar un modelo para predecir las alturas del Río de la Plata con varias horas de anticipación en la ciudad de Buenos Aires utilizando modelos predictivos basados en datos abiertos.

A tal fin, se presenta el desarrollo y evaluación de modelos lineales, como regresiones lineales y modelos autorregresivos de media móvil (SARIMAX), y no lineales, como las redes neuronales con memoria de corto largo plazo (LSTM).

Los modelos incorporaron datos de observaciones de alturas horarias y vientos, y pronósticos de vientos, presión y mareas, provistas por el Servicio de Hidrografía Naval (SHN), el Servicio Meteorológico Nacional (SMN) y el National Center for Atmospheric Research (NCAR). El conjunto de datos se generó utilizando cálculos de media móvil en distintas ventanas de tiempo y aplicando desplazamientos temporales sobre los datos provistos, y seleccionando iterativamente aquellas variables que explicaran el comportamiento de las alturas horarias en mayor grado.

Los modelos con mejor precisión de cada clase fueron comparados entre sí utilizando una métrica robusta de error de predicción acumulado adecuada para series de tiempo. De manera incremental, esta métrica incorpora observaciones y calcula la raíz del error cuadrático medio respetando la secuencia temporal.

Los resultados preliminares muestran que los modelos lineales obtuvieron mejores resul-

---

tados en términos de esta métrica que la red neuronal. Como parte de las conclusiones finales, estos resultados serán comparados con los resultados obtenidos de los modelos numéricos preexistentes SMARA y SHN/SMN, utilizados desde hace varias décadas en pronósticos del nivel del Río de la Plata.

Palabras claves: aprendizaje automático, predicción de alturas, Río de la Plata, sudestadas, regresión lineal, series temporales, SARIMAX, redes neuronales, LSTM

# Agradecimientos

Quisiera expresar mi más sincero agradecimiento a mi director de tesis, Dr. Rafael Grimson, por todo el tiempo dedicado a guiarme en el desarrollo de este trabajo. Su ayuda y consejos han significado un aporte invaluable para mí. También quiero agradecer a la Dra. Paula Etala por brindar de manera tan generosa su conocimiento y experiencia a través de sus numerosos trabajos en este área.

De manera especial, quiero agradecer a mi hija Valentina, su apoyo y entusiasmo permanentes fueron sin duda grandes motivadores en este camino largo y desafiante, así como también a mis padres, mi hermano y amigos, quienes me acompañaron y alentaron en todo momento a finalizar este proyecto.

Por último, agradezco a la Dra. María Elena Buemi por su guía y consejos, a la maestría y sus profesores, quienes me aportaron su gran experiencia y conocimientos.



# Índice general

|                                                           |           |
|-----------------------------------------------------------|-----------|
| <b>1. Introducción</b>                                    | <b>6</b>  |
| <b>2. Materiales y Métodos</b>                            | <b>10</b> |
| 2.1. Estuario del Río de la Plata . . . . .               | 10        |
| 2.1.1. Ubicación y características físicas . . . . .      | 10        |
| 2.1.2. Clima . . . . .                                    | 11        |
| 2.2. Datos . . . . .                                      | 12        |
| 2.2.1. Datos medidos (observaciones) . . . . .            | 12        |
| 2.2.2. Datos estimados (predicciones) . . . . .           | 15        |
| 2.3. Aprendizaje automático supervisado . . . . .         | 17        |
| 2.3.1. Regresión lineal . . . . .                         | 18        |
| 2.3.2. Modelos lineales sobre series de tiempo . . . . .  | 19        |
| 2.3.3. Redes neuronales . . . . .                         | 22        |
| 2.4. Procedimientos de la generación de modelos . . . . . | 26        |
| 2.4.1. Conjuntos de entrenamiento y evaluación . . . . .  | 26        |
| 2.4.2. Evaluación y selección de modelos . . . . .        | 27        |
| 2.5. Software utilizado . . . . .                         | 30        |
| <b>3. Análisis exploratorio y creación de variables</b>   | <b>32</b> |
| 3.1. Análisis exploratorio . . . . .                      | 32        |
| 3.1.1. Alturas horarias . . . . .                         | 32        |
| 3.1.2. Tablas de marea . . . . .                          | 36        |
| 3.1.3. Vientos . . . . .                                  | 37        |

|                                                                                       |           |
|---------------------------------------------------------------------------------------|-----------|
| 3.1.4. Pronósticos de viento y presión media sobre el nivel del mar (m.s.n.m.)        | 40        |
| 3.2. Creación de variables . . . . .                                                  | 41        |
| 3.3. Propiedades de la serie de tiempo de alturas del Río de la Plata . . . . .       | 41        |
| <b>4. Modelado</b>                                                                    | <b>43</b> |
| 4.1. Regresión lineal . . . . .                                                       | 43        |
| 4.2. Modelos lineales sobre series de tiempo . . . . .                                | 44        |
| 4.2.1. Formulación de clase general del modelo . . . . .                              | 45        |
| 4.2.2. Iteración 1: SARIMAX (3,0,12)x(2,1,1) <sub>24</sub> y una variable exógena . . | 50        |
| 4.2.3. Iteración 2: SARIMAX (3,0,12)x(2,1,1) <sub>24</sub> y 2 variables exógenas . . | 52        |
| 4.2.4. Iteración 3: SARIMAX (3,0,12)x(2,1,1) <sub>24</sub> y 3 variables exógenas . . | 53        |
| 4.2.5. Iteración 4: SARIMAX (3,0,12)x(2,1,1) <sub>24</sub> y 4 variables exógenas . . | 55        |
| 4.2.6. Iteración 5: SARIMAX (3,0,12)x(2,1,1) <sub>24</sub> y 5 variables exógenas . . | 56        |
| 4.3. Redes neuronales . . . . .                                                       | 56        |
| 4.3.1. Memoria de largo corto plazo (Long Short-Term Memory o LSTM) .                 | 56        |
| <b>5. Comparación de resultados</b>                                                   | <b>60</b> |
| 5.1. Comparación entre modelos lineales y no lineales generados . . . . .             | 60        |
| 5.2. Comparación con modelos numéricos preexistentes . . . . .                        | 61        |
| 5.2.1. Error medio de pronóstico del nivel de agua >1m . . . . .                      | 61        |
| 5.2.2. Error medio de pronóstico en períodos de sudestada y bajante . . .             | 62        |
| 5.2.3. Error de predicción acumulado entre 2014 y 2017 . . . . .                      | 64        |
| <b>6. Discusión y conclusiones</b>                                                    | <b>65</b> |
| <b>Bibliografía</b>                                                                   | <b>68</b> |
| <b>Índice de figuras</b>                                                              | <b>71</b> |
| <b>Índice de cuadros</b>                                                              | <b>72</b> |

# Capítulo 1

## Introducción

El nivel del Río de la Plata está afectado por el **efecto astronómico**, causado por las interacciones gravitacionales entre la Tierra, la Luna y el Sol, y el **efecto meteorológico**, generado por factores atmosféricos.

La gravedad del Sol y la Luna ejercen una atracción gravitatoria que afecta tanto la parte líquida como sólida de la Tierra, siendo en la primera donde los efectos son más notorios. Esta atracción genera cambios periódicos en el nivel del agua llamados mareas. La pleamar o marea alta es el momento en el que el agua alcanza su máxima altura dentro de un ciclo de mareas. De manera opuesta, la bajamar o marea baja es el momento donde alcanza su mínima altura.

En relación a los factores atmosféricos, vientos con dirección noroeste que soplan sobre la plataforma continental ocasionan ondas de tormenta (niveles del agua menos niveles de marea astronómica). Estas ondas de tormenta aumentan su magnitud al entrar en aguas poco profundas en la boca del estuario. Según [Perillo et al., 2012], el peor escenario se presenta durante las sudestadas, generadas por ciclogénesis centralizadas en el noreste de Argentina o Uruguay. Como describe el análisis estadístico de [Escobar G, 2004], estos eventos están caracterizados por variables como la altura y duración de las ondas de tormenta (Fig. 1.1).



Figura 1.1: Costanera Norte - Nov, 2014

Las sudestadas pueden durar varios días, lo que genera severas inundaciones en las áreas más densamente habitadas como la línea costera suroeste del estuario. Sumado a esto, las sudestadas están en general acompañadas por fuertes lluvias sostenidas que contribuyen a las inundaciones.

Si bien las crecidas en el Río de la Plata han sido estudiadas por numerosos autores ([Balay, 1961], [D’Onofrio et al., 1999], [Escobar G, 2004], [Cerne et al., 2005]), no ocurre lo mismo con las bajantes. En su trabajo, [Balay, 1961] muestra que grandes bajantes se registran con vientos fuertes de los sectores NNO, N y NNE.

En este aspecto, las bajantes u ondas de tormenta negativas afectan tanto a la navegación en el Río de la Plata como a la provisión de agua potable de una gran parte de la población que vive tanto en Buenos Aires como en las zonas aledañas (Fig. 1.2).



Figura 1.2: Costanera Norte - Jul, 2014

Hacia fines de la década del ‘50, a medida que en nuestro país se avanzaba en el desarrollo de distintas técnicas de pronóstico de olas en forma de métodos gráficos-estadísticos, se descubrió que el comportamiento de las olas podía ser descrito por una ecuación de balance de energía que daba la evolución temporal del espectro de las olas. De acuerdo a la publicación de [Etala, 1993], a partir de ahí se hicieron numerosos avances en la determinación de los procesos físicos asociados con el crecimiento y decaimiento, como la inestabilidad del flujo acoplado ola-viento, la interacción entre olas y, en aguas poco profundas, procesos relacionados con el fondo.

En el Servicio Meteorológico de la Armada Argentina (SMARA) funciona desde 1969 un modelo de olas de la llamada primera generación. Originalmente conducido por vientos analizados manualmente, desde 1988 se conduce por modelos meteorológicos. Hoy realiza una verificación en tiempo real con datos provenientes de mediciones visuales de las alturas, de baja confiabilidad y con información incompleta.

Este modelo continuó su evolución, incorporando técnicas de determinación del viento en superficie, mediante el diagnóstico y pronóstico de variables meteorológicas. El esquema permite la asimilación de datos de distintas fuentes de información (estaciones meteorológicas, satélites, buques, boyas y pronósticos globales).

En su trabajo publicado sobre la precisión de predicciones de ondas de tormenta extra tropicales, [Etala, 2009] menciona que el rango máximo de pronóstico en Buenos Aires

basado en las observaciones de nivel más recientes es de aproximadamente 12hs, siendo la mejor predicción de nivel del agua el pronóstico de 6hs (o *nowcast*).

La transición a la tercera generación de modelos se concretó operacionalmente en el 2004 [Etala et al., 2009]. La modelación marina está incluida en un sistema troncal y sostenido de pronóstico numérico del tiempo y desde el 2007 el Servicio de Hidrografía Naval (SHN) y el Servicio Meteorológico Nacional (SMN) difundieron públicamente el pronóstico numérico de olas para el Atlántico Sudoccidental y Mar Argentina y de ondas de tormenta en el litoral atlántico y Río de la Plata. Al mismo tiempo, el pronóstico de ondas de tormenta se comenzó a utilizar experimentalmente en el SMARA. Así, los dos núcleos del sistema operacional son el modelo de pronóstico de olas SMARA/WAM y el modelo de pronóstico de ondas de tormenta del SMARA, integrando pronósticos atmosféricos del National Centers for Environmental Prediction (NCEP) y del modelo Eta SMN.

Según menciona ese trabajo, la asimilación de datos de nivel del agua a tiempo real presenta buenas posibilidades de mejorar los pronósticos numéricos en cuanto a las ondas de tormenta, debido a la dinámica de propagación de la onda de tormenta en el Río de la Plata. En el mismo se demuestra que el pronóstico numérico de onda de tormenta permite extender el plazo de los pronósticos de nivel del agua disponibles, sin perder calidad.

Con respecto al uso de técnicas de aprendizaje supervisado en el modelado y predicción de niveles del agua en escenarios similares utilizando, se han realizado varios intentos con modelos en el pasado. En este sentido, el análisis de series de tiempo ha jugado un rol importante en la investigación hidrológica.

Tradicionalmente, los modelos autorregresivos de media móvil (ARIMA) han sido los métodos más utilizados en problemáticas similares como la predicción de la descarga de afluentes. Como publica [Fashae et al., 2019] en su trabajo sobre la descarga del Río Opeki, los modelos lineales obtuvieron mejores resultados que las redes neuronales.

Sin embargo, en ciertos casos los modelos lineales han sido deficientes en la representación de dinámicas no lineales donde las ANN han demostrado la capacidad de aproximar casi cualquier función continua a un grado de precisión, no conseguido con otras técnicas hidrológicas convencionales.

Los modelos ANN ya fueron aplicados en la predicción de mareas y ondas de tormenta. En [Tsai and Lee, 1999] y [Lee and Jeng, 2002] se utilizaron ANNs para el pronóstico de mareas basados en datos de campo. Más tarde, los resultados publicados en [Lee, 2004] [Lee, 2008] propusieron la aplicación de ANN utilizando algoritmos de propagación hacia atrás para la predicción de niveles de marea a largo plazo y de ondas de tormenta a corto plazo. A partir de su comparación con métodos numéricos y análisis de mareas convencionales, se pudo observar que tanto las mareas en el largo plazo como las ondas de tormenta en el corto plazo (entre 1h y 6hs previas) pueden predecirse de manera eficiente.

Como parte de este trabajo, se analizaron comparativamente modelos basados en técnicas lineales y no lineales de acuerdo a las predicciones correspondientes de las alturas del Río de la Plata y la precisión en términos de la Raíz del Error Cuadrático Medio (RMSE) y métricas robustas de error acumulado.

En este trabajo se utilizaron datos de predicciones de mareas realizadas por el SHN para distintos puertos en un área que se extiende desde Puerto Madryn (AR) hasta Chuy (UR), observaciones efectivas de los niveles alcanzados en estos puntos, llamadas alturas horarias y datos meteorológicos provistos por el SMN, el Instituto Uruguayo de Meteorología

(INUMET) y el National Center for Atmospheric Research (NCAR) Climate Forecast System (CFSv2). El período considerado en el estudio comprende desde Enero de 2014 a Diciembre de 2017.

Las mareas fueron modeladas con tres métodos: regresiones lineales, autorregresivos de media móvil y redes neuronales. Los datos se dividieron en una etapa de entrenamiento y otra de evaluación. La evaluación de los distintos modelos se realizó de dos formas complementarias. En una primera etapa, los resultados obtenidos por los modelos generados fueron comparados con las alturas horarias observadas por el SHN y posteriormente se realizó una comparación con los modelos preexistentes.

## Capítulo 2

# Materiales y Métodos

### 2.1. Estuario del Río de la Plata

#### 2.1.1. Ubicación y características físicas

El Río de la Plata está ubicado en la costa Este de América de Sur, entre los  $34^{\circ}$ - $36^{\circ}10'S$  y  $55^{\circ}$ - $58^{\circ}10'O$ . Es un estuario poco profundo que cubre una superficie cercana a los  $35.000\text{km}^2$ . Tiene  $320\text{km}$  de largo y el ancho varía entre  $38\text{km}$  en la región superior hasta  $230\text{km}$  en la boca entre Punta Rasa y Punta del Este [Perillo et al., 2012] (Fig 2.1).



Figura 2.1: Ubicación del Río de la Plata respecto a América de Sur y referencia de coordenadas

El Río de la Plata desagota la segunda cuenca más grande del continente, proveniente de los ríos Paraná-Paraguay y Uruguay, y la de mayor importancia en términos de población

y actividad económica.

Por su morfología y dinámica, el estuario puede dividirse en dos regiones principales, superior e inferior. Dichas regiones están divididas por la Barra del Indio, un área poco profunda que cruza en la línea Punta Piedras-Montevideo. El área superior tiene un régimen fluvial, casi en su totalidad ocupada por agua dulce. En el estuario inferior el agua dulce interactúa con la plataforma que se refleja en una turbidez máxima.

### 2.1.2. Clima

En las regiones superior y media, los vientos son responsables de niveles de agua extremos, mientras que en la region inferior la influencia del viento es menor. Las tormentas más frecuentes que afectan al Río de la Plata son las llamadas sudestadas. Estas se producen por la interacción entre sistemas frontales (con trayectorias SO-NE) y el sistema de baja presión del litoral, generando fuertes vientos del SE. En general las tormentas afectan el área por varios días, causando inundaciones a lo largo de la costa argentina del Río de la Plata [Perillo et al., 2012].

Por otro lado, investigaciones realizadas en los últimos decenios ([Sivaramanan, 2014], [Couharde et al., 2019], entre ellas) destacan la importante influencia que ejercen las interacciones entre la atmósfera y el océano en el cinturón tropical del océano Pacífico sobre las características del tiempo y del clima a escala mundial. Durante los episodios de El Niño, por ejemplo, la temperatura de la superficie del mar en las partes central y oriental del Pacífico tropical suele ser muy superior a la normal, mientras que, en esas mismas regiones, durante los episodios de La Niña la temperatura es inferior a la normal. Esas variaciones de temperatura pueden provocar fluctuaciones importantes del clima en el mundo entero y, una vez comenzadas, esas anomalías pueden durar un año, o incluso más [World Meteorological Organization (WMO) “El Niño/La Niña Background”]. Los episodios de El Niño y La Niña son extremos de lo que comúnmente se llama ciclo El Niño/Oscilación del Sur (ENSO por sus siglas en inglés). El ciclo tiene un período promedio de unos cuatro años, aunque en registros históricos el período ha variado entre dos y siete años [Changnon, 2000]. Desde el año 2014, período considerado en este trabajo, se observaron los eventos de El Niño y La Niña que se muestran en la Fig. 2.2.

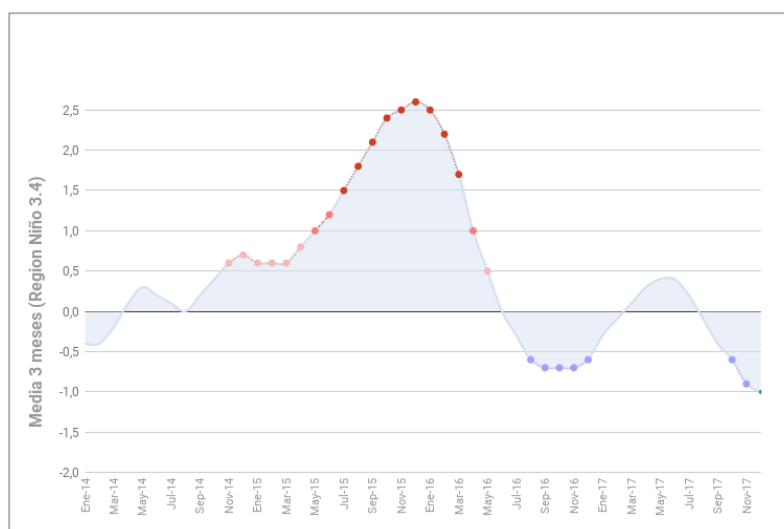


Figura 2.2: Índice Oceánico Niño (ONI). Eventos de El Niño y La Niña entre 2014 - 2017





Figura 2.3: Conjuntos de datos y organismos de procedencia

El Índice Oceánico Niño (ONI, por sus siglas en inglés) se ha convertido en el índice estándar para la identificación de dichos eventos en el Pacífico tropical.

## 2.2. Datos

Los datos utilizados en este trabajo son de tres tipos. Por un lado, utilizaremos las predicciones de mareas realizadas por el SHN para distintos puertos en un área que se extiende desde Puerto Madryn (AR) hasta Chuy (UR). Estas sólo tienen en cuenta el efecto astronómico y son regulares. Por otra parte, el mismo organismo recolecta observaciones efectivas de los niveles alcanzados en estos puntos, llamadas alturas horarias. El tercer tipo de datos corresponde a datos meteorológicos. Estos últimos fueron provistos por el SMN, el Instituto Uruguayo de Meteorología (INUMET) y el National Center for Atmospheric Research (NCAR) Climate Forecast System (CFSv2) [Saha et al., 2011]. Los mismos incluyen datos de vientos y presiones atmosféricas correspondientes a un dominio definido entre 33°S y 43°S y 53°O y 65°O. (Fig. 2.3)

Si bien los conjuntos de datos no presentaron en general datos faltantes (o estos eran remediados aplicando técnicas adecuadas como se explica más adelante), en algunos casos los faltantes se observaron en períodos extendidos, condicionando la selección del conjunto final de datos. Este fue el caso de la medición de vientos en algunas de las estaciones de Uruguay, donde la interpolación de los valores faltantes habría introducido distorsiones en los datos.

El período considerado en este trabajo abarca datos de observaciones y pronósticos entre los años 2014 y 2017, con un registro de frecuencia horaria continuo en distintos puntos del dominio referido al tipo de medición. El período capta así eventos de El Niño y La Niña mencionados en la sección anterior, que pudieran generar comportamientos particulares.

La Fig. 2.4 muestra las referencias geográficas de acuerdo a cada tipo de dato (observaciones y predicciones) detallado en las siguientes secciones.

### 2.2.1. Datos medidos (observaciones)

#### Alturas horarias

Las alturas horarias observadas del nivel de agua son parte central del análisis y modelado en este trabajo. Los datos corresponden al Puerto de Buenos Aires, Muelle Club



Figura 2.4: Referencias geográficas del dominio

de Pescadores frente a Aeroparque J. Newbery ( $34^{\circ}33'39S$ ,  $58^{\circ}23'56O$ ). Las alturas están expresadas en cm y refieren al plano de reducción que coincide con el cero del Riachuelo y pasa 79 cm debajo del nivel medio.

Siendo  $x(t)$  la medición de la altura del agua en el momento  $t$ , asumiremos que  $X(t)$  es un proceso estacionario que como muestra la Fig.2.5a sigue una forma Gausseana y una media observable alrededor de 95cm, en valores enteros de  $t$  en el intervalo  $[0, T]$  (Fig.2.5b).

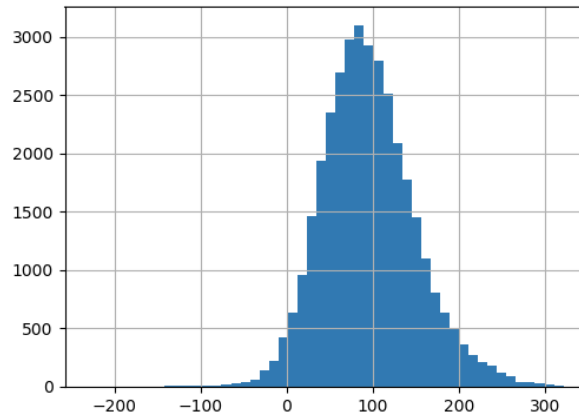
Si bien la muestra parecieran provenir de una distribución Gausseana, el test estadístico de normalidad Anderson-Darling rechazó la hipótesis nula (estadístico = 87.267 con un nivel de confianza del 5 %), probando que los datos no siguen una distribución normal. Dicha consideración es tomada en cuenta a lo largo de este trabajo.

Como se detallará más adelante, muchos de los modelos generados sobre series de tiempo trabajan sobre el supuesto de que la serie es estacionaria. Intuitivamente, si la serie se comporta de una forma determinada a lo largo del tiempo, existe una alta probabilidad que continúe haciéndolo en el futuro.

La Fig.2.6 muestra el registro de alturas horarias de Buenos Aires durante el año 2017. De forma visual, la serie muestra un comportamiento estacionario, no observándose una tendencia creciente o decreciente en la misma. Sin embargo, tomando una ventana menor en Enero de ese mismo año, la curva muestra picos en períodos de aproximadamente 12 horas (Fig.2.7).

De manera más formal, el análisis de la próxima sección utilizó los siguientes métodos para comprobar si efectivamente la serie era o no estacionaria:

- gráficos de estadísticos móviles. Se graficó la media y el desvío estándar móviles para comprobar si existían o no variaciones en el tiempo.
- uso de tests estadísticos. Se utilizaron tests estadísticos a fin de chequear si la expectativa estacionaria se cumplía o no.



(a) Histograma alturas horarias

|       |              |
|-------|--------------|
| count | 35064.000000 |
| mean  | 94.960529    |
| std   | 54.488846    |
| min   | -230.000000  |
| 25%   | 58.000000    |
| 50%   | 91.000000    |
| 75%   | 127.000000   |
| max   | 321.000000   |

(b) Estadísticos

Figura 2.5: Alturas horarias Buenos Aires - 2014-2017

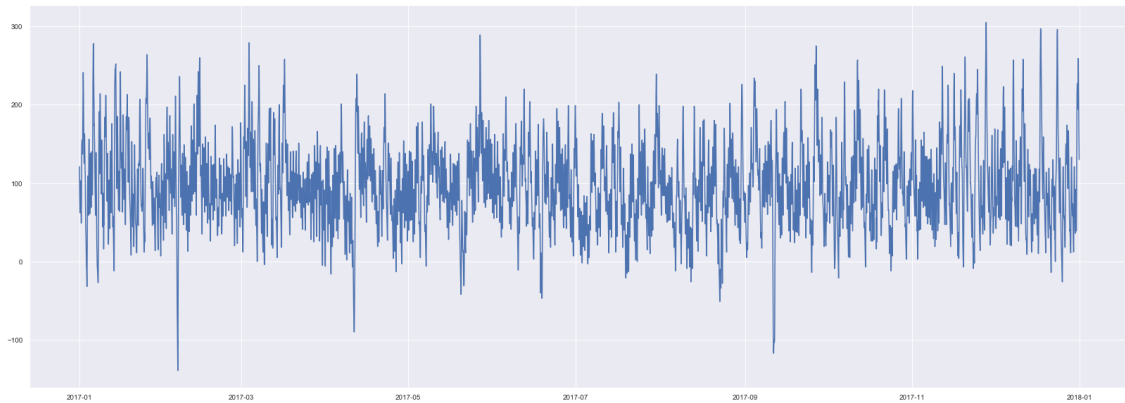


Figura 2.6: Alturas horarias Buenos Aires - 2017

## Vientos

Según menciona el trabajo de [Balay, 1961], el efecto de factores climáticos sobre la amplitud de marea del Río de la Plata es muy significativo, siendo uno de los factores más importantes la influencia de tormentas y vientos locales.

El conjunto de datos analizado fue provisto por el Servicio Meteorológico Nacional (SMN), con mediciones horarias para los siguientes puntos de referencia:

- San Fernando
- Buenos Aires
- Mar del Plata

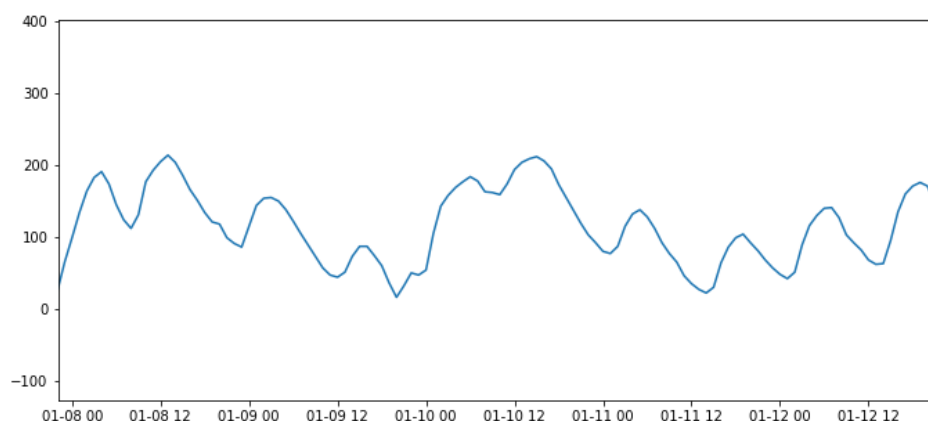


Figura 2.7: Alturas horarias Buenos Aires - Enero 2017

Los registros de viento están expresados en términos de velocidad (km/h) y dirección del viento. La dirección sigue la convención meteorológica, es decir, indica la dirección desde la cual sopla el viento. El valor se incrementa en el sentido de las agujas del reloj desde el punto norte y está expresado en grados en un intervalo entre  $0^\circ$  -  $360^\circ$ . El viento Este equivale entonces a  $90^\circ$ .

### 2.2.2. Datos estimados (predicciones)

#### Tablas de marea

Desde el aspecto astronómico, debido a las aguas poco profundas del Río de la Plata, la generación de mareas no es posible dentro de la cuenca interna. Las mayores fluctuaciones del nivel del agua son producidas por procesos oceánicos influenciados por patrones climáticos [Perillo et al., 2012].

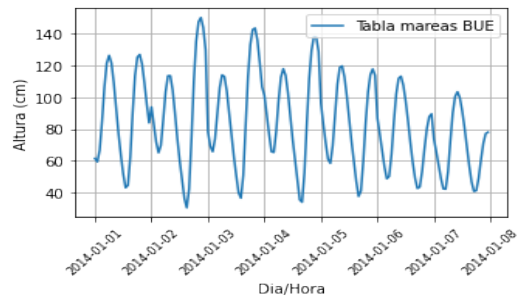
Las mareas del Río de la Plata tienen un régimen semi-diurno (el ciclo completo de pleamar y bajamar se completa en medio día) con desigualdades diurnas [Balay, 1961]. El tamaño de la cuenca y las doce horas que tardan las mareas en recorrer todo el Río de la Plata hacia arriba, genera momentos en que las mareas alta y baja ocurren simultáneamente en distintos lugares.

En este trabajo se incluyeron datos correspondientes a las tablas de mareas provistos por el Servicio de Hidrografía Naval (SHN), tanto para puertos de Argentina como así también de Uruguay (Fig. 2.4). Las alturas se registran en metros en referencia al plano de reducción que pasa 0,85m debajo del nivel medio.

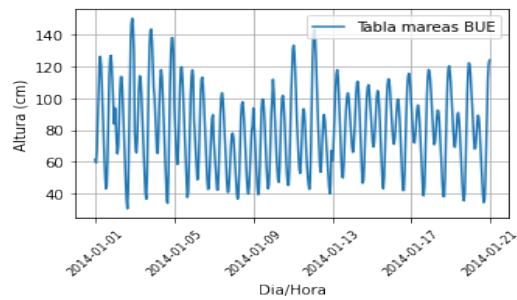
Las tablas de marea proveen los pronósticos diarios de pleamar y bajamar. La tabla de la Fig.2.8a lista los pronósticos correspondientes a la Dársena F de Buenos Aires durante la primera semana de Enero 2014. A partir de estos valores se genera una serie de tiempo horaria, que refleja cierta regularidad, y por lo tanto predictibilidad, asociada con un régimen semidiurno de mareas, con dos pleamares y dos bajamares en el transcurso de un día, como muestran los gráficos de las Fig.2.8b y 2.8c.

| Día | Hora:Min | Altura |
|-----|----------|--------|
| 1   | 0:48     | ▼ 0.59 |
|     | 5:59     | ▲ 1.26 |
|     | 13:25    | ▼ 0.42 |
|     | 18:42    | ▲ 1.27 |
| 2   | 2:05     | ▼ 0.60 |
|     | 6:53     | ▲ 1.22 |
|     | 14:17    | ▼ 0.38 |
| 3   | 19:44    | ▲ 1.38 |
|     | 3:17     | ▼ 0.61 |
|     | 7:47     | ▲ 1.16 |
| 4   | 15:11    | ▼ 0.33 |
|     | 20:49    | ▲ 1.46 |
|     | 4:27     | ▼ 0.60 |
|     | 8:43     | ▲ 1.10 |
| 5   | 16:05    | ▼ 0.30 |
|     | 21:57    | ▲ 1.51 |
|     | 5:32     | ▼ 0.59 |
|     | 9:41     | ▲ 1.04 |
| 6   | 17:00    | ▼ 0.28 |
|     | 23:06    | ▲ 1.54 |
|     | 6:34     | ▼ 0.58 |
| 7   | 10:40    | ▼ 0.99 |
|     | 17:56    | ▼ 0.27 |
|     | 0:12     | ▼ 1.55 |
|     | 7:33     | ▼ 0.57 |
|     | 11:40    | ▼ 0.96 |
|     | 18:52    | ▼ 0.28 |

(a) Tabla bajamar y pleamar 1ª semana



(b) Pronósticos marea horarios 1ª semana



(c) Pronósticos marea horarios del mes

Figura 2.8: Pronósticos de marea Dársena F - Buenos Aires, Enero 2014

Entre los puertos de Argentina, se incluyeron predicciones de marea de:

- Buenos Aires
- La Plata
- Mar del Plata
- Oyarvide
- Pilote Norden
- San Clemente
- San Fernando
- Santa Teresita

mientras que entre los puertos de Uruguay, se incluyeron de:

- Colonia
- Montevideo

## Vientos y presión media sobre el nivel del mar

Los pronósticos de viento y presión media reducida a nivel del mar fueron tomados del National Centers for Environmental Prediction (NCEP) Climate Forecast System (CFS) version 2. El CFSv2 mejora y extiende a su predecesor, el CFSv1, originalmente diseñado como una integración global, de alta resolución, del sistema atmósfera - océano - superficie terrestre - hielo marítimo para proveer la mejor estimación del estado de estos dominios acoplados.

El CFSv2 introdujo mejoras en componentes de su versión anterior, específicamente en los dos modelos de pronóstico y en los dos sistemas de asimilación de datos, junto con otras novedades. Pero por sobre todo, el CFSv2 fue diseñado para mejorar la consistencia entre estados del modelo y estados iniciales producidos por los sistemas de asimilación. El CFSv2 fue implementado y puesto en marcha en Marzo 2011 [Saha et al., 2011].

Los datos fueron suministrados en formato grib2 (GRIdded Binary or General Regularly-distributed Information in Binary form), formato conciso comúnmente usado en meteorología para almacenar datos climáticos. Estos archivos fueron procesados mediante un utilitario en C a fin de convertirlos en archivos con valores separados por coma (.csv). En este punto los archivos fueron procesados por un script para filtrar las referencias de latitud y longitud de interés.

El CFS es inicializado 4 veces al día (0000, 0600, 1200, and 1800 UTC), generando productos CFS de series de tiempo en intervalos horarios, combinando un análisis y los pronósticos subsiguientes entre 1 y 5 horas.

Los datos descargados del NCEP correspondieron al área comprendida entre los 33°-43° S y 53°-65°O, dentro de la cual se seleccionaron las siguientes ubicaciones:

De Argentina:

- Bahía Blanca
- Buenos Aires
- La Plata
- Mar del Plata
- Villa Gesell

De Uruguay:

- Colonia
- Montevideo
- Punta del Este

## 2.3. Aprendizaje automático supervisado

El análisis de series de tiempo cumple un rol fundamental en el estudio de fenómenos climatológicos. Se utilizaron diversas técnicas para el modelado y predicción de los mismos.

Algunos de los modelos generados mostraron mayor eficiencia a la hora de representar eventos no lineales asociados con estos fenómenos. Técnicas como las redes neuronales muestran mayor efectividad a la hora de aproximar series continuas. Este trabajo comienza trazando una línea base definida por modelos lineales básicos, avanzando luego hacia modelos autorregresivos de media móvil para finalizar haciendo una comparación con redes neuronales, con el fin de encontrar el mejor modelo predictor de alturas del Río de la Plata.

### 2.3.1. Regresión lineal

Como método inicial de aprendizaje supervisado, en este trabajo se aplicaron modelos de regresión lineal. El objetivo de la regresión es predecir el valor de una o más variables continuas dado el valor de un vector  $D$ -dimensional  $x$  de variables de entrada.

En su forma más simple, el modelo de regresión lineal es también una función lineal de las variables de entrada

$$y(X, W) = w_0 + w_1x_1 + \dots + w_Dx_D$$

donde  $X = (x_1, \dots, x_D)^T$ . Esto se conoce simplemente como regresión lineal. Este modelo es una función lineal de los parámetros  $w_0, \dots, w_D$ . Es además, una función lineal de las variables  $x_i$ .

Sin embargo, podemos obtener una clase de funciones de mayor utilidad tomando combinaciones lineales de funciones fijas no lineales de las variables de entrada

$$y(X, W) = w_0 + \sum_{j=1}^{M-1} w_j\phi_j(X)$$

donde  $\phi_j(X)$  son conocidas como *funciones base*. El total de parámetros de este modelo será entonces  $M$ . El parámetro  $w_0$  permite desplazamientos fijos, llamados *bias*. Por conveniencia, a menudo se define una función base  $\phi_0(X) = 1$ , de manera que

$$y(X, W) = \sum_{j=0}^{M-1} w_j\phi_j(X) = W^T\Phi(X)$$

donde  $W = (w_0, \dots, w_{M-1})^T$  y  $\Phi = (\phi_0, \dots, \phi_{M-1})^T$ . Funciones de esta forma son llamadas modelos lineales, ya que son lineales en  $W$ . Es esta linealidad en los parámetros lo que simplifica el análisis de esta clase de modelos.

A pesar de las limitaciones de los modelos lineales como técnicas prácticas de reconocimiento de patrones, en particular para problemas que involucran espacios de dominio de alta dimensionalidad, poseen buenas propiedades analíticas y sirven de base a modelos más sofisticados [Bishop, 2006].

### 2.3.2. Modelos lineales sobre series de tiempo

Si consideramos las observaciones de nivel del agua como un conjunto de observaciones secuenciales en el tiempo, dichos datos forman una serie de tiempo. Según [Box and Jenkins, 1994], un aspecto intrínseco de una serie de tiempo es que observaciones adyacentes son típicamente dependientes, lo que permite aplicar distintas técnicas de análisis.

El uso en un momento  $t$  de observaciones disponibles de una serie de tiempo para pronosticar su valor en un momento futuro  $t + 1$  provee la base para procesos como planificación de producción y control y producción de inventario entre otros. Los pronósticos son generalmente necesarios para un período conocido como tiempo hacia adelante (o *lead time* por su nombre en inglés), que varía para cada problema.

Suponemos que las observaciones se disponen en intervalos de tiempo discretos y equidistantes  $z_1, z_2, \dots, z_t$ . Dichas observaciones podrán utilizarse para estimar tiempos hacia adelante  $l = 1, 2, \dots, n$ . Definimos  $\hat{z}_t(l)$  al pronóstico realizado en el momento  $t$  de la observación  $z_{t+l}$  del momento  $t + l$ . El objetivo será obtener una función de predicción tal que la media de los desvíos cuadrados  $z_{t+l} - \hat{z}_t(l)$  entre el valor real y el predicho sea lo menor posible para cada tiempo  $t$ .

**Series estacionarias.** Muchas de las series de tiempo representativas de fenómenos naturales exhiben comportamientos que se mantienen en equilibrio en un nivel medio constante. Estas series de tiempo se denominan estacionarias, y pueden ser descritas por su media, varianza y función de autocorrelación (o de manera equivalente, por su función de densidad espectral).

En particular, se dice que una serie es estrictamente estacionaria [Cryer and Kellet, 1991] si por cualesquiera sean los puntos de la serie  $t_1, \dots, t_n$  y cualquiera sea el retraso  $k$ , la distribución conjunta de  $x_{t_1}, \dots, x_{t_n}$  y  $x_{t_1-k}, \dots, x_{t_n-k}$  es la misma. Una definición relacionada, referida a veces como débil o estacionaria de segundo orden, es aquella que define a la serie como tal, si la media permanece constante y la autocorrelación solo depende de la posición relativa de los valores en la serie.

**Series estacionales.** La estacionalidad en una serie de tiempo se presenta como un patrón de cambios repetitivos en  $S$  intervalos de tiempo, donde  $S$  define el número de períodos de tiempo hasta que el patrón se repite nuevamente. Si por ejemplo en una serie con frecuencia mensual, valores altos suelen ocurrir en ciertos meses y valores bajos en otros meses,  $S = 12$  es el lapso del comportamiento periódico estacional.

### Modelos estocásticos

Con el fin de introducir la definición de modelo estocástico, incluiremos aquí el concepto de modelo matemático determinístico como aquel que describe el comportamiento de fenómenos físicos con exactitud.

Como se menciona en [Box and Jenkins, 1994], la deducción de un modelo totalmente determinístico probablemente no exista, dados factores desconocidos que pudieran alterar las condiciones. Sin embargo, es posible generar modelos que calculen la probabilidad de cierto valor futuro entre dos límites especificados. A dicho modelo se lo denomina modelo de probabilidad o *estocástico*.



**Modelos estocásticos estacionarios.** Una clase especial de los modelos estocásticos es la que asume que el proceso se mantiene en este equilibrio estadístico particular. Estos modelos son denominados modelos estacionarios.

### Modelo Autorregresivo (AR)

Un modelo estocástico de utilidad para la representación de determinadas series de tiempo es el modelo autorregresivo (o *AR* por sus siglas en inglés). En este modelo el valor actual es expresado como una combinación lineal finita de valores previos y un ruido  $a_t$ . Este ruido  $a_t$ , llamado ruido blanco o shock en tiempo  $t$ , denota nueva información en el momento  $t$  de la serie de tiempo, y se caracteriza porque sus valores en dos tiempos diferentes no guardan correlación.

Sean valores del modelo  $z_t, z_{t-1}, z_{t-2}, \dots$  en tiempos igualmente espaciados  $t, t-1, t-2, \dots$  sean  $\tilde{z}_t, \tilde{z}_{t-1}, \tilde{z}_{t-2}, \dots$  desvíos de  $\mu$ , tal que  $\tilde{z}_t = z_t - \mu$ , entonces:

$$\tilde{z}_t = \phi_1 \tilde{z}_{t-1} + \phi_2 \tilde{z}_{t-2} + \dots + \phi_p \tilde{z}_{t-p} + a_t$$

se denomina proceso autorregresivo de orden  $p$  o  $AR(p)$ , donde  $\mu$  es la media alrededor de la cual el proceso varía, siempre que el proceso sea estacionario y  $\phi_1, \phi_2, \dots, \phi_p$  son parámetros del modelo. El proceso es llamado autorregresivo ya que la variable  $z$  depende de valores previos de sí misma. El modelo contiene  $p + 2$  parámetros desconocidos  $\mu, \phi_1, \phi_2, \dots, \phi_p, \sigma_a^2$ , donde el parámetro adicional  $\sigma_a^2$  es la varianza del ruido blanco  $a_t$ .

Utilizando el operador de retroceso o *backshift*  $Bz_t = z_{t-1}$ , podemos reescribir la definición anterior como

$$\tilde{z}_t - \phi_1 \tilde{z}_{t-1} - \phi_2 \tilde{z}_{t-2} - \dots - \phi_p \tilde{z}_{t-p} = a_t$$

y aplicando la definición de operador autorregresivo de orden  $p$

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p$$

el proceso autorregresivo  $AR(p)$  se puede reescribir utilizando la notación polinómica como

$$\phi(B)\tilde{z}_t = a_t$$

### Modelo de Media Móvil (MA)

Otro de los modelos utilizados en el análisis de series de tiempo es el de media móvil (o *MA* por sus siglas en inglés). Este modelo especifica que la variable de salida  $\tilde{z}$  es linealmente dependiente de un número finito  $q$  de ruidos  $a$  previos. El proceso

$$\tilde{z}_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \dots - \theta_q a_{t-q}$$

se denomina proceso de media móvil de orden  $q$  o  $MA(q)$ . El modelo tiene  $q+2$  parámetros  $\mu, \theta_1, \dots, \theta_q, \sigma_a^2$  que se estiman a partir de los datos. Con la misma notación abreviada utilizada con el modelo AR, podemos escribir el proceso de media móvil de manera equivalente como

$$\tilde{z}_t = (1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q) a_t$$

o en notación polinómica equivalente

$$\tilde{z} = \theta(B) a_t$$

### Modelo combinado Autorregresivo - Media Móvil (ARMA)

Por último, a fin de alcanzar mayor flexibilidad en la generación del modelo, los modelos autorregresivos de media móvil (o ARMA) utilizan una combinación de los modelos anteriores:

$$\tilde{z}_t = \phi_1 \tilde{z}_{t-1} + \phi_2 \tilde{z}_{t-2} + \dots + \phi_p \tilde{z}_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \dots - \theta_q a_{t-q}$$

con  $p + q + 2$  parámetros desconocidos  $\mu, \phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q, \sigma_a^2$  (varianza del ruido  $a$ ) estimados a partir de los datos.

### Modelo combinado estacional (SARIMA)

Con frecuencia las series de tiempo exhiben características estacionales. Se dice que una serie exhibe comportamiento estacional con período  $S$  cuando ocurren similitudes en la serie luego de  $S$  intervalos básicos de tiempo [Box and Jenkins, 1994].

Si bien los modelos ARIMA pueden modelar datos con cierta tendencia, son limitados frente a series de tiempo con componentes estacionales. Para ello se utilizaron modelos estacionales o Seasonal ARIMA (SARIMA), una extensión de los modelos ARIMA que agrega términos estacionales específicos.

Los modelos SARIMA intentan modelar pronósticos a partir de datos univariados que contienen tendencias y estacionalidad. Por otro lado, muchas de estas series tienen influencia de variables externas o exógenas. En estos casos la implementación de modelos SARIMAX permiten incorporar dichas variables al modelo, lo que contribuye a aumentar la precisión del mismo. Como ejemplo en la predicción de alturas horarias en el Río de la Plata, la incorporación de variables exógenas como vientos y/o tablas de mareas podrían potencialmente contribuir a la explicación del comportamiento de la serie.

Siguiendo las ideas planteadas por [Box and Jenkins, 1994], este trabajo consideró para la construcción de los modelos lineales sobre las series de tiempo dos conceptos básicos:

- el **principio de parsimonia**, por el cual el enfoque utilizó el mínimo número posible de parámetros para representaciones adecuadas. Esta idea es similar al principio de

la navaja de Ockham, según el cual existiendo dos explicaciones para un evento, la que requiera la menor cantidad de supuestos es generalmente la correcta.

- un **enfoque iterativo** en la selección del modelo, utilizando conocimiento teórico incompleto para identificar la clase de modelos apropiados a partir de los datos experimentales (Fig.2.9).

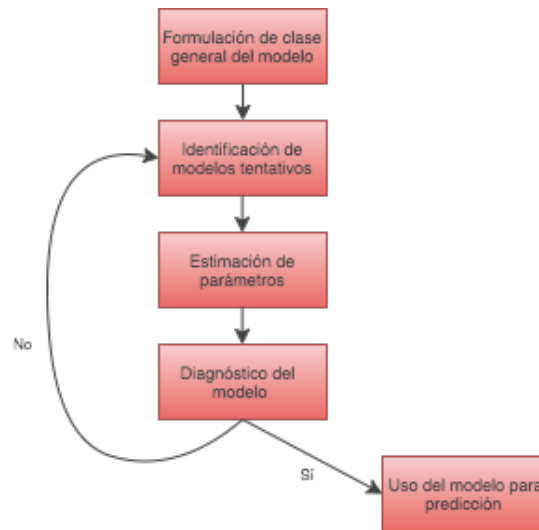


Figura 2.9: Proceso iterativo de selección de modelo lineal

Los modelos anteriormente presentados en esta sección proveen una clase de modelos adecuados para la representación de series de tiempo. Este enfoque iterativo, basado en la identificación de modelos, estimación de parámetros y diagnóstico desarrollado en las secciones siguientes, permitió la selección del mejor modelo representativo de los datos.

### 2.3.3. Redes neuronales

Si bien los modelos que comprenden combinaciones lineales poseen propiedades analíticas útiles, su aplicación práctica en el reconocimiento de patrones es limitada. Esto ocurre, por ejemplo, en casos prácticos donde se deben explorar espacios de alta dimensionalidad que comprenden una gran cantidad de variables de entrada.

Dicha dificultad, presente en estos espacios, es generalmente conocida como *maldición de la dimensionalidad*, y es un factor clave al momento de diseñar las técnicas de reconocimiento de patrones. La complejidad se hace evidente en numerosos problemas al considerar sus extensiones a problemas con mayor número de variables de entrada. En casos con enfoques simples a problemas de baja dimensionalidad, las problemáticas tienden a un crecimiento en el orden de potencias o exponencial, al tiempo que la dimensionalidad del espacio aumenta.

Sin embargo, esto no impide encontrar técnicas de reconocimiento efectivas aplicables a estos espacios. Esto se debe a que:

- los datos reales con frecuencia estarán circunscriptos a una región del espacio de menor dimensionalidad, y en particular, las direcciones sobre las cuales variaciones

importantes en las variables dependientes ocurren, también pueden estar circunscritas;

- los datos reales en general muestran ciertas propiedades de suavizamiento tal que en general pequeños cambios en las variables de entrada producirán pequeños cambios en las variables de salida, lo que permite el uso de técnicas de interpolación que ayuden a predecir nuevos valores

Los modelos lineales descritos en la sección 2.3.2 abarcan combinaciones lineales de funciones base fijas  $\phi_j(X)$ . Para poder aplicar estos modelos a problemas que presentan la complejidad mencionada, es necesario adaptar las funciones base a los datos. El objetivo es extender este modelo haciendo que las funciones base  $\phi_j(X)$  dependan de parámetros y permitir que estos parámetros sean ajustados juntamente con los coeficientes  $w_j$  durante el entrenamiento. El modelo básico de red neuronal puede describirse así como un conjunto de transformaciones.

Como primer paso se construyen  $M$  combinaciones lineales de las variables de entrada  $x_1, \dots, x_D$  de la siguiente forma

$$a_j = \sum_{i=1}^D w_{ji}^{(1)} x_i + w_{j0}^{(1)} \quad (2.1)$$

donde  $j = 1, \dots, M$  y el superíndice (1) indican que los parámetros correspondientes están en la primera «capa» de la red. Los parámetros  $w_{ji}^{(1)}$  se denominan *pesos* y los parámetros  $w_{j0}^{(1)}$  *biases*. Los resultados  $a_j$  se denominan *activaciones*. Cada uno de ellos es a su vez transformados utilizando una función de activación  $h(\cdot)$ , lo que resulta

$$z_j = h(a_j)$$

Estas cantidades corresponden a las salidas de las funciones base en (2.1) llamadas *unidades ocultas* en el contexto de redes neuronales. Estas funciones no lineales  $h(\cdot)$  son en general funciones sigmoideas, como la función logística sigmoideal o la tanh. Estos nuevos valores son a su vez combinados en forma lineal

$$a_k = \sum_{i=1}^M w_{ki}^{(2)} x_i + w_{k0}^{(2)}$$

dando como resultado *activaciones de unidad de salida*, donde  $k = 1, \dots, K$  y  $K$  es el número total de salidas. Finalmente, las activaciones de unidad de salida son transformadas usando la función de activación apropiada para dar como resultado las salidas de la red  $y_k$ . La elección de la función de activación está determinada por la naturaleza de los datos. En el caso de problemas de regresión lineal, dicha función de activación es la identidad, por lo que  $y_k = a_k$ .

El proceso completo se puede interpretar como una propagación de información hacia adelante en la red (o *feed forward propagation* por su denominación en inglés).

Una característica clave es el uso de funciones sigmoidales no lineales continuas, lo que implica que la función de la red es diferenciable con respecto a los parámetros de la red, lo que juega un rol fundamental en el entrenamiento.

Dentro de las redes neuronales, las redes neuronales recurrentes (o *RNN* por sus siglas en inglés) constituyen una clase de redes particularmente aptas para el modelado de datos secuenciales como son las series de tiempo. A alto nivel, una RNN itera sobre los pasos de la secuencia a la vez que mantiene un estado interno que codifica información acerca de los pasos recorridos hasta ese momento.

De forma más ejemplificada, en cada instante de tiempo una neurona recurrente recibe la entrada  $x$  de la capa anterior así como su propia salida en el instante de tiempo anterior para generar su salida  $y$ . Matemáticamente el proceso de acarreo de memoria hacia adelante puede describirse de la siguiente manera:

$$h_t = \phi(Wx_t + Uh_{t-1})$$

donde  $h_t$  es el estado oculto en el momento  $t$ . Este es una función de la entrada en el mismo momento  $x_t$ , modificada por la matriz de pesos  $W$  sumada al estado oculto del momento anterior  $h_{t-1}$  multiplicado por su propia matriz de pesos  $U$  también conocida como matriz de transición. El error generado retorna en forma de propagación hacia atrás (o *back propagation* por su denominación en inglés) y es utilizado para ajustar los pesos hasta que el error no puede ser más reducido. La suma es abarcada por la función  $\phi$ , la cual es una función logística sigmoidea o tanh a fin de facilitar el manejo de gradientes en dicha propagación hacia atrás.

Debido a que este ciclo ocurre en cada momento de tiempo de la serie, cada estado oculto contiene rastros del estado oculto anterior y de todos los precedentes mientras la memoria persista (Fig. 2.10)

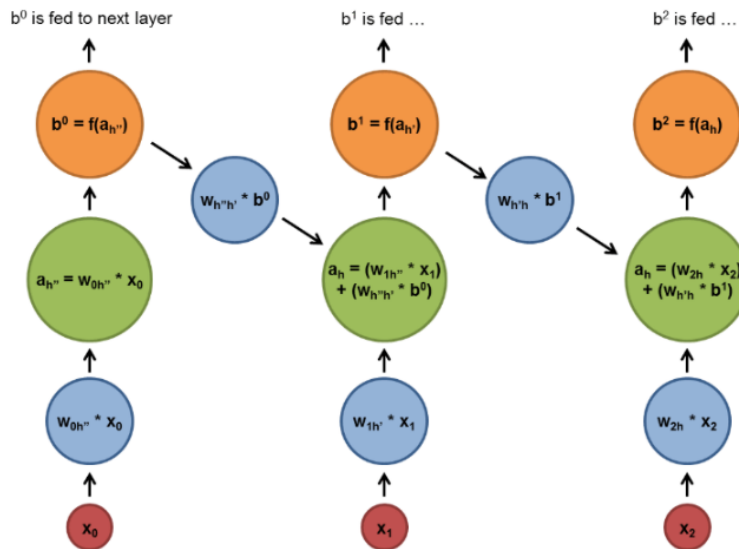


Figura 2.10: Redes neuronales recurrentes - (Cortesía de imgur)

Uno de los problemas de performance de las RNN es el de los gradientes apagados. El gradiente expresa el cambio de los pesos en relación al cambio en el error. Si no conocemos el gradiente, no podremos ajustar los pesos en la dirección que disminuya el error, y la red deja de aprender.

En la búsqueda para establecer conexiones entre una salida final y eventos ocurridos muchos pasos atrás, las RNN muchas veces se estancan ya que es muy difícil saber cuánto peso se debe asignar a entradas remotas. Esto se debe en parte a que la información que fluye por una red neuronal atraviesa varias fases de multiplicación, donde las derivadas son susceptibles de desaparecer o hacer «explosión».

Para el caso en el que las derivadas hacen esta explosión, la solución es truncar o aplastar los gradientes, pero la solución no es tan sencilla en el primer caso.

Cuando los gradientes tienden a desaparecer, tienden a ser demasiado chicos para que la red continúe aprendiendo. La propuesta presentada en [Hochreiter and Schmidhuber, 1997] implementó una capa RNN específica llamada memoria de corto largo plazo (*Long Short Term Memory* o LSTM). Esta capa mantiene el error que se propaga hacia atrás en el tiempo y capas. Al mantener un error más constante, la red continúa aprendiendo por muchos pasos, abriendo un canal que relaciona causas y efectos de manera remota. La Fig. 2.11 muestra el flujo de los datos a través de una célula de memoria en LSTM.

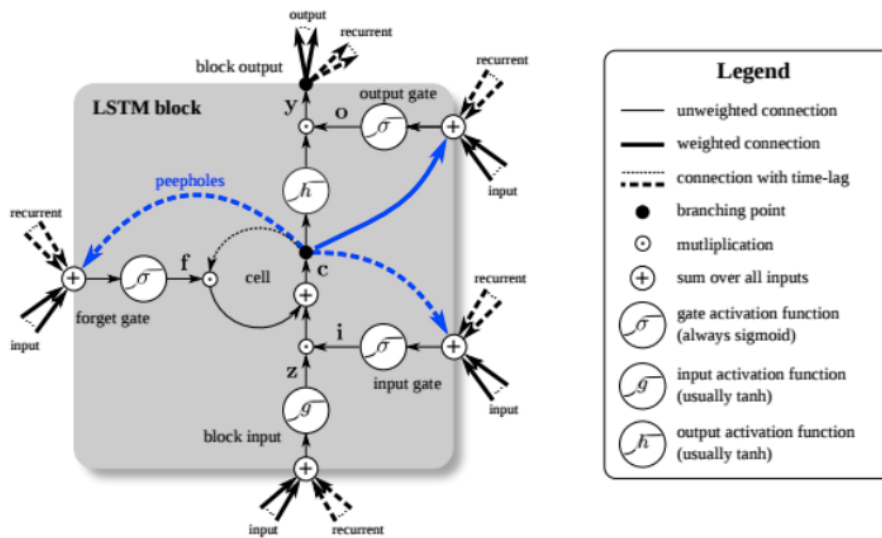


Figura 2.11: Bloque LSTM (Long Short-Term Memory)

La combinación de entrada presente y estado de la célula pasada alimenta a la célula misma así como a sus tres puertas, que deciden cómo se manejará la entrada.

Los círculos negros son las puertas en sí que determinan si permiten la entrada, borran el estado presente de la célula y/o permiten que el estado afecte la salida de la red en el paso presente. El signo (+) central es el concepto básico de las LSTMs. Esto es lo que permite mantener un error constante al momento de ser propagada hacia atrás. En lugar de calcular el siguiente estado de la célula multiplicando el estado actual por la nueva entrada, estos se suman.

## 2.4. Procedimientos de la generación de modelos

### 2.4.1. Conjuntos de entrenamiento y evaluación

La validación cruzada es un método de evaluación de modelos que proveen una estimación de qué tan bien el modelo va a predecir cuando se le presenten datos que no haya visto aún. Para entrenar el modelo se usa una parte del conjunto de datos, removiendo parte de los datos antes del proceso. Cuando el entrenamiento se completa, los datos removidos se utilizan para evaluar el rendimiento del modelo sobre nuevos datos. Esta es la idea básica de los métodos denominados de validación cruzada (*Cross validation* o *CV* por su nombre en inglés). Esta habilidad del modelo para categorizar nuevos ejemplos se denomina generalización [Bishop, 2006], parte central del reconocimiento de patrones.

La forma más simple de estos métodos es la CV con retención (o *hold-out* por su traducción en inglés). El conjunto de datos se divide en un conjunto llamado de entrenamiento y otro conjunto llamado de evaluación (Fig. 2.12).

Para el ajuste y selección de parámetros se calculan y comparan sobre el conjunto de entrenamiento criterios de información y errores obtenidos, seleccionando los mejores de acuerdo a los criterios que se explican en la siguiente sección. Finalmente, se calcula el error en el conjunto de evaluación.

La evaluación puede tener una alta varianza, ya que depende fuertemente de los datos que quedaron en el conjunto de entrenamiento y en el de evaluación, por lo que las pruebas pueden ser significativamente diferentes dependiendo de la forma en que se hizo la división.



Figura 2.12: Ejemplo de CV con retención

Al tratarse de series de tiempo, ciertas técnicas de validación cruzada generalmente utilizadas en la separación del conjunto de datos no son las apropiadas en estos casos. Esto se debe a que:

- al trabajar con series de tiempo, se debe realizar una división de los datos de manera tal de prevenir pérdida de información. Para simular de manera precisa condiciones de predicción reales, se deben mantener todos los datos siguientes a los eventos usados para entrenar el modelo. Esto asegura que las dependencias temporales se mantienen. Por esto, para series de tiempo se utiliza validación cruzada con retención, donde el conjunto de evaluación sigue cronológicamente al conjunto de entrenamiento, y el de validación al subconjunto de entrenamiento
- la selección de un conjunto de datos de evaluación arbitrario puede conducir a que el cálculo de error sobre este subconjunto sea una estimación pobre del error en un subconjunto de evaluación independiente. Para resolver este punto, se utiliza un método llamado validación cruzada anidada (o *Nested CV* en inglés). Esta validación contiene un ciclo externo para estimar el error y un ciclo interno para ajustar

parámetros como se menciona anteriormente. El ciclo externo divide al conjunto de datos en múltiples subconjuntos de entrenamiento y evaluación, y el error en cada paso se promedia a fin de obtener una estimación del error del modelo robusta.

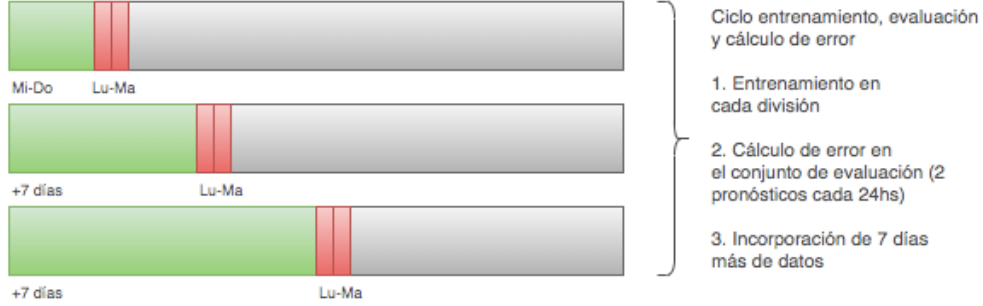


Figura 2.13: Validación cruzada anidada de origen móvil con recalibración

En particular, es de interés el método de validación cruzada anidado de origen móvil con recalibración (en inglés *rolling-origin-recalibration*) [Bergmeir and Benítez, 2012]. Este método considera cada período  $p$  siguiente como conjunto de evaluación, asignando los valores del período de evaluación previo al conjunto de entrenamiento (Fig. 2.13).

#### 2.4.2. Evaluación y selección de modelos

La evaluación del rendimiento de la generalización es de gran importancia en la práctica, en la medida que ofrece una medida de la calidad del modelo seleccionado, permite comparar modelos y orienta la selección de un método de aprendizaje o modelo.

Entre las distintas métricas sugeridas para seleccionar modelos estadísticos, este trabajo utilizó distintos criterios:

- Criterio de Información Akaike (AIC)
- Error Absoluto Medio (MAE)
- Raíz del Error Cuadrático Medio (RMSE)
- Error de Predicción Acumulado (APE)

#### Criterio de Información Akaike

AIC pertenece a una familia de estadísticas de ajuste de modelos penalizados (que se denominan "AIC") aplicadas a modelos de regresión ajustados por máxima verosimilitud, de la forma:

$$AIC_j = -2\log_e L(\hat{\theta}_j) + cs_j$$

donde  $L(\hat{\theta}_j)$  es la máxima verosimilitud bajo el modelo  $M_j$ ;  $\theta_j$  es el vector de parámetros del modelo, incluyendo por ejemplo, los coeficientes de regresión y un error de varianza



o parámetro de dispersión, y  $\hat{\theta}_j$  es el vector de estimaciones por máxima verosimilitud (MLEs) de los parámetros;  $s_j$  es el número de parámetros en  $\theta_j$ ; y  $c$  es una constante que difiere de un criterio de selección de modelo a otro. El término  $-2\log_e L(\hat{\theta}_j)$  es el desvío residual del modelo (o diferencia del desvío residual por una constante). La *magnitud* de \*IC generalmente no es interpretable, el interés se enfoca en las diferencias entre valores de los distintos modelos. Nuestro interés se enfocará en los modelos con menor \*IC, que son aquellos que tienen mayor soporte de los datos.

AIC se define de la siguiente forma [Fox, 2015]:

$$AIC_j = -2\log_e L(\hat{\theta}_j) + 2s_j$$

Para modelos lineales con errores normales, el MLE de los coeficientes de regresión es el estimador de cuadrados mínimos y el MLE de la varianza de error es  $\hat{\sigma}_e^{j^2} = (\sum E_i^{(j)^2})/n$ , donde  $E_i^{(j)}$  son los residuos mínimos cuadrados del modelo  $M_j$ , y:

$$AIC_j = n\log_e \hat{\sigma}_e^{(j^2)} + 2s_j$$

El error calculado en el conjunto de entrenamiento

$$\overline{err} = \frac{1}{N} \sum L(y_i, \hat{f}(x_i))$$

será menor que el error de generalización, ya que los mismos datos son usados para ajustar el algoritmo y evaluar el error [Hastie et al., 2009]. El algoritmo de ajuste generalmente se adapta a los datos de entrenamiento y por lo tanto el error aparente o de entrenamiento es una estimación sobre optimista del error de generalización. En parte esta discrepancia se debe a que los vectores de entrada del conjunto de prueba pueden no coincidir con los vectores de entrenamiento.

Generalmente este error *in-sample* no es el interés final, pero a fin de comparar modelos es un método efectivo de selección, ya que lo importante es el tamaño relativo del error (vs el absoluto).

Dentro de cada clase de modelo se seleccionó aquel con menor AIC. La comparación final entre los mejores modelos de cada clase utilizó la métrica de Error de Predicción Acumulado que se describe a continuación.

### Error Absoluto Medio (MAE)

Esta es una medida común del error de pronóstico en análisis de series de tiempo. Se calcula como la diferencia entre dos variables continuas. Dadas dos series de datos ( $X$  e  $Y$ ) relativas a un mismo fenómeno, el MAE es la distancia vertical promedio entre cada uno de los puntos y la recta identidad ( $y = x$ ), o también la distancia horizontal promedio entre cada punto y la recta identidad:

$$MAE = \frac{\sum_{j=1}^n (y_j - x_j)}{n}$$

El MAE tiene una interpretación clara como la diferencia absoluta promedio entre  $y_j$  y  $x_j$ , y es habitualmente utilizado ya que utiliza la misma escala que los datos que se están midiendo.

El MAE fue utilizado en la comparación de los resultados presentados en este trabajo con los obtenidos primeros modelos aún productivos a la fecha. Si bien no se obtuvieron pronósticos específicos de estos modelos, se realizó un análisis del MAE obtenido en distintos horizontes temporales.

### **Raíz del Error Cuadrático Medio (RMSE)**

El RMSE mide también la magnitud promedio del error. Es la raíz cuadrada del promedio de las diferencias cuadradas entre predicción y observación.

$$RMSE = \sqrt{\frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2}$$

El RMSE expresa el error en unidades de la variable de interés y en ambos valores menores son mejores. En el caso de RMSE, como el error es potenciado antes de ser promediado, el cálculo asigna pesos relativamente altos a errores grandes. Esto lo convierte en una métrica útil cuando esta clase de errores son indeseables.

Desde la interpretación, RMSE no describe el error promedio solamente y tiene implicancias que son más difíciles de interpretar.

### **Error de Predicción Acumulado (APE)**

Dentro de los modelos seleccionados, aquel con mejor capacidad de separar la estructura replicable del ruido, es el que tendrá el mejor rendimiento predictivo. El criterio más ampliamente aplicado para la selección de modelo es el de maximización de la generalización, o de forma equivalente, el de minimización del error de predicción de datos futuros [Wagenmakers et al., 2006].

El método APE utilizado para la selección final del modelo, está basado en la intuición que un buen indicador del error de predicción para datos futuros es la suma (o de forma equivalente, el promedio) de los errores de predicción previos. Considerando una serie de tiempo de  $n$  observaciones  $x^n = (x_1, x_2, \dots, x_n)$ , el método APE calcula pronósticos de  $k$ -pasos basado en conjuntos de datos que aumentan gradualmente. Por ejemplo, para calcular el APE de un modelo  $M_j$  los pasos son los siguientes:

1. Determinar la menor cantidad de observaciones  $s$  que permiten identificar el modelo ( $i = s + k$ , tal que  $i - k = s$ )
2. En base a las primeras  $i - k$  observaciones, calcular la predicción  $\hat{p}_i$  para la observación  $i$
3. Calcular el error de predicción para la observación  $i$  (por ejemplo, el error cuadrático entre la observación  $x_i$  y el valor predicho  $\hat{p}$ )

4. Incrementar  $i$  en  $l$  (donde  $l$  puede o no ser igual a  $k$ ) y repetir los pasos 2 y 3 hasta que  $i = n$
5. Sumar todos los errores de predicción de  $k$ -pasos calculados en el paso 3. El resultado es el APE (Fig. 2.14)

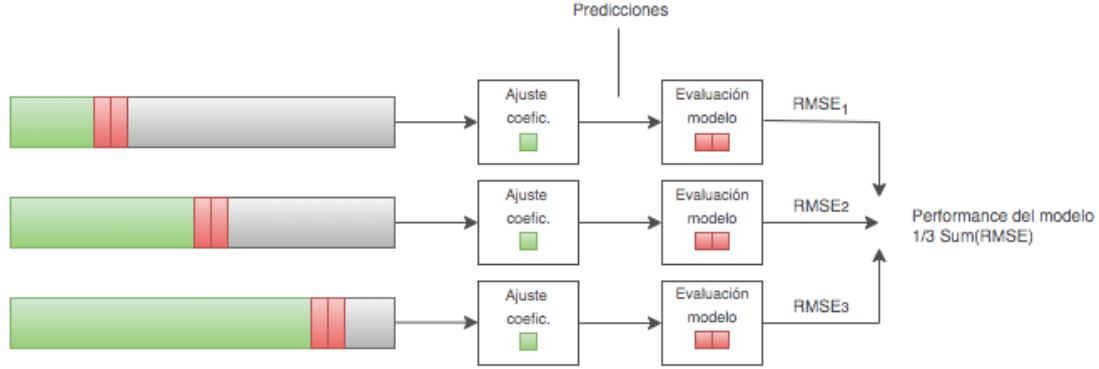


Figura 2.14: Cálculo del error de predicción acumulado (APE)

Este procedimiento se completó en forma separada para los distintos modelos, seleccionando el modelo con menor APE de todos. En forma resumida, podemos decir que el APE está definido como:

$$APE(M_j) = \sum_{i=s+l}^n d[x_i, (\hat{p}_i | x^{i-k})]$$

donde  $d$  indica la función específica que cuantifica la discrepancia entre la observación y la predicción.

## 2.5. Software utilizado

En el procesamiento, análisis de datos y modelado se utilizó Python v3.7 bajo ambiente Jupyter Notebook. Jupyter Notebook es una aplicación web de código abierto que permite crear y compartir documentos y generar salidas interactivas, con amplia integración de herramientas de Big Data.

Las librerías Python específicas de análisis, procesamiento y modelado de los datos usadas en esta tesis incluyen:

- **sklearn**: librería específica de aprendizaje automático, hace uso de otras librerías como numpy y matplotlib. Esta herramienta implementa modelos para el análisis predictivo de datos, como regresiones, clustering y clasificación, además de métodos para el cálculo de coeficientes de determinación y error cuadrático medio entre otros. Utilizada también para llevar las variables del conjunto de datos a una escala similar.
- **statsmodel**: librería para la generación de diversos modelos y tests estadísticos. Esta librería fue utilizada específicamente para el análisis y generación de modelos de series de tiempo y los tests estadísticos de soporte.

- tensorflow: librería de aprendizaje automático específica para la construcción y entrenamiento de redes neuronales.
- matplotlib: librería para la generación de gráficos a partir de estructuras de datos.
- seaborn: librería para visualización de datos y elaboración de gráficos estadísticos basada en matplotlib.
- numpy: librería para el manejo de arrays y matrices multi-dimensionales, junto con una extensa colección de funciones matemáticas de alto nivel para operar sobre los mismos.
- pandas: extensión de la librería Numpy para manipulación y análisis de datos, con métodos de conveniencia para manejo de estructuras de datos y herramientas de análisis específicos

Los archivos grib2 de predicciones de Presión y Vientos (formato comúnmente usado para datos climáticos) provistos por el National Centers for Environmental Prediction (NCEP) fueron procesados utilizando librerías C y el utilitario wgrib2, herramienta también desarrollada en C por el National Oceanic and Atmospheric Administration (NOAA).

La manipulación de los archivos csv generados a partir de estos archivos y la extracción de los puntos geográficos de interés se realizó en bash script bajo iOS High Sierra v10.13.6.

Para la generación de mapas y ubicación de las referencias geográficas seleccionadas se utilizó el Sistema de Información Geográfica (SIG) QGIS3 v3.4.5-Madeira licenciado bajo la GNU (General Public License).

El proyecto completo (datos, código fuente y recursos adicionales generados) se encuentra disponible en el siguiente repositorio: <https://gitlab.com/lolafrias/prediccionmareardp>

## Capítulo 3

# Análisis exploratorio y creación de variables

### 3.1. Análisis exploratorio

A continuación se describen los procedimientos para obtener los atributos utilizados en la generación de los modelos.

Los datos de entrada fueron preprocesados transformándolos en un nuevo espacio de variables, con el objeto de que el problema de reconocimiento de patrones fuera uno más simple de resolver. Este preprocesamiento, también llamado extracción de características, tiene entre otros beneficios acelerar los tiempos de cómputo.

Por otro lado, en determinados casos donde se facilitaba el aprendizaje de los algoritmos, se aplicaron métodos para preprocesar los datos llevando las variables a una escala o rango de valores similar. Facilitar en este contexto significa que los algoritmos funcionen mejor o converjan más rápido, como es el caso de las regresiones lineales y las redes neuronales.

Siendo la escala aplicada (0,1), el cálculo sobre cada valor  $x_i$  de una variable es el siguiente:

$$x_{si} = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

Con respecto a los valores perdidos, en varios de los conjuntos de datos se hallaron períodos consecutivos de faltantes. Esto obligó a acotar los datos a aquellos años sin faltantes o con faltantes donde fue posible aplicar técnicas de interpolación que no introdujeran distorsión en los datos. En estos casos se aplicó el tipo de interpolación spline, una interpolación dentro de las llamadas *spline* que aplica una función polinomial compuesta por múltiples subfunciones que corrigen oscilaciones entre puntos que pueden darse al utilizar polinomios de alto grado.

#### 3.1.1. Alturas horarias

Como se mencionó anteriormente, la marea astronómica muestra cierta regularidad predecible asociada con un régimen semidiurno (Fig.3.1).

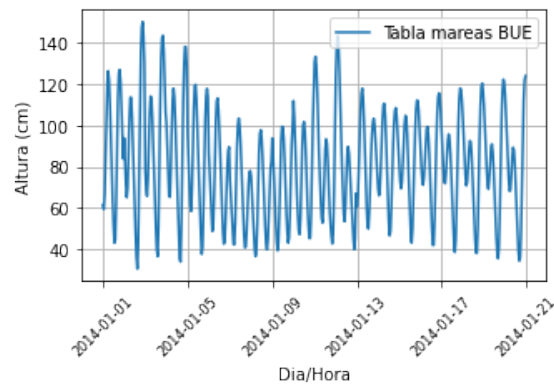


Figura 3.1: Tabla mareas Buenos Aires - Enero 2014

El efecto astronómico sobre el nivel del Río de la Plata puede observarse al comparar esta serie con las tablas de marea, destacándose picos en las alturas primariamente asociados con vientos intensos y persistentes.

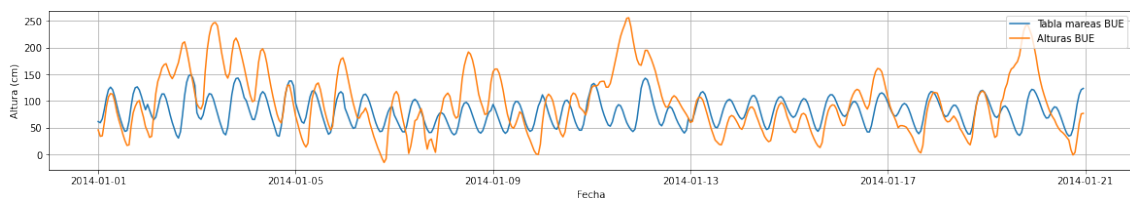


Figura 3.2: Alturas horarias y tabla mareas Buenos Aires - Enero 2014

Por otro lado, factores como el viento tiene un efecto meteorológico que afecta también las alturas del agua, pudiendo generar en ocasiones ondas de tormenta. En particular, Noviembre de 2014 mostró aumentos significativos en las alturas del agua ( $> 3, 10$  mts), coincidentes con registros de sudestadas ocurridas a principios de ese mes (Fig.3.3).

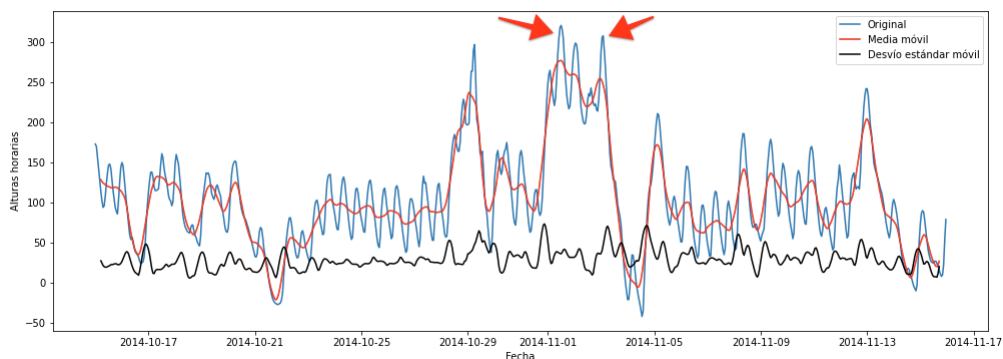


Figura 3.3: Valores extremos de alturas horarias Buenos Aires - Noviembre 2014

Analizando aspectos característicos principales, la mayoría de los modelos aplicados a series de tiempo asume que la series son estacionarias, es decir, que propiedades estadísticas de la serie no cambian a lo largo del tiempo. Estas propiedades se verificaron mediante dos procedimientos:

- graficando estadísticos móviles. Se graficó la media y el desvío estándar móviles para comprobar si existían o no variaciones en el tiempo.

En cualquier instante  $t$ , se tomó el promedio y el desvío estándar centrados de  $n$  observaciones anteriores. En este caso se tomó una ventana de tiempo de 12 observaciones (equivalente a medio día) con el fin de observar cualquier efecto asociado con mareas semi-diurnas. Se utilizó el desvío en lugar de la varianza para mantener la misma unidad de medida que la media.

De forma visual, en la Fig.3.3 se observa que a lo largo del mes no hay una tendencia creciente o decreciente significativa en el desvío estándar o en la media móviles. Por otro lado, esta ventana de tiempo hace notorio un efecto estacional que será explorado durante la generación del modelo para analizar este factor específico.

- realizando el test de Dickey-Fuller, asumiendo como hipótesis nula que la serie es no-estacionaria:

El test estadístico de Dickey-Fuller se utilizó para chequear la condición estacionaria de la serie de tiempo, determinando la existencia de raíces unitarias. En un modelo autorregresivo de orden (1):

$$y_t = p y_{t-1} + u_t$$

donde  $y_t$  es la variable de interés,  $t$  es el índice de tiempo,  $p$  es un coeficiente y  $u_t$  es el término de error, el test se define entonces como:

$$\Delta y_t = (p - 1)y_{t-1} + u_t$$

donde la hipótesis nula es que la serie es no-estacionaria, es decir, que existe una raíz unitaria  $p = 1$ . Si el estadístico es menor que el valor crítico, se rechaza la hipótesis nula y podemos afirmar que la serie es estacionaria.

Para la serie completa, el test arrojó un valor = -23.076441, menor al valor crítico (1%) igual a -3.43, por lo que con un nivel de confianza del 99% se rechazó la hipótesis nula, confirmando que efectivamente la serie era estacionaria (Fig.3.4).

```

Dickey-Fuller Test:
Test estadístico          -23.076441
p-valor                   0.000000
#Obs hacia atrás usadas   52.000000
Número de observaciones usadas 35011.000000
Valor crítico (1%)       -3.430537
Valor crítico (5%)       -2.861623
Valor crítico (10%)      -2.566814
dtype: float64

```

Figura 3.4: Test de Dickey-Fuller (Bs As 2014-2017)

Por otro lado, el análisis de autocorrelación de la serie, que determina la existencia de correlaciones significativa entre la serie y valores desplazados de la misma, mostró que existe una correlación positiva con  $n$  valores anteriores, siendo  $n > 18$ , lo que equivale a decir, que para cada momento  $t$  existe una correlación entre  $x_t$  y  $x_{t-1}, x_{t-2} \dots x_{t-n}$  en un intervalo de confianza del 99%, mostrando la dependencia entre una observación  $x_t$  y dichas observaciones anteriores. El hecho de que  $n > 18$  sería indicio de la presencia de factores estacionales, por lo que en un siguiente paso la serie fue diferenciada.

Al tratarse de series de tiempo donde se asume un modelo de media móvil, la generación de los intervalos de confianza sigue la fórmula de Barlett:

$$\pm z_{1-\alpha/2} \sqrt{\frac{1}{N} \left( 1 + 2 \sum_{i=1}^k r_i^2 \right)}$$

donde  $k$  es el dato pasado. Se observa que los intervalos de confianza aumentan a medida que el retraso aumenta.

En particular, se observó que para cada momento  $t$  la correlación entre la observación  $x_t$  y las precedentes  $x_{t-24}$  muestra un aumento significativo, que sugiere una correspondencia con una estacionalidad diaria.

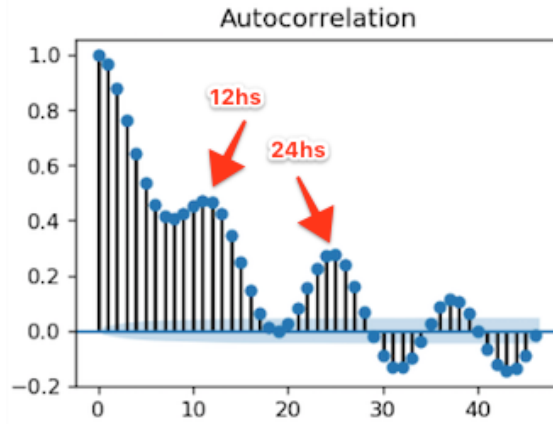


Figura 3.5: Autocorrelación de alturas horarias Buenos Aires (2014-2017)

### Datos faltantes y valores atípicos

La falta de registros de alturas observadas en intervalos de tiempo consecutivos prolongados obligó a descartar algunas estaciones mareológicas del dominio seleccionado para este trabajo. En el caso de Buenos Aires, se contó con registros consecutivos completos durante el período abarcado. En los restantes casos, aplicar métodos de interpolación sobre series de tiempo habría introducido distorsiones en el análisis posterior, por lo que estos datos fueron descartados.

En referencia a las alturas horarias de Buenos Aires, se observó la presencia de anomalías en la serie de tiempo, evaluándose la aplicación de una transformación de Winsor. Esta transformación limita los valores extremos en los datos y reduce los efectos causados por los valores atípicos. Dichos valores son fijados a un percentil específico de los datos.

A fin de analizar en detalle estos valores atípicos, se corrió una prueba aplicando una transformación de un 99.95 %, fijando los datos por debajo del percentil 0.05 al percentil 0.05 y por encima del percentil 99.95 al 99.95.

El boxplot de la serie de tiempo original y resultante se muestran en la Fig. 3.6.

Sobre la nueva serie, se aplicó una operación de convolución lineal discreta. Esta operación suaviza las fluctuaciones en plazos cortos, y resalta tendencias en plazos largos. La convo-



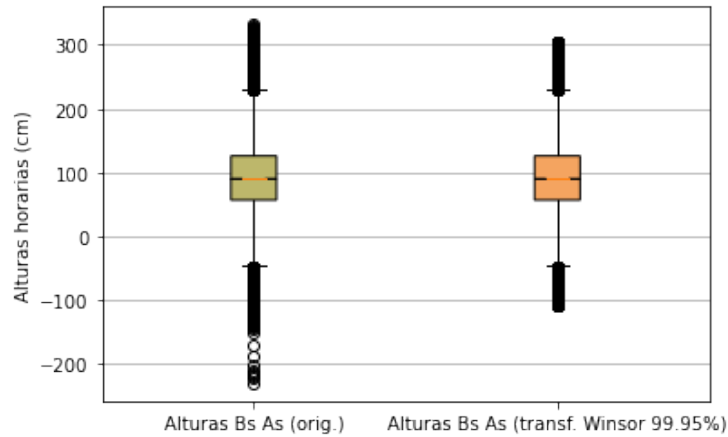


Figura 3.6: Boxplot de serie original y transformada de alturas Buenos Aires (2011-2018)

lución modela el efecto de un sistema de invariación lineal en el tiempo sobre la serie. La operación, definida como:

$$(a * v)[n] = \sum_{m=-\infty}^{\infty} a[m]v[n - m]$$

calcula una media móvil central, utilizando datos anteriores y posteriores del cual se quiere obtener la media. Se definió una ventana móvil de tamaño  $n=7$ , considerando 3 observaciones anteriores y posteriores a fin de detectar posibles errores del mareógrafo. Como se observa en la Fig.3.7, la nueva serie suaviza valores extremos.

Sin embargo, analizando puntualmente los períodos con anomalías detectadas, en cada uno de ellos los valores mostraban una secuencia aumento-disminución (y viceversa) ordenada en el tiempo, dando la pauta que no se debieron a errores de los mareógrafos sino efectivamente a eventos reales.

Por otro lado, los eventos fueron coincidentes con distintos eventos meteorológicos, como la bajante ocurrida en Agosto/2014 ([Agencia de Protección Ambiental, 2014]), por lo que se asumió que los valores detectados fueron válidos, descartándose la serie transformada y conservándose los valores originales.

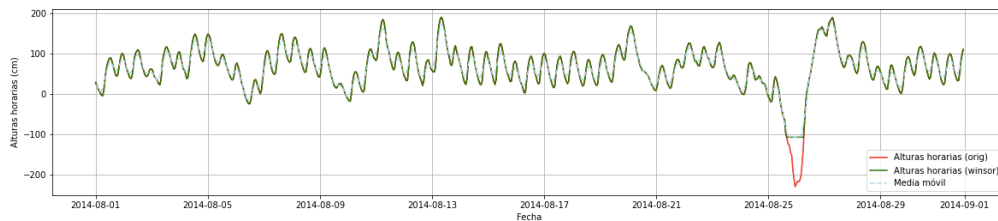


Figura 3.7: Alturas horarias y media móvil Buenos Aires

### 3.1.2. Tablas de marea

En un primer análisis a partir de los pronósticos de marea, los datos correspondientes a la costa norte y sur muestran una clara diferencia de amplitud, que se puede observar en

el rango mayor de las mareas de puntos como el puerto de San Clemente y Santa Teresita en comparación con Montevideo y Colonia (Fig.3.8).

Esta diferencia de amplitud sin embargo, disminuye a medida que nos desplazamos hacia la parte superior del Río de la Plata. Como se muestra en el Cuadro 3.1, el desvío estándar disminuye desde San Clemente (0.3217) en la parte inferior del Río de la Plata, pasando por Oyarvide (0.2807) hasta Pilote Norden (0.2107). La disminución en la diferencia también se puede ver en la Fig.3.9, donde la diferencia de alturas entre San Clemente y Montevideo (3.9a) es mucho mayor que hacia la parte superior del Río de la Plata, a la altura de Buenos Aires y Colonia (3.9b).

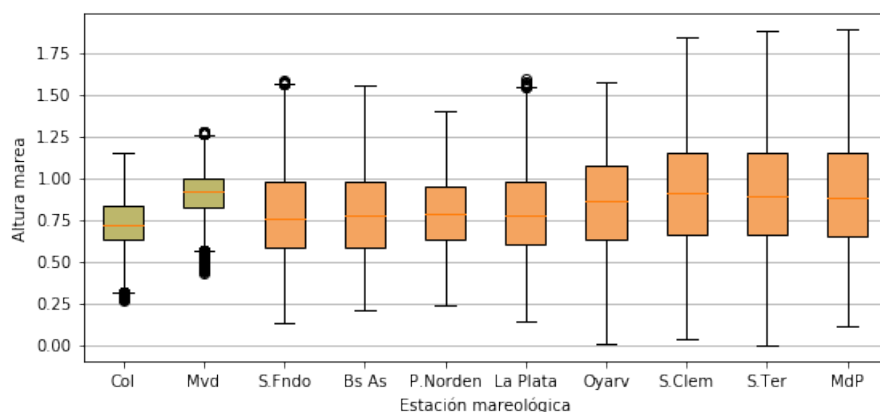


Figura 3.8: Boxplot de tablas de marea (AR y UR)

|         | Estaciones |              |               |          |          |              |              |               |          |          |
|---------|------------|--------------|---------------|----------|----------|--------------|--------------|---------------|----------|----------|
|         | San Frndo  | Buenos Aires | Pilote Norden | La Plata | Oyarvide | San Clemente | Sta Teresita | Mar del Plata | Colonia  | Mvdo     |
| media   | 0.790008   | 0.790376     | 0.790067      | 0.790762 | 0.850195 | 0.910182     | 0.909974     | 0.911439      | 0.729962 | 0.910003 |
| dsv est | 0.262329   | 0.253566     | 0.210710      | 0.255233 | 0.280744 | 0.321767     | 0.332920     | 0.328954      | 0.153228 | 0.131027 |
| mín     | 0.132118   | 0.205299     | 0.239335      | 0.140326 | 0.006919 | 0.036711     | -0.007718    | 0.108255      | 0.264210 | 0.430530 |
| máx     | 1.581246   | 1.552149     | 1.400636      | 1.590516 | 1.573253 | 1.844059     | 1.883398     | 1.892048      | 1.148045 | 1.277007 |

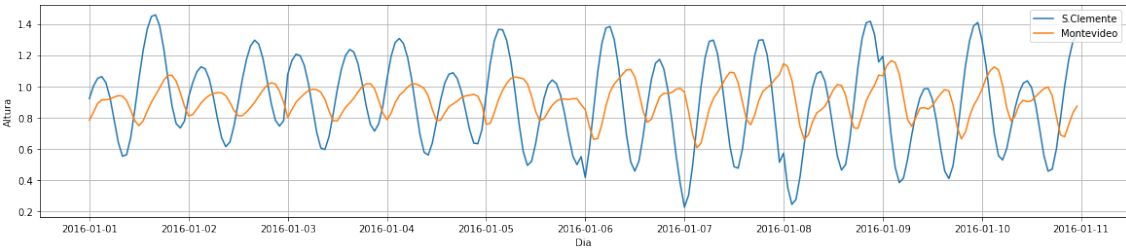
Cuadro 3.1: Estadísticos de tablas de marea de los puertos referenciados

En relación a la bajamar y pleamar en distintas estaciones del Río de la Plata, el gráfico de mareas entre Oyarvide y La Plata (Fig.3.10) mostró un desplazamiento de onda de aproximadamente 6hs entre ambos puntos. En particular, este desplazamiento se estima en aproximadamente 4hs/5hs hasta Buenos Aires, dato que se utilizó posteriormente para generar nuevas variables predictoras del nivel del agua en el punto de interés.

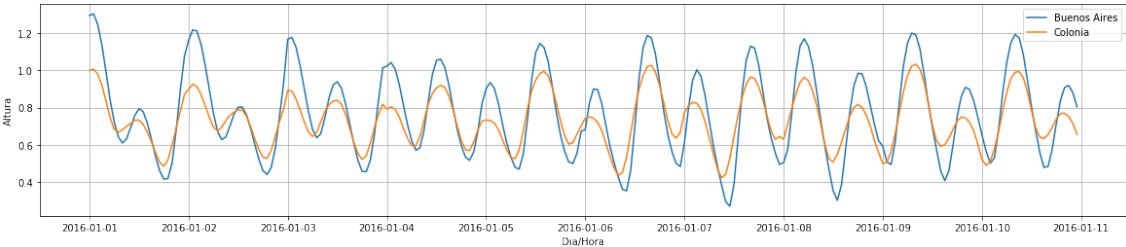
### 3.1.3. Vientos

Los valores del Cuadro 3.2 muestran la frecuencia mensual de la dirección de los vientos en Buenos Aires analizados en el período 2014-2017. Como se puede observar, existe una predominancia de vientos del este desde principios de primavera hasta finales de verano.

El mismo período desde el punto de vista de intensidad y dirección de los vientos, en la Fig.3.11 muestra que los vientos de mayor intensidad ocurren durante los meses de Septiembre a Noviembre y en el verano, predominando los viento de dirección este-sureste, que alcanzan intensidades entre los 30 km/h y 45 km/h, superando los 45 km/h en algunos casos. La escala que se muestra en los ejes de cada gráfico (de 0 a 20) indica el porcentaje



(a) San Clemente - Montevideo



(b) Buenos Aires / Colonia

Figura 3.9: Comparación de la amplitud de mareas en el comienzo (a) y la salida (b) del Río de la Plata - Enero 2016

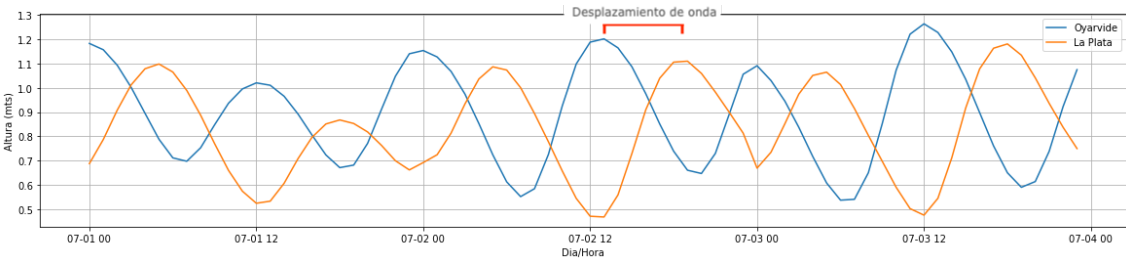


Figura 3.10: Tablas de marea entre Oyarvide y La Plata - Julio 2016

|          | Ene    | Feb    | Mar    | Abr    | May    | Jun    | Jul    | Ago    | Sep    | Oct    | Nov    | Dic    |
|----------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| N        | 24,50% | 18,14% | 17,14% | 14,27% | 19,09% | 21,15% | 20,97% | 18,15% | 13,75% | 13,61% | 18,54% | 23,72% |
| NE       | 10,69% | 9,14%  | 11,26% | 10,42% | 8,00%  | 5,66%  | 7,43%  | 6,92%  | 8,78%  | 8,84%  | 11,91% | 11,83% |
| E        | 27,55% | 33,15% | 32,86% | 26,98% | 25,87% | 14,62% | 22,51% | 20,93% | 31,32% | 33,13% | 27,01% | 26,71% |
| SE       | 7,59%  | 9,29%  | 6,25%  | 10,94% | 9,64%  | 4,20%  | 9,01%  | 9,85%  | 8,51%  | 11,29% | 8,19%  | 6,82%  |
| S        | 13,84% | 13,16% | 15,63% | 12,78% | 13,84% | 12,74% | 15,69% | 14,05% | 19,41% | 17,94% | 12,33% | 13,31% |
| SO       | 4,30%  | 3,76%  | 5,95%  | 6,32%  | 6,15%  | 9,90%  | 6,35%  | 9,31%  | 6,94%  | 5,54%  | 5,97%  | 4,03%  |
| O        | 4,81%  | 8,08%  | 5,44%  | 10,56% | 9,34%  | 18,85% | 9,81%  | 11,46% | 6,94%  | 6,12%  | 7,64%  | 7,33%  |
| NO       | 6,35%  | 4,31%  | 4,67%  | 6,46%  | 6,45%  | 10,42% | 7,53%  | 6,99%  | 3,58%  | 2,92%  | 7,60%  | 5,48%  |
| S/Viento | 0,37%  | 0,96%  | 0,81%  | 1,28%  | 1,61%  | 2,47%  | 0,71%  | 2,35%  | 0,76%  | 0,60%  | 0,80%  | 0,77%  |

Cuadro 3.2: Frecuencia mensual de la dirección de vientos - Buenos Aires, 2014-2017

que cada rango representa del total.

Estas observaciones están en línea con análisis previos que establecen que las sudestadas ocurren con más frecuencia durante los meses de verano que en los de invierno, con dos máximas en frecuencia al principio de la primavera y durante el verano. Como menciona el trabajo de [Escobar G, 2004], la forma del Río de la Plata, de 50km de ancho al inicio y 300km de ancho en su salida al Océano Atlántico y su curso noroeste-sureste, hacen que, al soplar viento fuerte del sureste, el nivel de agua suba.

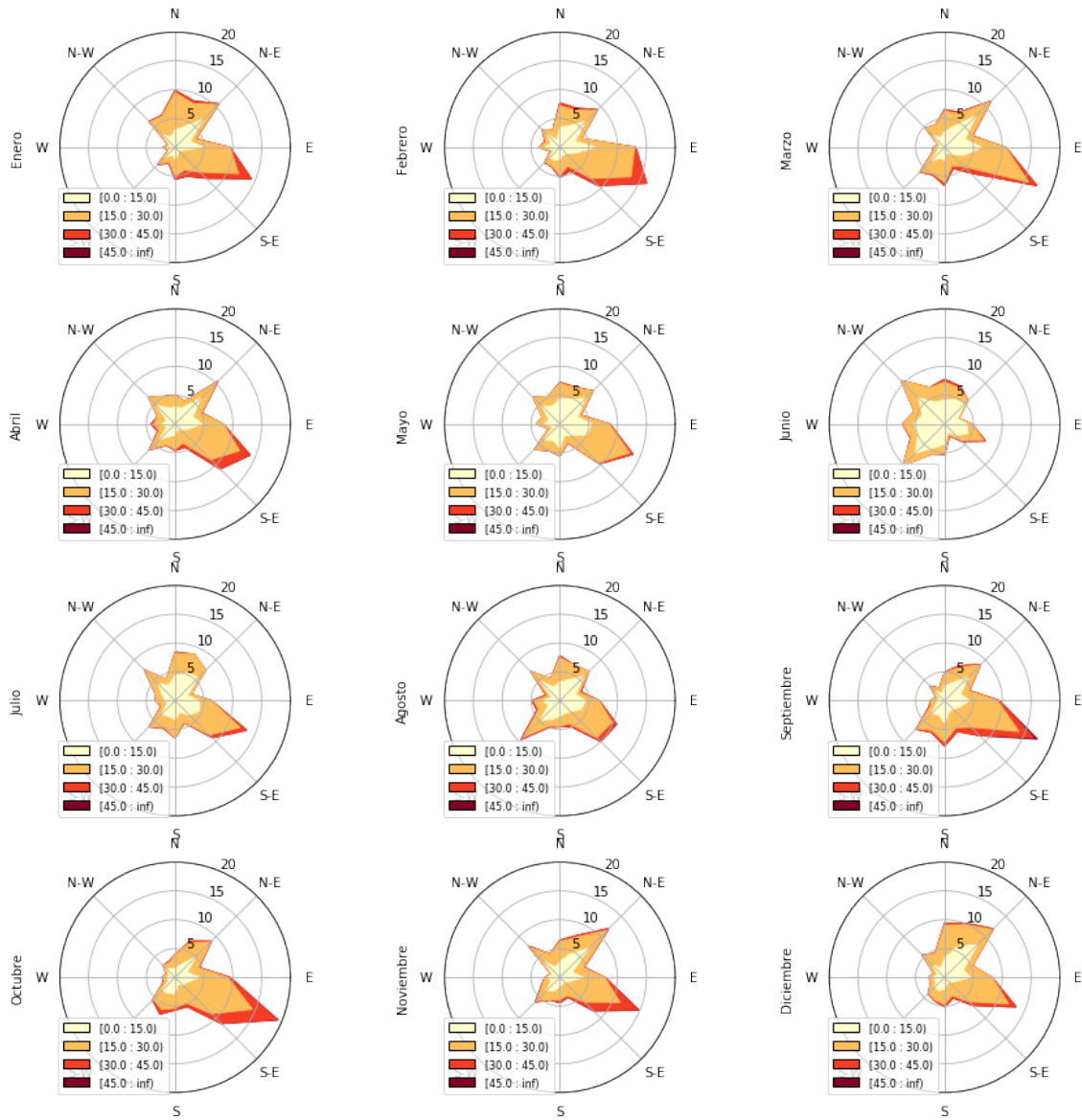


Figura 3.11: Intensidad y dirección mensual de vientos - Buenos Aires, 2014-2017

### Conversión coordenadas a componentes ortogonales

Los registros de viento fueron provistos en términos de velocidad y dirección del viento. La dirección sigue la convención meteorológica, indicando la dirección desde la cual sopla el viento con términos como “viento este”. El valor se incrementa en el sentido de las agujas del reloj desde el punto Norte y está expresado en grados en un intervalo entre  $0^\circ$  y  $360^\circ$ . Así, el viento este equivale entonces a  $90^\circ$ .

Para pasar de la convención velocidad/dirección a coordenadas ortogonales, el uso de funciones trigonométricas asume que los ángulos están expresados en radianes (vs grados), por lo que se convirtió este dato mediante la relación:

$$\phi(rad) = \frac{\pi}{180} \phi(deg)$$

Para obtener las componentes ortogonales de velocidad, donde:

- $u$  es la velocidad zonal, es decir, la componente del viento horizontal hacia el Este
- $v$  es la velocidad meridional, es decir, la componente del viento horizontal hacia el Norte

se aplicaron las relaciones entre componentes  $x$  e  $y$  con el seno y coseno del ángulo definido por el vector respectivamente:

$$u = -|v_H| \sin \left[ \frac{\pi}{180} \phi_{MET}(deg) \right]$$
$$v = -|v_H| \cos \left[ \frac{\pi}{180} \phi_{MET}(deg) \right]$$

#### 3.1.4. Pronósticos de viento y presión media sobre el nivel del mar (m.s.n.m.)

Los puntos geográficos seleccionados en el análisis de los pronósticos de viento y presión media sobre el nivel del mar fueron los siguientes:

- Buenos Aires
- La Plata
- Villa Gesell
- Mar del Plata
- Bahía Blanca
- Colonia
- Montevideo
- Punta del Este

Ambos pronósticos se desplazaron -3 horas, haciéndolos corresponder así con la zona horaria de registro de los datos restantes, es decir, UTC-03.

Durante el análisis inicial y posterior modelado, se utilizaron los pronósticos de viento como reemplazo de las observaciones reales. Esta decisión estuvo guiada por dos motivos: por un lado, existían períodos extensos de valores faltantes en las observaciones de vientos que no podían ser generados sin introducir sesgos en el análisis; por otro, la necesidad de utilizar pronósticos como variables adicionales de los modelos para una mejor predicción de las alturas horarias.

Para verificar que el uso de pronósticos de viento no introdujera sesgos en el análisis y modelado al reemplazar a las observaciones, se analizaron las correlaciones entre ambos conjuntos de variables para los puntos geográficos de Buenos Aires y Mar del Plata tomados como referencia.

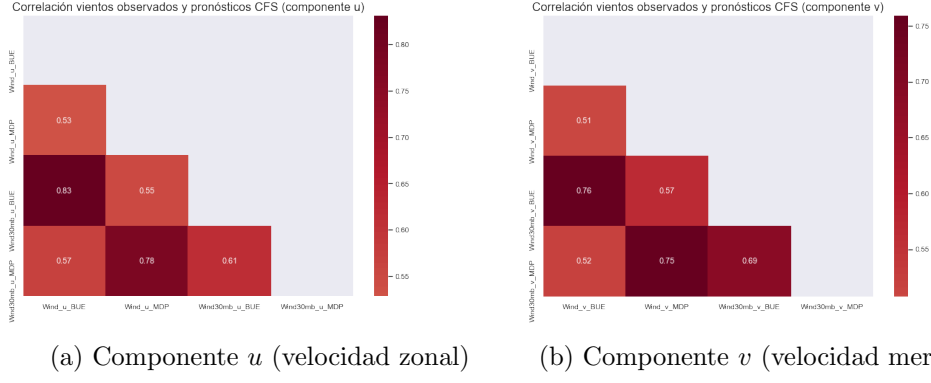


Figura 3.12: Correlación de vientos y pronósticos CFS Bs As y Mar del Plata, 2014-2017

Como se observa en el gráfico 3.12, las variables de observaciones y sus correspondientes de pronósticos de vientos, tanto en Buenos Aires como en Mar del Plata, muestran correlación positiva significativa entre sus componentes ( $\geq 0,75$ ) que dio sustento a la sustitución de las primeras.

### 3.2. Creación de variables

A partir de las observaciones y pronósticos recolectados, se generó una serie de nuevas variables, considerando medias móviles en distintas ventanas de tiempo y desplazamientos de  $n$  horas hacia atrás / adelante (Fig.3.13).

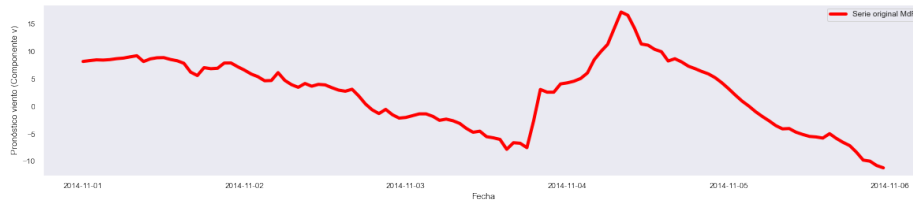
Al contar con pronósticos como el caso de mareas, vientos y presión m.s.n.m, además de generar variables para tiempos anteriores a  $x_t$ , fue posible generar variables en tiempo futuro en relación al mismo.

Por otro lado, se exploraron las correlaciones entre los residuos resultantes de cada modelo y los pronósticos de vientos de cada punto geográfico recorriendo direcciones entre  $0^\circ$  y  $360^\circ$  en incrementos de  $10^\circ$  y desplazamientos que variaron entre -5hs y +6hs con respecto al registro de hora del pronóstico. De esta forma, se generaron nuevas variables a partir de la mejor correlación obtenida.

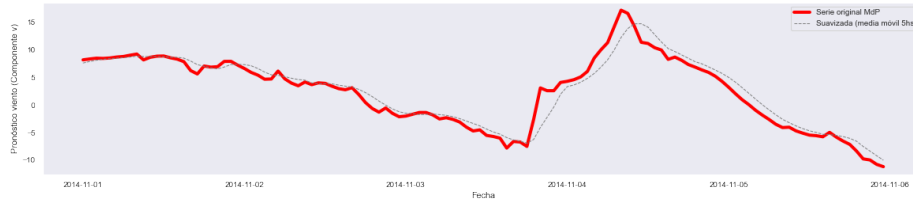
### 3.3. Propiedades de la serie de tiempo de alturas del Río de la Plata

Como menciona [Box and Jenkins, 1994], observaciones adyacentes en una serie de tiempo son típicamente dependientes. A partir del análisis exploratorio, se comprobaron las siguientes propiedades de la serie correspondiente a las alturas del Río de la Plata:

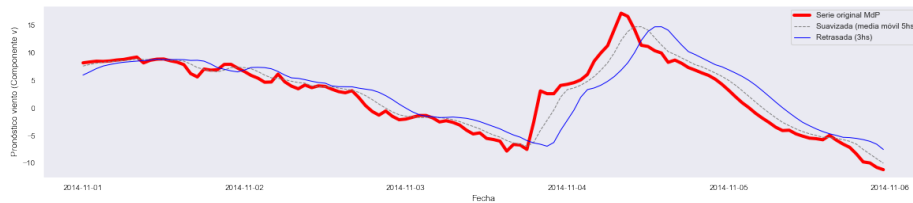
- de manera visual, se observó que la serie de tiempo de alturas del Río de la Plata no muestra **tendencia**, las observaciones en promedio no muestran crecimiento o decrecimiento a lo largo del tiempo que pudiera requerir una diferenciación de la serie
- por otro lado, la serie exhibe un comportamiento en equilibrio, donde su media y



(a) Serie de tiempo original



(b) Serie suavizada (media móvil 5hs)



(c) Serie desplazada (adelantada 3hs)

Figura 3.13: Generación variable sobre pronóstico CFS - Mar del Plata, 2014-2017

varianza se mantienen constantes. Dicha **estacionaridad** de la serie se comprobó mediante el test de Dickey-Fuller

- el gráfico de autocorrelación identificó la presencia de un patrón repetitivo regular, que exhibió una **estacionalidad** de subas / bajas con una frecuencia de 24 hs
- por último, como se mostrará en la siguiente sección (Fig.4.6), el gráfico de valores ajustados de la regresión vs residuos mostró una varianza constante a lo largo del tiempo, por lo que inicialmente se asumió **homocedasticidad** de la serie

## Capítulo 4

# Modelado

En esta sección se analizaron las distintas técnicas de modelado y aprendizaje automático para la predicción de la altura del nivel de agua en el Río de la Plata.

Partiendo de un modelo de regresión lineal básico de mínimos cuadrados (*OLS* por sus siglas en inglés), el objetivo fue trazar una línea base a partir de la cual medir la performance de modelos más específicos aplicables a series de tiempo, como ARIMA, SARIMAX y redes neuronales.

Esto permitió realizar una comparación relativa de las distintas clases de modelo en términos de precisión.

### 4.1. Regresión lineal

Desde el punto de vista de implementación, la regresión lineal aplicada implementa el método de mínimos cuadrados básico a fin de minimizar el error residual de la suma de cuadrados entre las observaciones y las predicciones de la aproximación lineal. Se aplicó a la serie una normalización, restando la media y dividiendo por la norma Euclídea o *l2*.

Con el fin de establecer la línea base más óptima dentro de las regresiones lineales, se generaron 4 modelos combinando distintos predictores (Cuadro 4.1):

| # | Grupos                                                                    | Puntos geográficos                       | # Vars | Variables                                                |
|---|---------------------------------------------------------------------------|------------------------------------------|--------|----------------------------------------------------------|
| 1 | Alturas horarias (Pa)                                                     | Buenos Aires                             | 7      | Obs Alt (-6hs a -1h)                                     |
| 2 | Tabla de mareas (Fut)                                                     | Todos                                    | 70     | Pron (+1h a +6hs)                                        |
| 3 | Alturas horarias (Pa) + Tabla mareas (Fut)                                | Buenos Aires (Alt) +<br>Todos (TM)       | 77     | Obs alt (-6hs a -1h) + Pron<br>TM (+1h a +6hs)           |
| 4 | Tabla mareas (Fut) + Alturas horarias (Pa) + Presión/Vientos CFS (Pa+Fut) | Buenos Aires (Alt) +<br>Todos (TM + CFS) | 413    | Obs (Alt -6hs a -1h) + Pron<br>(-6hs a -1h y +1h a +6hs) |

Cuadro 4.1: Detalle de conjuntos de datos generados

Los modelos generados se entrenaron sobre los nuevos conjuntos de datos y se predijeron las alturas horarias correspondientes a la ciudad de Buenos Aires, obteniéndose el siguiente gráfico de curvas comparativo con pronósticos a 6hs (Fig. 4.1).

Como se ve en el gráfico y a partir del análisis de las métricas del Cuadro 4.2, se pudo ver que el modelo #4 (línea amarilla del gráfico) predice a 6hs con la mayor precisión el



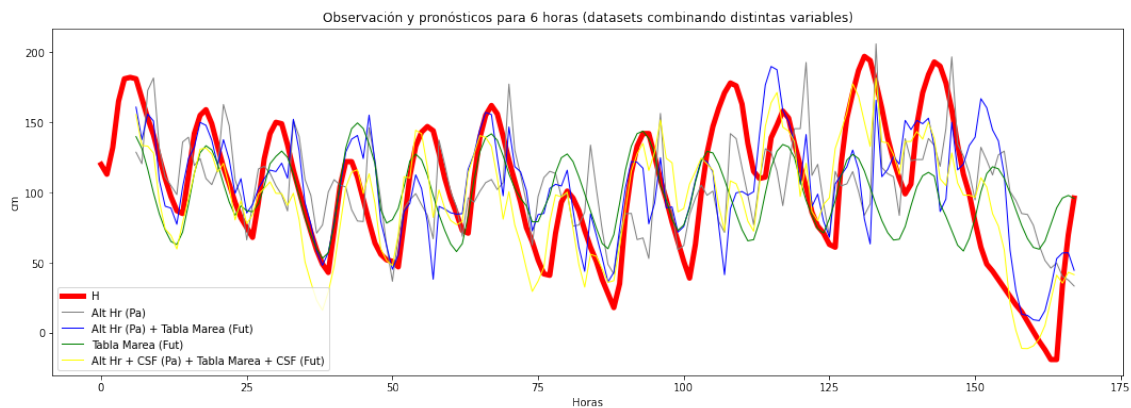


Figura 4.1: Observación y predicciones a 6h - Buenos Aires, 1<sup>a</sup> semana Enero 2018

nivel del Río de la Plata, con el menor RMSE de aproximadamente 30cm y un coeficiente de determinación mayor que los restantes modelos ( $=0.709$ ). Si bien el total de variables incluidas en el conjunto de datos ( $=413$ ) podría haber incidido en el criterio de selección de este modelo, el AIC del mismo fue además el menor.

| # | Regresión lineal                                                          | RMSE  | $R^2$ | AIC        |
|---|---------------------------------------------------------------------------|-------|-------|------------|
| 1 | Alturas Horarias (Pa)                                                     | 36.67 | 0.506 | 624289.831 |
| 2 | Tabla Mareas (Fut)                                                        | 48.95 | 0.120 | 663165.504 |
| 3 | Alturas Horarias (Pa) + Tabla Mareas (Fut)                                | 30.62 | 0.656 | 607470.064 |
| 4 | Tabla Mareas (Fut) + Alturas Horarias (Pa) + Presión/Vientos CFS (Pa+Fut) | 30.00 | 0.709 | 495350.354 |

Cuadro 4.2: RMSE, coeficiente de determinación y AIC de regresiones lineales (predicciones a 6h)

Se observó además que el modelo que consideró únicamente pronósticos de marea resultó con el mayor error y menor coeficiente  $R^2$ , lo que evidenció un menor aporte de información al momento de explicar la varianza de la serie. En otras palabras, el fenómeno astronómico relacionado a las mareas no ofrecería en la regresión lineal un aporte significativo a la hora de predecir alturas del nivel de agua.

Tomando como base el modelo con mejor resultado en términos de AIC, se corrió una nueva predicción de las alturas del agua con pronósticos a 1h, 3hs, 6hs y 12hs para determinar la capacidad de predicción comparativamente a períodos menores y mayores, obteniéndose los resultados que se muestran en la Fig. 4.2 y el Cuadro 4.3).

| Predicción (hs) | RMSE | $R^2$ |
|-----------------|------|-------|
| +1              | 7.4  | 0.982 |
| +3              | 20.5 | 0.864 |
| +6              | 30.0 | 0.709 |
| +12             | 28.7 | 0.734 |

Cuadro 4.3: RMSE y coeficiente de determinación (1h, 3hs, 6hs, 12hs)

## 4.2. Modelos lineales sobre series de tiempo

Una clase importante de procesos estocásticos es la de los procesos estacionarios que en particular varían alrededor de una media fija. Una serie de tiempo se puede descomponer de acuerdo a:

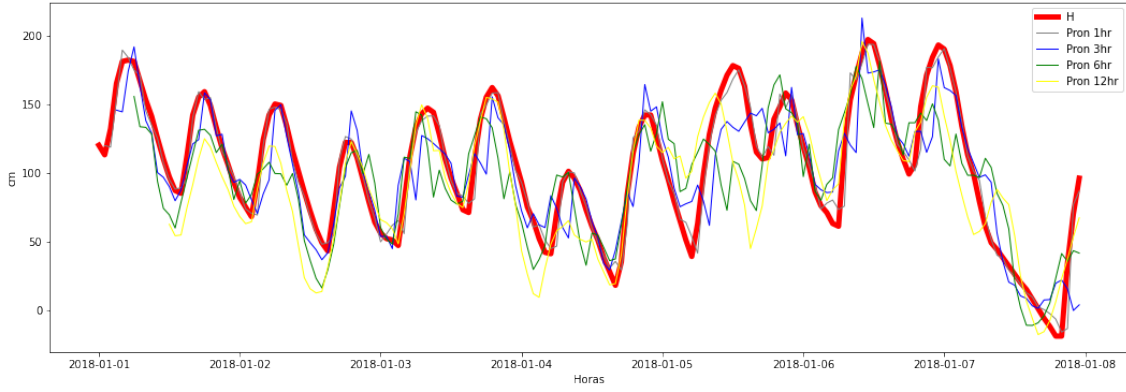


Figura 4.2: Observación y pronósticos a 1h, 3hs, 6hs y 12hs - Buenos Aires, 1<sup>a</sup> semana Enero 2018

- Tendencia: aumento / disminución de las observaciones a lo largo de períodos extendidos de tiempo
- Estacionalidad: variaciones específicas de una ventana de tiempo particular
- Ruido: picos y depresiones en intervalos aleatorios

Como se menciona en la sección *SARIMA*, durante el proceso iterativo, la identificación del modelo y la estimación de parámetros se superponen en el sentido de que la estimación condujo a la simplificación, siendo la identificación un método inexacto donde no hay una formulación precisa del problema. En esta fase se utilizaron métodos gráficos para considerar tentativamente una clase de modelos a ser luego entrenados.

#### 4.2.1. Formulación de clase general del modelo

##### Funciones de autocorrelación (ACF) y autocorrelación parcial (PACF)

El cálculo de las estimaciones iniciales de un proceso  $ARMA(p, q)$  está basado en las primeras  $p + q + 1$  autocovarianzas  $c_j[j = 0, 1, \dots, (p + q)]$  de  $w_t = \Delta^d z_t$ . Los pasos del cálculo son los siguientes:

1. Los parámetros  $\phi_1, \phi_2, \dots, \phi_p$  se calculan a partir de las autocovarianzas  $c_{q-p+1}, \dots, c_{q+1}, c_{q+2}, \dots, c_{q+p}$
2. Con las estimaciones  $\hat{\phi}$  obtenidas, se calculan las primeras  $q + 1$  autocovarianzas  $c'_j(j = 0, 1, \dots, q)$  de la serie derivada

$$w'_t = w_t - \hat{\phi}w_{t-1} - \dots - \hat{\phi}_p w_{t-p}$$

3. Las autocovarianzas  $c'_0, c'_1, \dots, c'_q$  se utilizan iterativamente para calcular las estimaciones iniciales de los parámetros de media móvil  $\theta_1, \theta_2, \dots, \theta_q$  y de la varianza residual  $\sigma_a^2$

Para un proceso  $ARMA(p, q)$  existirán  $q$  autocorrelaciones  $\rho_q, \rho_{q-1}, \dots, \rho_1$  cuyos valores dependen directamente de la elección de los  $q$  parámetros  $\theta$  de la media móvil, así como también de los  $p$  parámetros autorregresivos  $\phi$ .

Se estudió la forma de las funciones de autocorrelación (ACF) y autocorrelación parcial (PACF) estimadas para obtener una primera estimación acerca de la elección del orden de  $p$  y  $q$  para los operadores autorregresivos y de media móvil:

- **Función de Autocorrelación (ACF):** muestra la medida de correlación entre la serie y valores pasados de sí misma:

$$ACF(x_t, x_{t-h}) = \frac{Cov(x_t, x_{t-h})}{Var(x_t)}$$

- **Función de Autocorrelación Parcial (PACF):** muestra la medida de la correlación entre la serie con valores pasados de sí misma luego de eliminar las variaciones explicadas por las comparaciones intervinientes. Una correlación parcial de 3 orden puede escribirse como:

$$PACF(x_t, x_{t-3}) = \frac{Cov(x_t, x_{t-3} | x_{t-1}, x_{t-2})}{\sqrt{Var(x_t | x_{t-1}, x_{t-2}) Var(x_{t-3} | x_{t-1}, x_{t-2})}}$$

Repasando el comportamiento característico de la teoría de estas funciones en procesos combinados ARMA, mientras que la función de autocorrelación de un proceso autorregresivo de orden  $p$  se afina, su función de autocorrelación parcial tiene un corte a partir de la observación pasada  $q$ . Inversamente, la función de autocorrelación de un proceso de media móvil de orden  $q$  muestra un corte mas allá de la observación pasada  $q$ , mientras que su autocorrelación parcial se afina. Si ambas funciones se afinan, esto indica un proceso combinado conteniendo un componente autorregresivo de orden  $p$  y un componente de media móvil de orden  $q$ .

Para un modelo AR, las autocorrelaciones parciales son iguales a 0 pasando el orden del modelo, o lo que es igual, el número de autocorrelaciones parciales  $q$  da el orden del modelo AR. Para un modelo MA, un patrón más claro lo muestran las autocorrelaciones totales. ACF tendrá autocorrelaciones distintas de 0 para los valores retrasados intervinientes en el modelo MA.

Analizando los gráficos correspondientes a la serie de estudio (Fig. 4.3) se observa que tanto la autocorrelación como la autocorrelación parcial se afinan en los primeros 40 valores pasados, indicando que trataría de un proceso combinado (ARMA).

Con respecto a la tendencia de la serie, y en base a lo analizado en la sección 3.1 (Figs. 3.3 y 3.4) no se encontró tendencia creciente o decreciente marcada en el período. Por otro lado, el gráfico de la Fig. 4.3 mostró un patrón regular que se repetía cerca de las 24 horas, lo que indicó una cantidad de períodos  $S = 24$  (estacionalidad diaria).

A fin de analizar este efecto estacional en más detalle, se buscaron patrones significativos cercanos a frecuencias de 24hs. Se diferenciò la serie con respecto a  $\nabla_{24}$ , calculando  $x_t - x_{t-24}$  y analizando nuevamente los gráficos ACF y PACF para la nueva serie generada (Fig. 4.4).

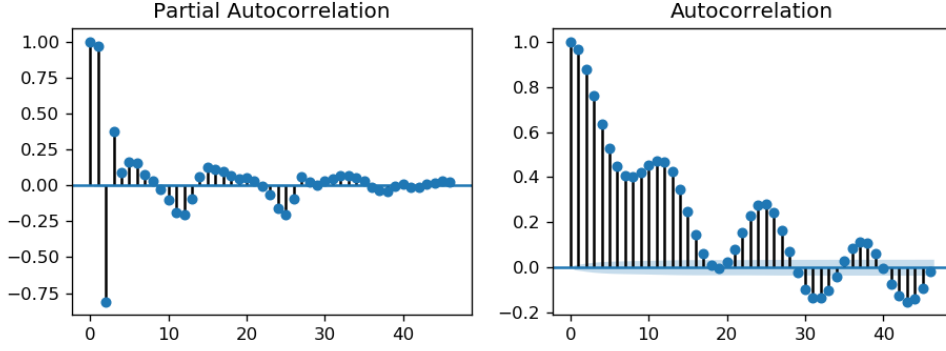


Figura 4.3: PACF y ACF de nivel de alturas Buenos Aires (2014-2017)

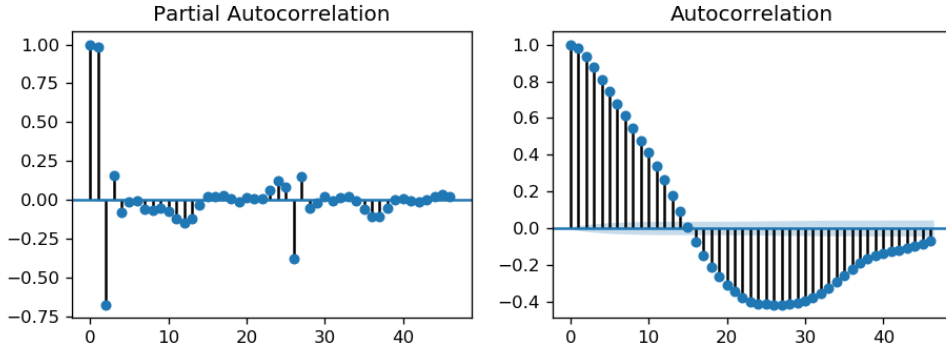


Figura 4.4: PACF y ACF de la serie diferenciada a 24hs

- **términos no estacionales.** Los retrasos significativos cercanos al inicio (y menores a  $S = 24$ ), tanto en el ACF y PACF, dan una aproximación a los términos no estacionales que pueden contribuir al modelo. En el PACF, los tres primeros saltos (positivo y negativo) indicaron un posible modelo AR(3); en los valores siguientes los valores se amortiguan. En el ACF, los valores significativos hasta el 14vo indicaron un posible modelo MA(14).
- **términos estacionales.** Para los términos estacionales, se observó si existían patrones cercanos a valores múltiplos de la diferencia tomada ( $\nabla_{24}$ ). En el PACF se observan valores significativos cercanos a 24, por lo que se consideró un posible modelo AR(1). El ACF mostró un grupo de valores negativos significativos cercanos a 24, luego del cual no se observa otro patrón significativo, indicando un posible modelo MA(1).

Sobre esta base se identificó un modelo SARIMA(3,0,14)x(1,1,1)<sub>24</sub> como modelo inicial. Los modelos SARIMA de los siguientes análisis ajustan por máxima verosimilitud utilizando el filtro Kalman. El filtro Kalman es un algoritmo recursivo de estimación de estados en sistemas dinámicos lineales con ruido, que trabaja sobre la base de predicción - corrección.

Asumiendo que el estado del sistema evoluciona de acuerdo a la ecuación lineal

$$x_k = Ax_{k-1} + w_{k-1}$$

donde A es una matriz que relaciona con el estado anterior  $x_{k-1}$  y  $w_{k-1} \in \mathbb{R}^n$  modela el

ruido del sistema, el filtro Kalman asume que las mediciones pueden modelarse en relación a los estados mediante la ecuación

$$z_k = Hx_k + v_k$$

donde la medición  $z_k \in \mathbb{R}^m$  en el momento  $k$  depende linealmente del estado del sistema  $x_k$ , la matriz  $H$  relaciona el estado actual  $x_k$  a la medición  $z_k$  y el vector  $v_k \in \mathbb{R}^m$  modela el ruido. El filtro utiliza este modelo a fin de corregir la predicción del estado cuando existe una medición. Así, dado un estado real del sistema, el modelo puede describir cuál será la medición correspondiente, para poder comparar con la medición real y corregir la predicción del estado.

### Búsqueda de hiperparámetros (*Grid search*)

Con el objetivo de realizar un primer análisis y aproximar la configuración de hiperparámetros a los valores que mejor se ajustaban a la serie de tiempo, se ejecutó una búsqueda sobre un subconjunto acotado de valores iniciales (o *grid search* por su nombre en inglés).

En base a los valores iniciales obtenidos en la sección anterior, la búsqueda se acotó en los componentes AR MA e I a valores de orden bajo, (entre 0 y 3), a excepción del término no estacional MA que se estimó en 14. Un total de 4 modelos que generaron alertas por falta de convergencia, no fue incluido en los resultados.

Se compararon 56 configuraciones de modelos, analizando el criterio AIC, además de RMSE como se muestra en el Cuadro 4.4.

El modelo con mejores resultados iniciales en términos de AIC fue el SARIMA (3,0,12)x(2,1,1)<sub>24</sub> (#51), donde todos los coeficientes excepto uno mostraron un nivel de significancia  $p < 0,05$ , el AIC fue el menor registrado, y el RMSE fue de 7.7 cm. Los siguientes modelos en términos de criterio de información y error estimado fueron el SARIMA (3,0,13)x(1,1,1)<sub>24</sub> (#53) y el SARIMA (3,0,8)x(2,1,1)<sub>24</sub> (#35).

Analizada más en detalle, la salida del modelo (3,0,12)x(2,1,1)<sub>24</sub> (#51) (Fig. 4.5), muestra que los parámetros estimados en este primer modelo, tanto estacionales como no estacionales, son significativos en un nivel de confianza  $p < 0,05$ , con excepción de los términos no estacionales MA(8) y MA(10), cuyos coeficientes fueron relativamente despreciables en cuanto a su peso en el modelo.

Por el lado de la varianza de los residuos, de forma visual se observó que se cumplía la expectativa de que los mismos mantuvieran una varianza constante, mostrando una recta de ajuste casi horizontal (Fig. 4.6).

A partir de la formulación de este modelo inicial, se analizó iterativamente la contribución de distintas variables exógenas a la explicación de los residuos del modelo y en consecuencia, de la dinámica de la serie de tiempo.

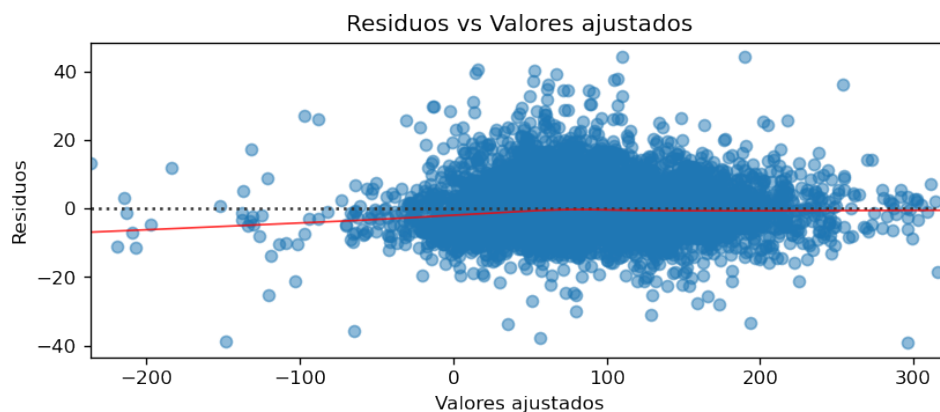
| #         | p        | d        | q         | P        | D        | Q        | S         | CI y errores     |            |
|-----------|----------|----------|-----------|----------|----------|----------|-----------|------------------|------------|
|           |          |          |           |          |          |          |           | AIC              | RMSE       |
| 1         | 2        | 0        | 7         | 1        | 1        | 1        | 24        | 59838.862        | 8.4        |
| 2         | 2        | 0        | 7         | 1        | 1        | 2        | 24        | 59839.594        | 8.4        |
| 3         | 2        | 0        | 7         | 2        | 1        | 1        | 24        | 59838.687        | 8.4        |
| 4         | 2        | 0        | 7         | 2        | 1        | 2        | 24        | 59842.363        | 8.4        |
| 5         | 2        | 0        | 8         | 1        | 1        | 1        | 24        | 59667.101        | 8.3        |
| 6         | 2        | 0        | 8         | 1        | 1        | 2        | 24        | 59668.928        | 8.3        |
| 7         | 2        | 0        | 8         | 2        | 1        | 1        | 24        | 59668.797        | 8.3        |
| 8         | 2        | 0        | 8         | 2        | 1        | 2        | 24        | 59670.285        | 8.3        |
| 9         | 2        | 0        | 9         | 1        | 1        | 1        | 24        | 59564.730        | 8.3        |
| 10        | 2        | 0        | 9         | 1        | 1        | 2        | 24        | 59565.872        | 8.3        |
| 11        | 2        | 0        | 9         | 2        | 1        | 1        | 24        | 59565.386        | 8.3        |
| 12        | 2        | 0        | 9         | 2        | 1        | 2        | 24        | 59567.972        | 8.3        |
| 13        | 2        | 0        | 10        | 1        | 1        | 1        | 24        | 59515.115        | 8.2        |
| 14        | 2        | 0        | 10        | 1        | 1        | 2        | 24        | 59514.903        | 8.2        |
| 15        | 2        | 0        | 10        | 2        | 1        | 1        | 24        | 59513.697        | 8.2        |
| 16        | 2        | 0        | 10        | 2        | 1        | 2        | 24        | 59518.204        | 8.2        |
| 17        | 2        | 0        | 11        | 1        | 1        | 1        | 24        | 59514.088        | 8.2        |
| 18        | 2        | 0        | 11        | 1        | 1        | 2        | 24        | 59513.918        | 8.2        |
| 19        | 2        | 0        | 11        | 2        | 1        | 1        | 24        | 59522.480        | 8.2        |
| 20        | 2        | 0        | 11        | 2        | 1        | 2        | 24        | 59517.093        | 8.2        |
| 21        | 2        | 0        | 12        | 1        | 1        | 1        | 24        | 59322.478        | 8.2        |
| 23        | 2        | 0        | 12        | 2        | 1        | 1        | 24        | 59307.053        | 8.2        |
| 24        | 2        | 0        | 12        | 2        | 1        | 2        | 24        | 59326.225        | 8.2        |
| 25        | 2        | 0        | 13        | 1        | 1        | 1        | 24        | 59245.817        | 8.1        |
| 26        | 2        | 0        | 13        | 1        | 1        | 2        | 24        | 59249.329        | 8.1        |
| 27        | 2        | 0        | 13        | 2        | 1        | 1        | 24        | 58531.951        | 7.9        |
| 28        | 2        | 0        | 13        | 2        | 1        | 2        | 24        | 59250.159        | 8.1        |
| 29        | 3        | 0        | 7         | 1        | 1        | 1        | 24        | 58149.814        | 7.8        |
| 30        | 3        | 0        | 7         | 1        | 1        | 2        | 24        | 58154.966        | 7.8        |
| 31        | 3        | 0        | 7         | 2        | 1        | 1        | 24        | 58147.214        | 7.8        |
| 32        | 3        | 0        | 7         | 2        | 1        | 2        | 24        | 58153.658        | 7.8        |
| 33        | 3        | 0        | 8         | 1        | 1        | 1        | 24        | 58145.741        | 7.7        |
| 34        | 3        | 0        | 8         | 1        | 1        | 2        | 24        | 58206.511        | 7.8        |
| <b>35</b> | <b>3</b> | <b>0</b> | <b>8</b>  | <b>2</b> | <b>1</b> | <b>1</b> | <b>24</b> | <b>58143.231</b> | <b>7.7</b> |
| 36        | 3        | 0        | 8         | 2        | 1        | 2        | 24        | 58172.986        | 7.8        |
| 37        | 3        | 0        | 9         | 1        | 1        | 1        | 24        | 59396.332        | 8.2        |
| 39        | 3        | 0        | 9         | 2        | 1        | 1        | 24        | 59347.194        | 8.2        |
| 40        | 3        | 0        | 9         | 2        | 1        | 2        | 24        | 59399.313        | 8.2        |
| 41        | 3        | 0        | 10        | 1        | 1        | 1        | 24        | 59373.314        | 8.2        |
| 43        | 3        | 0        | 10        | 2        | 1        | 1        | 24        | 59363.544        | 8.2        |
| 44        | 3        | 0        | 10        | 2        | 1        | 2        | 24        | 59377.139        | 8.2        |
| 45        | 3        | 0        | 11        | 1        | 1        | 1        | 24        | 59362.574        | 8.2        |
| 46        | 3        | 0        | 11        | 1        | 1        | 2        | 24        | 59283.017        | 8.1        |
| 47        | 3        | 0        | 11        | 2        | 1        | 1        | 24        | 59351.573        | 8.2        |
| 48        | 3        | 0        | 11        | 2        | 1        | 2        | 24        | 59365.703        | 8.2        |
| 49        | 3        | 0        | 12        | 1        | 1        | 1        | 24        | 59312.728        | 8.2        |
| 50        | 3        | 0        | 12        | 1        | 1        | 2        | 24        | 59386.749        | 8.2        |
| <b>51</b> | <b>3</b> | <b>0</b> | <b>12</b> | <b>2</b> | <b>1</b> | <b>1</b> | <b>24</b> | <b>57908.764</b> | <b>7.7</b> |
| 52        | 3        | 0        | 12        | 2        | 1        | 2        | 24        | 59316.228        | 8.2        |
| <b>53</b> | <b>3</b> | <b>0</b> | <b>13</b> | <b>1</b> | <b>1</b> | <b>1</b> | <b>24</b> | <b>57928.468</b> | <b>7.7</b> |
| 55        | 3        | 0        | 13        | 2        | 1        | 1        | 24        | 59474.855        | 8.2        |
| 56        | 3        | 0        | 13        | 2        | 1        | 2        | 24        | 59251.831        | 8.1        |

Cuadro 4.4: Búsqueda hiperparámetros y comparación de modelos SARIMA

### Limitación de modelos SARIMAX con incorporacion de exógenas

Es importante notar que dado que los pronósticos de esta clase de modelos se realizan utilizando variables exógenas disponibles en las horas a predecir, el horizonte pronosticado se limita a los datos existentes para cada momento. Dado que el conjunto de datos utilizado incluyó pronósticos de vientos y presión de CFS, que realizan reanálisis 4 veces y 5 pronósticos en las horas sucesivas, el límite de horas a predecir es 6.

| Statespace Model Results |                                 |                   |            |       |        |        |
|--------------------------|---------------------------------|-------------------|------------|-------|--------|--------|
| Dep. Variable:           | Level_BUE_winsorized            | No. Observations: | 8760       |       |        |        |
| Model:                   | SARIMAX(3, 0, 12)x(2, 1, 1, 24) | Log Likelihood    | -28935.382 |       |        |        |
| Date:                    | Sat, 20 Mar 2021                | AIC               | 57908.764  |       |        |        |
| Time:                    | 22:01:21                        | BIC               | 58043.193  |       |        |        |
| Sample:                  | 01-01-2014                      | HQIC              | 57954.575  |       |        |        |
|                          | - 12-31-2014                    |                   |            |       |        |        |
| Covariance Type:         | opg                             |                   |            |       |        |        |
|                          | coef                            | std err           | z          | P> z  | [0.025 | 0.975] |
| ar.L1                    | 2.6417                          | 0.008             | 347.930    | 0.000 | 2.627  | 2.657  |
| ar.L2                    | -2.5591                         | 0.013             | -192.153   | 0.000 | -2.585 | -2.533 |
| ar.L3                    | 0.8908                          | 0.008             | 117.778    | 0.000 | 0.876  | 0.906  |
| ma.L1                    | -0.8081                         | 0.011             | -76.264    | 0.000 | -0.829 | -0.787 |
| ma.L2                    | -0.1166                         | 0.011             | -10.901    | 0.000 | -0.138 | -0.096 |
| ma.L3                    | 0.2882                          | 0.013             | 22.631     | 0.000 | 0.263  | 0.313  |
| ma.L4                    | 0.2537                          | 0.013             | 18.917     | 0.000 | 0.227  | 0.280  |
| ma.L5                    | 0.0397                          | 0.015             | 2.735      | 0.006 | 0.011  | 0.068  |
| ma.L6                    | 0.0986                          | 0.014             | 6.816      | 0.000 | 0.070  | 0.127  |
| ma.L7                    | 0.0778                          | 0.015             | 5.097      | 0.000 | 0.048  | 0.108  |
| ma.L8                    | 0.0139                          | 0.014             | 0.971      | 0.331 | -0.014 | 0.042  |
| ma.L9                    | -0.0309                         | 0.014             | -2.258     | 0.024 | -0.058 | -0.004 |
| ma.L10                   | 0.0226                          | 0.013             | 1.798      | 0.072 | -0.002 | 0.047  |
| ma.L11                   | 0.0882                          | 0.012             | 7.608      | 0.000 | 0.065  | 0.111  |
| ma.L12                   | 0.0397                          | 0.010             | 3.940      | 0.000 | 0.020  | 0.059  |
| ar.S.L24                 | 0.0305                          | 0.008             | 3.608      | 0.000 | 0.014  | 0.047  |
| ar.S.L48                 | -0.0378                         | 0.010             | -3.811     | 0.000 | -0.057 | -0.018 |
| ma.S.L24                 | -0.9861                         | 0.003             | -380.396   | 0.000 | -0.991 | -0.981 |
| sigma2                   | 43.6351                         | 0.416             | 104.963    | 0.000 | 42.820 | 44.450 |
| Ljung-Box (Q):           | 745.23                          | Jarque-Bera (JB): | 7249.53    |       |        |        |
| Prob(Q):                 | 0.00                            | Prob(JB):         | 0.00       |       |        |        |
| Heteroskedasticity (H):  | 0.95                            | Skew:             | 0.84       |       |        |        |
| Prob(H) (two-sided):     | 0.21                            | Kurtosis:         | 7.13       |       |        |        |

Figura 4.5: Salida SARIMA (3,0,12)x(2,1,1)<sub>24</sub>Figura 4.6: SARIMA (3,0,12)x(2,1,1)<sub>24</sub> - Residuos vs valores ajustados

#### 4.2.2. Iteración 1: SARIMAX (3,0,12)x(2,1,1)<sub>24</sub> y una variable exógena

##### Identificación de modelo tentativo

En una primera iteración se analizaron las correlaciones con variables exógenas con el objeto de explicar las variaciones de alturas como un proceso autorregresivo donde otras variables (observaciones o predicciones) se asumen como regresoras.

En base al modelo anterior, se analizó la correlación del error residual de la variable principal (*Residual*) con distintas variables, ajustando mediante el coeficiente de Pearson que se independiza de la escala de medida de las variables.

Entre las variables generadas de mayor correlación con el error residual, las generadas a

partir de los pronósticos de vientos fueron las que mostraron una correlación más fuerte. En este grupo, se seleccionó la velocidad meridional (componente  $v$ ) del pronóstico de vientos de Montevideo, con una media móvil de 3 horas y desplazamiento de 1 hora (es decir, promedio de 3 horas pasadas de vientos pronosticados a un momento dado y desplazado 1 hora hacia adelante al momento de cada observación), correspondiente a la variable *Wind30mb\_v\_MVD\_3\_1*, como muestra la Fig 4.7). La correlación con esta variable fue de 0.13, y si bien se obtuvieron correlaciones más fuertes a esta en términos absolutos, los modelos resultantes no fueron mejores en relación a criterios de información ni errores calculados.

Así, la velocidad meridional de los pronósticos de viento de Montevideo estimados una hora antes explican parte del error residual no captada por el modelo base.

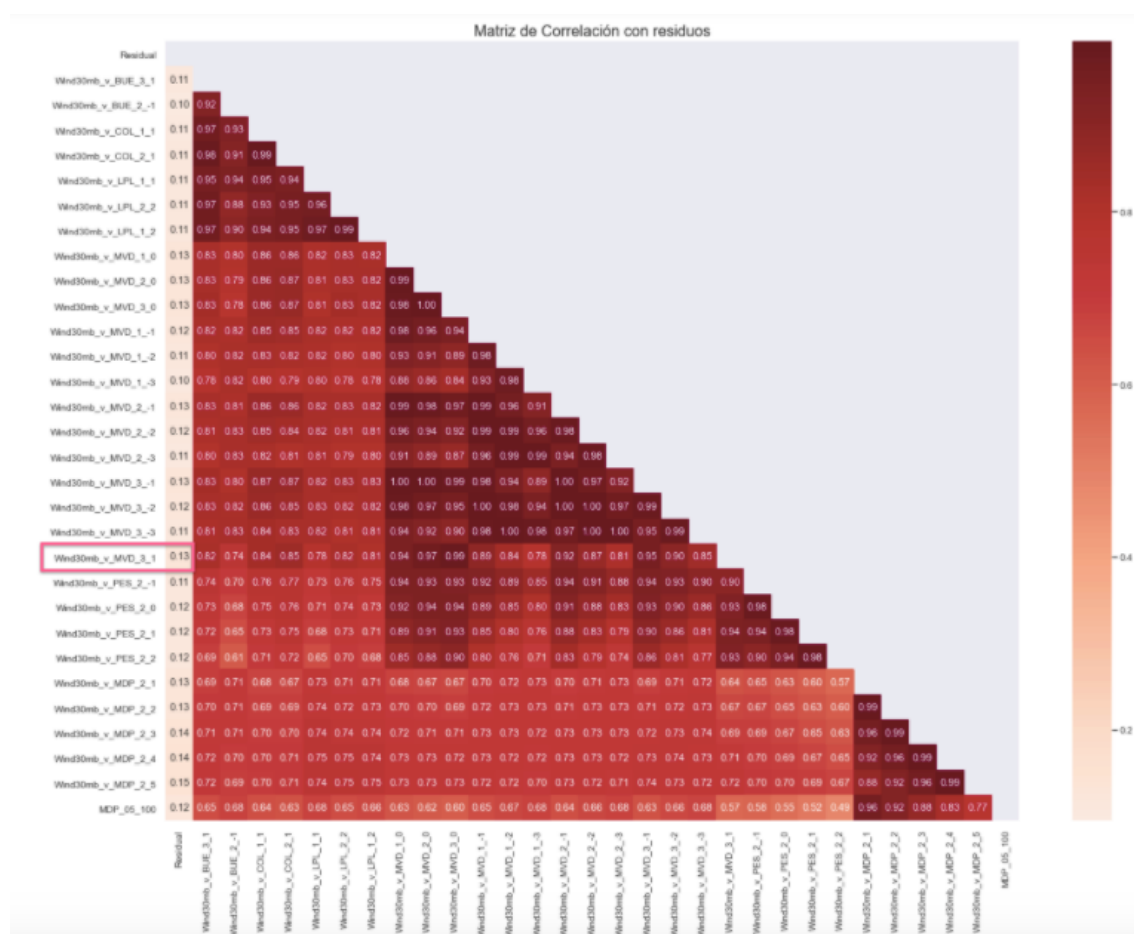


Figura 4.7: Correlación de residuos SARIMA(3,0,12)x(2,1,1)<sub>24</sub> vs. pronósticos de vientos

## Estimación de parámetros

La salida de este nuevo modelo se muestra en la Fig. 4.8.

Si bien en relación al modelo base el AIC obtenido es mayor (59274.441 respecto al anterior 57908.764), se realizó primeramente un análisis comparativo de los valores de AIC entre los modelos que incorporan variables exógenas para finalmente realizar una evaluación con el modelo base en un paso final (Cuadro 4.5 iteración #2). Como se verá en las siguientes iteraciones, la combinación de sucesivas variables exógenas mejorará la precisión total del



| Statespace Model Results |                                 |         |                   |       |            |        |
|--------------------------|---------------------------------|---------|-------------------|-------|------------|--------|
| =====                    |                                 |         |                   |       |            |        |
| Dep. Variable:           | Level_BUE                       |         | No. Observations: |       | 8760       |        |
| Model:                   | SARIMAX(3, 0, 12)x(2, 1, 1, 24) |         | Log Likelihood    |       | -29617.220 |        |
| Date:                    | Sat, 01 May 2021                |         | AIC               |       | 59274.441  |        |
| Time:                    | 11:13:32                        |         | BIC               |       | 59415.945  |        |
| Sample:                  | 01-01-2014                      |         | HQIC              |       | 59322.663  |        |
|                          | - 12-31-2014                    |         |                   |       |            |        |
| Covariance Type:         | opg                             |         |                   |       |            |        |
| =====                    |                                 |         |                   |       |            |        |
|                          | coef                            | std err | z                 | P> z  | [0.025     | 0.975] |
| -----                    |                                 |         |                   |       |            |        |
| x1                       | 0.0757                          | 0.013   | 5.826             | 0.000 | 0.050      | 0.101  |
| ar.L1                    | 1.9745                          | 0.012   | 162.559           | 0.000 | 1.951      | 1.998  |
| ar.L2                    | -1.9490                         | 0.013   | -149.038          | 0.000 | -1.975     | -1.923 |
| ar.L3                    | 0.9037                          | 0.012   | 78.378            | 0.000 | 0.881      | 0.926  |
| ma.L1                    | -0.0392                         | 0.015   | -2.650            | 0.008 | -0.068     | -0.010 |
| ma.L2                    | 0.5738                          | 0.015   | 39.363            | 0.000 | 0.545      | 0.602  |
| ma.L3                    | 0.7074                          | 0.023   | 31.222            | 0.000 | 0.663      | 0.752  |
| ma.L4                    | 0.4983                          | 0.029   | 16.898            | 0.000 | 0.441      | 0.556  |
| ma.L5                    | 0.1642                          | 0.035   | 4.706             | 0.000 | 0.096      | 0.233  |
| ma.L6                    | 0.0894                          | 0.035   | 2.577             | 0.010 | 0.021      | 0.157  |
| ma.L7                    | -0.0209                         | 0.034   | -0.613            | 0.540 | -0.088     | 0.046  |
| ma.L8                    | -0.1227                         | 0.031   | -3.943            | 0.000 | -0.184     | -0.062 |
| ma.L9                    | -0.1880                         | 0.025   | -7.550            | 0.000 | -0.237     | -0.139 |
| ma.L10                   | -0.1614                         | 0.018   | -9.055            | 0.000 | -0.196     | -0.126 |
| ma.L11                   | -0.1050                         | 0.013   | -7.991            | 0.000 | -0.131     | -0.079 |
| ma.L12                   | -0.1208                         | 0.011   | -11.038           | 0.000 | -0.142     | -0.099 |
| ar.S.L24                 | 0.1516                          | 0.009   | 17.007            | 0.000 | 0.134      | 0.169  |
| ar.S.L48                 | 0.0825                          | 0.010   | 7.994             | 0.000 | 0.062      | 0.103  |
| ma.S.L24                 | -0.9989                         | 0.010   | -103.116          | 0.000 | -1.018     | -0.980 |
| sigma2                   | 50.7945                         | 0.645   | 78.701            | 0.000 | 49.530     | 52.060 |
| =====                    |                                 |         |                   |       |            |        |
| Ljung-Box (Q):           | 2410.24                         |         | Jarque-Bera (JB): |       | 8191.26    |        |
| Prob(Q):                 | 0.00                            |         | Prob(JB):         |       | 0.00       |        |
| Heteroskedasticity (H):  | 0.95                            |         | Skew:             |       | 1.02       |        |
| Prob(H) (two-sided):     | 0.14                            |         | Kurtosis:         |       | 7.28       |        |
| =====                    |                                 |         |                   |       |            |        |

Figura 4.8: Salida SARIMA (3,0,12)x(2,1,1)<sub>24</sub> con exóg=Wind30mb\_MVD\_3.1

modelo.

### Diagnóstico del modelo

Si bien en relación al RMSE del modelo anterior los valores obtenidos tienen un leve aumento respecto a los valores anteriores (6 %), como se menciona en la sección anterior se evaluarán los valores RMSE resultantes en las sucesivas iteraciones. (Cuadro 4.5).

#### 4.2.3. Iteración 2: SARIMAX (3,0,12)x(2,1,1)<sub>24</sub> y 2 variables exógenas

##### Identificación de modelo tentativo

En una nueva iteración sobre la selección de variables exógenas generadas, se analizó la matriz de correlaciones con el error residual del modelo anterior (*Residual + Wind30mb\_MVD\_3.1*). En particular, se realizaron diversas pruebas con variables generadas a partir de los pronósticos de mareas de puntos geográficos cercanos a Buenos Aires, entre las cuales se obtuvieron correlaciones positivas significativas (Fig.4.9).

En este paso se seleccionó el pronóstico de marea en Buenos Aires, para una ventana de tiempo de 1 hora retrasada 1 hora con respecto a cada observación de la altura del Río de la Plata (*Tide\_BUE\_1-1*). Esto significa que para la predicción de la altura del Río de la Plata se utilizó la información aportada por el pronóstico de marea en Buenos Aires de la hora siguiente.

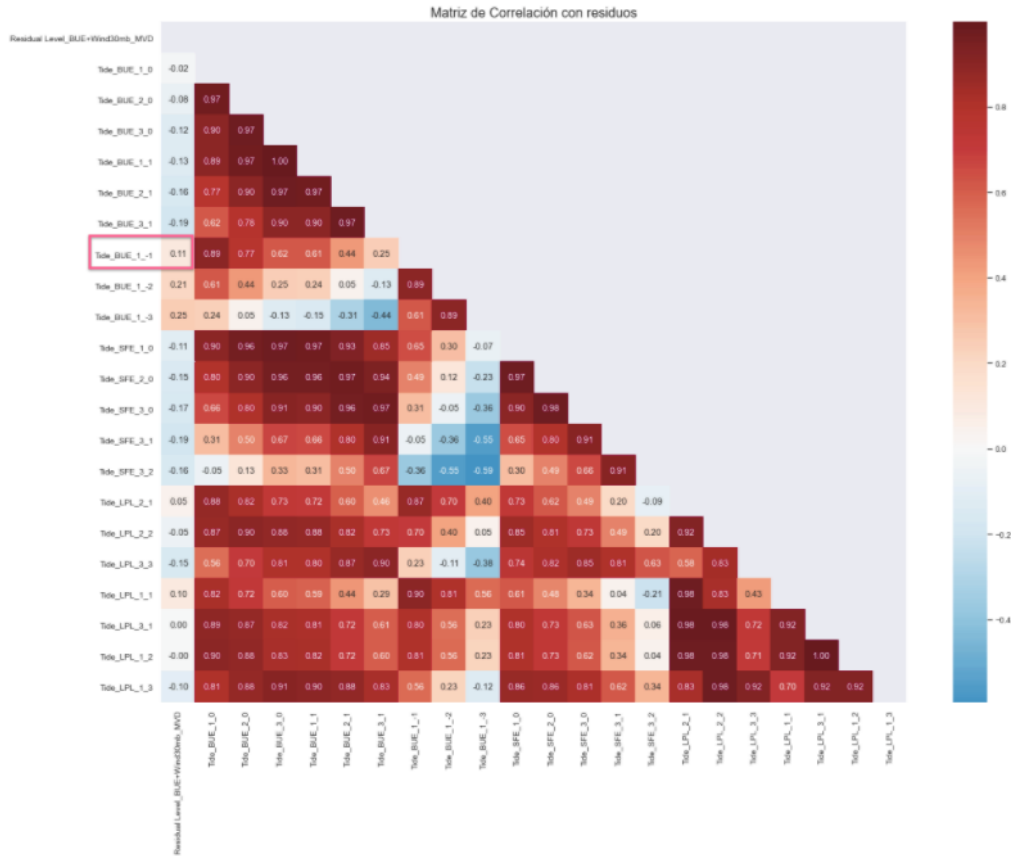


Figura 4.9: Correlación de residuos SARIMAX(3,0,12)x(2,1,1)<sub>24</sub> con exóg. vs pronósticos de marea

## Estimación de parámetros

Como muestra la salida del modelo en la Fig. 4.10, el AIC mostró una considerable mejora con respecto al anterior, indicando que la inclusión del pronóstico de marea en Buenos Aires mejora el modelo previo (Cuadro 4.5 iteración #3).

## Diagnóstico del modelo

Los resultados de las pruebas indicaron que la incorporación de la variable *Tide\_BUE\_1\_1* mejoró la precisión del modelo. El RMSE mostró una disminución del 8 % con respecto al anterior y del 3 % con respecto al modelo base, siendo así el menor obtenido entre los modelos SARIMA y SARIMAX (Cuadro 4.5).

### 4.2.4. Iteración 3: SARIMAX (3,0,12)x(2,1,1)<sub>24</sub> y 3 variables exógenas

#### Identificación de modelo tentativo

En esta iteración, se incorporó la variable generada a partir de la componente *v* del pronóstico de vientos de Mar del Plata (*Wind30mb\_MDP\_05\_100*). Esta variable generada, resultante del cálculo de las componentes vectoriales *u* y *v* del viento Este en Mar

```

Statespace Model Results
=====
Dep. Variable:                Level_BUE      No. Observations:      8760
Model:                SARIMAX(3, 0, 12)x(2, 1, 1, 24)  Log Likelihood      -29061.591
Date:                Sun, 11 Jul 2021      AIC                58165.182
Time:                23:02:38      BIC                58313.761
Sample:                01-01-2014      HQIC               58215.815
                  - 12-31-2014
Covariance Type:                opg
=====
              coef      std err          z      P>|z|      [0.025      0.975]
-----
x1              0.0364      0.012        3.110      0.002        0.013      0.059
x2              0.1130      0.014        7.846      0.000        0.085      0.141
ar.L1           2.6526      0.007    388.094      0.000        2.639      2.666
ar.L2          -2.5769      0.012   -214.836      0.000       -2.600     -2.553
ar.L3           0.9012      0.007    133.151      0.000        0.888      0.914
ma.L1          -0.8337      0.010   -82.303      0.000       -0.854     -0.814
ma.L2          -0.1074      0.011   -9.852      0.000       -0.129     -0.086
ma.L3           0.2878      0.013    22.382      0.000        0.263      0.313
ma.L4           0.2502      0.013    18.809      0.000        0.224      0.276
ma.L5           0.0244      0.014     1.694      0.090       -0.004      0.053
ma.L6           0.0966      0.014     6.895      0.000        0.069      0.124
ma.L7           0.0711      0.015     4.771      0.000        0.042      0.100
ma.L8           0.0063      0.014     0.440      0.660       -0.022      0.034
ma.L9          -0.0337      0.014     -2.430      0.015       -0.061     -0.007
ma.L10          0.0238      0.013     1.879      0.060       -0.001      0.049
ma.L11          0.0978      0.012     8.411      0.000        0.075      0.121
ma.L12          0.0215      0.010     2.117      0.034        0.002      0.041
ar.S.L24       -0.0064      0.010     -0.627      0.530       -0.026      0.014
ma.S.L48       -0.0620      0.011     -5.665      0.000       -0.083     -0.041
ma.S.L24       -0.9167      0.008   -120.969      0.000       -0.932     -0.902
sigma2         45.1294      0.450    100.317      0.000       44.248      46.011
=====
Ljung-Box (Q):                727.68      Jarque-Bera (JB):                6561.96
Prob(Q):                      0.00      Prob(JB):                      0.00
Heteroskedasticity (H):        0.94      Skew:                          0.82
Prob(H) (two-sided):          0.13      Kurtosis:                     6.92
=====

```

Figura 4.10: Salida SARIMA (3,0,12)x(2,1,1)<sub>24</sub> y exóg= *Wind30mb\_v\_MVD\_3.1*, *Ti-de\_BUE\_1-1*

del Plata, adelantado 5 horas, fue la de mayor correlación con los residuos en relación a las variables calculadas a distintos ángulos de dirección y desplazamientos horarios (correlación Pearson=0.11).

### Estimación de parámetros

Con la incorporación de esta variable, el AIC mejora aún más (57911.604 respecto al anterior 58165.182), acercándose al AIC del modelo base (57908.764) (Fig.4.11 y cuadro 4.5 iteración #4).

### Diagnóstico del modelo

En términos de RMSE, la incorporación de la variable *Wind30mb\_MDP\_05\_100* igualó el mismo error que el modelo anterior. Siendo que el criterio de corte considerado fue llegar a un valor mayor en relación al valor de AIC anterior, se procedió a una nueva iteración.

| Statespace Model Results |                                 |         |                   |                   |            |        |
|--------------------------|---------------------------------|---------|-------------------|-------------------|------------|--------|
| Dep. Variable:           | Level_BUE                       |         |                   | No. Observations: | 8760       |        |
| Model:                   | SARIMAX(3, 0, 12)x(2, 1, 1, 24) |         |                   | Log Likelihood    | -28933.802 |        |
| Date:                    | Thu, 06 May 2021                |         |                   | AIC               | 57911.604  |        |
| Time:                    | 09:10:34                        |         |                   | BIC               | 58067.259  |        |
| Sample:                  | 01-01-2014                      |         |                   | HQIC              | 57964.648  |        |
|                          | - 12-31-2014                    |         |                   |                   |            |        |
| Covariance Type:         | opg                             |         |                   |                   |            |        |
|                          | coef                            | std err | z                 | P> z              | [0.025     | 0.975] |
| x1                       | 0.0421                          | 0.012   | 3.628             | 0.000             | 0.019      | 0.065  |
| x2                       | 0.1094                          | 0.014   | 7.666             | 0.000             | 0.081      | 0.137  |
| x3                       | 0.0028                          | 0.006   | 0.513             | 0.608             | -0.008     | 0.014  |
| ar.L1                    | 2.6452                          | 0.007   | 368.836           | 0.000             | 2.631      | 2.659  |
| ar.L2                    | -2.5646                         | 0.013   | -203.775          | 0.000             | -2.589     | -2.540 |
| ar.L3                    | 0.8940                          | 0.007   | 125.491           | 0.000             | 0.880      | 0.908  |
| ma.L1                    | -0.8215                         | 0.010   | -79.527           | 0.000             | -0.842     | -0.801 |
| ma.L2                    | -0.1059                         | 0.011   | -9.813            | 0.000             | -0.127     | -0.085 |
| ma.L3                    | 0.2876                          | 0.013   | 22.772            | 0.000             | 0.263      | 0.312  |
| ma.L4                    | 0.2493                          | 0.013   | 19.021            | 0.000             | 0.224      | 0.275  |
| ma.L5                    | 0.0281                          | 0.014   | 1.981             | 0.048             | 0.000      | 0.056  |
| ma.L6                    | 0.1011                          | 0.014   | 7.226             | 0.000             | 0.074      | 0.128  |
| ma.L7                    | 0.0745                          | 0.015   | 5.023             | 0.000             | 0.045      | 0.104  |
| ma.L8                    | 0.0104                          | 0.014   | 0.734             | 0.463             | -0.017     | 0.038  |
| ma.L9                    | -0.0311                         | 0.014   | -2.274            | 0.023             | -0.058     | -0.004 |
| ma.L10                   | 0.0197                          | 0.013   | 1.574             | 0.115             | -0.005     | 0.044  |
| ma.L11                   | 0.0956                          | 0.011   | 8.335             | 0.000             | 0.073      | 0.118  |
| ma.L12                   | 0.0270                          | 0.010   | 2.666             | 0.008             | 0.007      | 0.047  |
| ar.S.L24                 | 0.0214                          | 0.009   | 2.449             | 0.014             | 0.004      | 0.038  |
| ar.S.L48                 | -0.0411                         | 0.010   | -4.074            | 0.000             | -0.061     | -0.021 |
| ma.S.L24                 | -0.9747                         | 0.004   | -260.739          | 0.000             | -0.982     | -0.967 |
| sigma2                   | 43.6491                         | 0.420   | 104.038           | 0.000             | 42.827     | 44.471 |
| Ljung-Box (Q):           | 699.47                          |         | Jarque-Bera (JB): | 7079.65           |            |        |
| Prob(Q):                 | 0.00                            |         | Prob(JB):         | 0.00              |            |        |
| Heteroskedasticity (H):  | 0.95                            |         | Skew:             | 0.81              |            |        |
| Prob(H) (two-sided):     | 0.18                            |         | Kurtosis:         | 7.10              |            |        |

Figura 4.11: Salida SARIMA  $(3,0,12) \times (2,1,1)_{24}$  y  $\text{exóg} = \text{Wind30mb\_v\_MVD\_3\_1}$ ,  $\text{Tide\_BUE\_1\_1}$  y  $\text{Wind30mb\_MDP\_05\_100}$

#### 4.2.5. Iteración 4: SARIMAX $(3,0,12) \times (2,1,1)_{24}$ y 4 variables exógenas

##### Identificación de modelo tentativo

La variable incorporada en esta iteración fue la generada a partir del pronóstico de mareas de La Plata ( $\text{Tide\_LPL\_2\_0}$ ). Esta fue un promedio de los pronósticos de las dos últimas horas sin desplazamiento. La correlación entre los residuos del modelo anterior con esta variable fue 0.08 (Fig.4.12).

##### Estimación de parámetros

De acuerdo al valor de AIC obtenido, la incorporación de la nueva variable exógena contribuye a la precisión del modelo con respecto al modelo anterior (57907.124 vs 57911.604) (Fig.4.13 y Cuadro 4.5 iteración #5). Asimismo, en este paso se mejora el valor de AIC con respecto al modelo base.

##### Diagnóstico del modelo

En términos de RMSE, la incorporación de la variable  $\text{Tide\_LPL\_2\_0}$  mantiene el mismo orden de error.

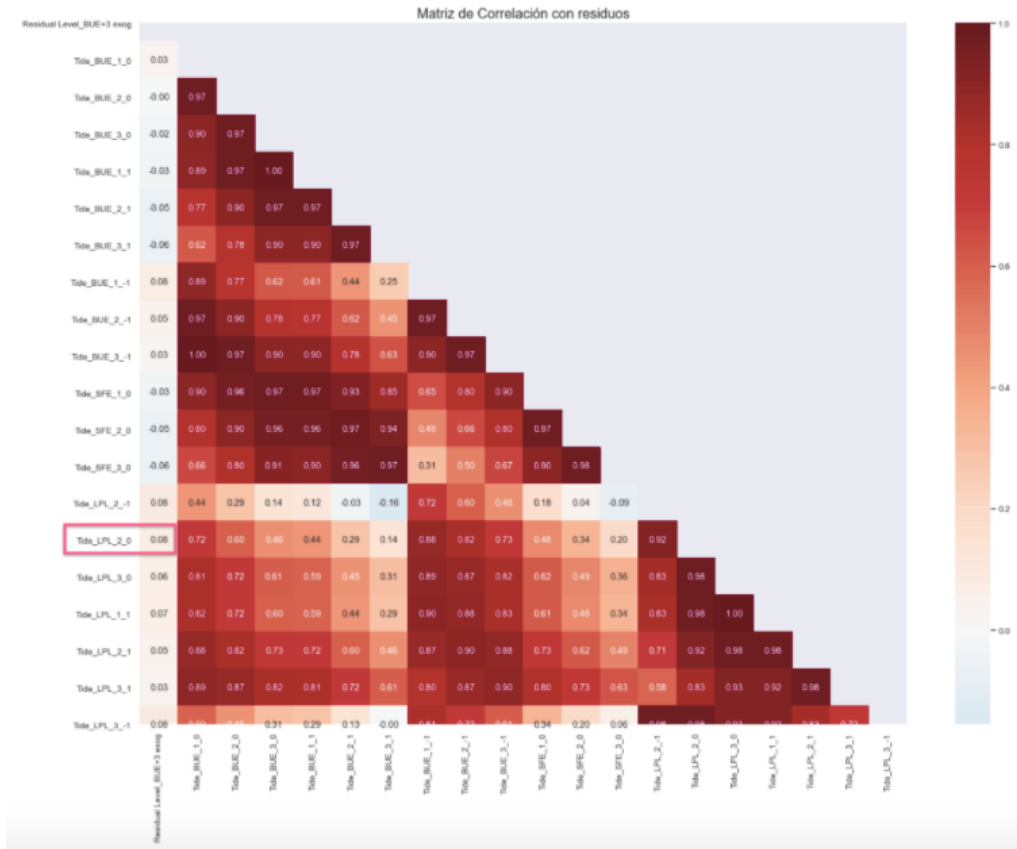


Figura 4.12: Correlación de residuos SARIMAX(3,0,12)x(2,1,1)<sub>24</sub> con 3 exóg. vs tabla de mareas

#### 4.2.6. Iteración 5: SARIMAX (3,0,12)x(2,1,1)<sub>24</sub> y 5 variables exógenas

##### Identificación de modelo tentativo

En la iteración final se llega al criterio de corte, al obtenerse un valor de AIC mayor al modelo anterior. En distintas pruebas las variables incorporadas generaron que los modelos resultantes dieran valores mayores en este criterio con respecto al del último modelo obtenido. En el Cuadro 4.5 se muestra como ejemplo los resultados al incorporar los pronósticos de viento de Villa Gesell 5 horas antes del registro de alturas del Río de la Plata.

### 4.3. Redes neuronales

#### 4.3.1. Memoria de largo corto plazo (Long Short-Term Memory o LSTM)

El modelado de las alturas del Río de la Plata implementado con LSTM consideró 3 pasos básicos: la construcción de la red, el entrenamiento del modelo y el diagnóstico del mismo. El modelado abarcó el mismo período y variables de pronósticos de marea y vientos que los modelos anteriores para una mejor comparación posterior.

Es importante aclarar que si bien la preparación de los datos consideró aplicar una dife-

| Statespace Model Results |                                 |         |                   |       |            |        |
|--------------------------|---------------------------------|---------|-------------------|-------|------------|--------|
| =====                    |                                 |         |                   |       |            |        |
| Dep. Variable:           | Level_BUE                       |         | No. Observations: |       | 8760       |        |
| Model:                   | SARIMAX(3, 0, 12)x(2, 1, 1, 24) |         | Log Likelihood    |       | -28930.562 |        |
| Date:                    | Sun, 09 May 2021                |         | AIC               |       | 57907.124  |        |
| Time:                    | 08:57:33                        |         | BIC               |       | 58069.854  |        |
| Sample:                  | 01-01-2014                      |         | HQIC              |       | 57962.579  |        |
|                          | - 12-31-2014                    |         |                   |       |            |        |
| Covariance Type:         | opg                             |         |                   |       |            |        |
| =====                    |                                 |         |                   |       |            |        |
|                          | coef                            | std err | z                 | P> z  | [0.025     | 0.975] |
| -----                    |                                 |         |                   |       |            |        |
| x1                       | 0.0424                          | 0.012   | 3.653             | 0.000 | 0.020      | 0.065  |
| x2                       | 0.1117                          | 0.015   | 7.589             | 0.000 | 0.083      | 0.141  |
| x3                       | -0.0008                         | 0.027   | -0.031            | 0.975 | -0.054     | 0.052  |
| x4                       | 0.0027                          | 0.006   | 0.489             | 0.625 | -0.008     | 0.014  |
| ar.L1                    | 2.6433                          | 0.007   | 365.068           | 0.000 | 2.629      | 2.658  |
| ar.L2                    | -2.5609                         | 0.013   | -201.545          | 0.000 | -2.586     | -2.536 |
| ar.L3                    | 0.8921                          | 0.007   | 124.330           | 0.000 | 0.878      | 0.906  |
| ma.L1                    | -0.8198                         | 0.010   | -78.992           | 0.000 | -0.840     | -0.799 |
| ma.L2                    | -0.1063                         | 0.011   | -9.859            | 0.000 | -0.127     | -0.085 |
| ma.L3                    | 0.2872                          | 0.013   | 22.696            | 0.000 | 0.262      | 0.312  |
| ma.L4                    | 0.2487                          | 0.013   | 18.972            | 0.000 | 0.223      | 0.274  |
| ma.L5                    | 0.0280                          | 0.014   | 1.978             | 0.048 | 0.000      | 0.056  |
| ma.L6                    | 0.0995                          | 0.014   | 7.124             | 0.000 | 0.072      | 0.127  |
| ma.L7                    | 0.0743                          | 0.015   | 5.024             | 0.000 | 0.045      | 0.103  |
| ma.L8                    | 0.0113                          | 0.014   | 0.799             | 0.424 | -0.016     | 0.039  |
| ma.L9                    | -0.0298                         | 0.014   | -2.181            | 0.029 | -0.057     | -0.003 |
| ma.L10                   | 0.0198                          | 0.012   | 1.586             | 0.113 | -0.005     | 0.044  |
| ma.L11                   | 0.0965                          | 0.011   | 8.421             | 0.000 | 0.074      | 0.119  |
| ma.L12                   | 0.0276                          | 0.010   | 2.718             | 0.007 | 0.008      | 0.048  |
| ar.S.L24                 | 0.0234                          | 0.009   | 2.725             | 0.006 | 0.007      | 0.040  |
| ar.S.L48                 | -0.0364                         | 0.010   | -3.634            | 0.000 | -0.056     | -0.017 |
| ma.S.L24                 | -0.9803                         | 0.003   | -307.648          | 0.000 | -0.987     | -0.974 |
| sigma2                   | 43.5277                         | 0.418   | 104.097           | 0.000 | 42.708     | 44.347 |
| =====                    |                                 |         |                   |       |            |        |
| Ljung-Box (Q):           | 698.00                          |         | Jarque-Bera (JB): |       | 7288.59    |        |
| Prob(Q):                 | 0.00                            |         | Prob(JB):         |       | 0.00       |        |
| Heteroskedasticity (H):  | 0.95                            |         | Skew:             |       | 0.83       |        |
| Prob(H) (two-sided):     | 0.19                            |         | Kurtosis:         |       | 7.15       |        |

Figura 4.13: Salida SARIMA (3,0,12)x(2,1,1)<sub>24</sub> y exóg=Wind30mb\_v\_MVD\_3.1, Tide\_BUE\_1-1, Wind30mb\_MDP\_05\_100, Tide\_LPL\_2.0

| CI y errores |   |   |    |   |   |   |    |                                                                                                  |           |      |
|--------------|---|---|----|---|---|---|----|--------------------------------------------------------------------------------------------------|-----------|------|
| It.          | p | d | q  | P | D | Q | S  | Exóg                                                                                             | AIC       | RMSE |
| 1            | 3 | 0 | 12 | 2 | 1 | 1 | 24 | -                                                                                                | 57908.764 | 7.7  |
| 2            | 3 | 0 | 12 | 2 | 1 | 1 | 24 | Wind30mb_v_MVD_3.1                                                                               | 59274.441 | 8.2  |
| 3            | 3 | 0 | 12 | 2 | 1 | 1 | 24 | Wind30mb_v_MVD_3.1<br>Tide_BUE_1-1                                                               | 58165.182 | 7.5  |
| 4            | 3 | 0 | 12 | 2 | 1 | 1 | 24 | Wind30mb_v_MVD_3.1<br>Tide_BUE_1-1<br>Wind30mb_MDP_05_100                                        | 57911.604 | 7.5  |
| 5            | 3 | 0 | 12 | 2 | 1 | 1 | 24 | Wind30mb_v_MVD_3.1<br>Tide_BUE_1-1<br>Wind30mb_MDP_05_100<br>Tide_LPL_2.0                        | 57907.124 | 7.5  |
| 6            | 3 | 0 | 12 | 2 | 1 | 1 | 24 | Wind30mb_v_MVD_3.1<br>Tide_BUE_1-1<br>Wind30mb_MDP_05_100<br>Tide_LPL_2.0<br>Wind30mb_VGL_05_100 | 57924.400 | 7.4  |

Cuadro 4.5: Criterio de información y RMSE de modelos

renciación de -12hs a fin de eliminar la estacionalidad de la serie, el error final calculado sobre las predicciones fue mucho mayor que en el caso que se presenta, y por lo tanto esta variación no fue incluida en el presente análisis.

### Construcción de la red

En la fase de construcción se definió un modelo de red neuronal secuencial de 3 capas (entrada, oculta y salida), siendo este un diseño adoptado en muchos estudios pre-

vios que prueban que esta arquitectura modela datos no lineales de manera satisfactoria [Fashae et al., 2019].

Se generó un modelo secuencial con la siguiente estructura organizada en una pila o *stack*:

- una capa de entrada LSTM que recibe arreglos de dimensión (24, 5), correspondiente a 1 día de observaciones horarias de las alturas del agua en Buenos Aires (*Level\_BUE*), pronóstico de vientos CFS de Montevideo (*Wind30mb\_v\_MVD\_3-1*), pronóstico de mareas de Buenos Aires (*Tide\_BUE\_1-1*), pronóstico de vientos de Mar del Plata (*Wind30mb\_MDP\_05\_100*) y pronóstico de mareas de La Plata (*Tide\_LPL\_2-0*). La función de activación aplicada fue la tangente hiperbólica (*tanh*).
- una capa oculta densa, totalmente conectada con función de activación lineal. Esta capa que permite calcular múltiples pronósticos en un paso, con 60 unidades de salida (predicciones a 12 horas \* 5 variables).

$$h_{w,b}(X) = X \cdot w + b$$

- una capa de salida que retorna 12 horas de predicciones de las alturas del agua en Buenos Aires.

### Entrenamiento del modelo

El entrenamiento de la red aplicó un algoritmo de optimización RMSprop, el cual mantiene una media móvil del cuadrado de los gradientes para cada peso. Esto evita problemas de convergencia como ocurre en el caso de otros algoritmos que acumulan todos los gradientes desde el inicio del entrenamiento. Tanto la función de pérdida aplicada en el proceso de entrenamiento como la métrica de rendimiento del modelo aplican el cálculo de error cuadrático medio (MSE). Matemáticamente, es la función preferida para problemas de regresión donde la variable dependiente es gausseana. A fin de evitar sobreajuste del conjunto de entrenamiento, se configuró una función de detención temprana (o *early stopping* por su nombre en inglés), que define el límite de iteraciones luego del cual de no observarse mejora en el conjunto de validación el entrenamiento se interrumpe.

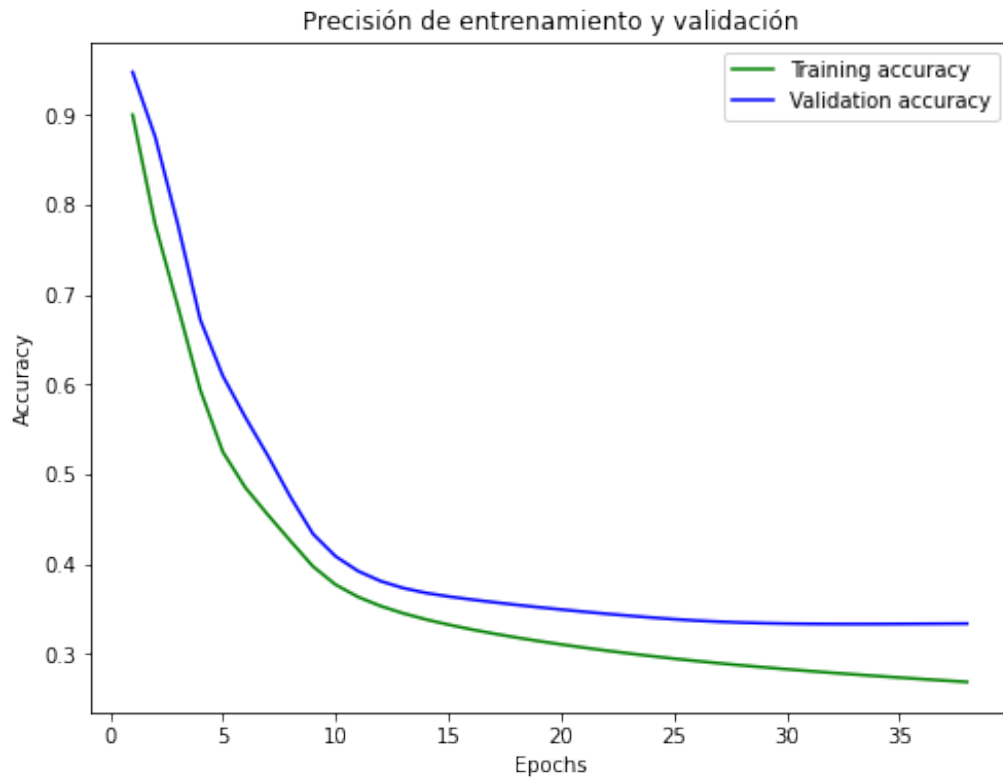


Figura 4.14: Configuración del modelo LSTM - Métrica rendimiento (MSE)

El modelo fue entrenado con 100 iteraciones o *epochs*. La pérdida en el conjunto de entrenamiento es menor que la del conjunto de validación y continua disminuyendo hasta la iteración #40, a diferencia del conjunto de validación que se mantiene casi constante a partir de la iteración #15. Esto indicaría cierto nivel de sobreajuste del modelo en el conjunto de entrenamiento a partir de ese punto.

### Diagnóstico del modelo

Como se puede observar en la Fig. 4.14, desde el punto de vista de la aplicación del modelo la precisión empieza a mantenerse en un valor cercano a 0.35 desde la iteración #15, donde el mismo mostró una mejor precisión en entrenamiento que en validación, indicando un probable sobreajuste del modelo en ese conjunto de datos.



## Capítulo 5

# Comparación de resultados

### 5.1. Comparación entre modelos lineales y no lineales generados

Para la evaluación de los modelos generados, se calculó el error de precisión acumulado (APE). Como se mencionara en la sección 2.4.2, el APE calcula el RMSE promedio de las predicciones de un modelo sobre series que iterativamente incorporan observaciones a la vez que mantienen la dependencia temporal de las mismas. Esto permite que la evaluación del modelo sea más robusta, contemplando su precisión tanto en períodos regulares como en eventos extremos de ondas de tormenta y ondas de tormenta negativas.

Para cada clase de modelo, se entrenó inicialmente un modelo base entre Enero y Diciembre de 2014, y sucesivos modelos que de manera incremental incorporaron datos en períodos de 1 semana hasta Diciembre de 2017.

En el Cuadro 5.1, se listan los resultados obtenidos para cada uno de los modelos generados para horizontes de tiempo entre 1 y 12 horas. Con fines comparativos, se incluyeron resultados de dos regresiones lineales adicionales, donde los pronósticos fueron la última hora y el promedio anual respectivamente:

| Predicc.<br>(hs) | Regr.lin.<br>(Últ.<br>altura) | Regr.lin.<br>(Prom.<br>anual) | Regr.lin. | Srmx<br>(3,0,12)x<br>(2,1,1) <sub>24</sub> | LSTM(3<br>capas) |
|------------------|-------------------------------|-------------------------------|-----------|--------------------------------------------|------------------|
| 1                | 0.13                          | 0.56                          | 0.06      | 0.06                                       | 0.08             |
| 2                | 0.26                          | 0.57                          | 0.12      | 0.13                                       | 0.14             |
| 3                | 0.36                          | 0.57                          | 0.17      | 0.19                                       | 0.18             |
| 4                | 0.44                          | 0.57                          | 0.21      | 0.24                                       | 0.22             |
| 5                | 0.50                          | 0.57                          | 0.24      | 0.28                                       | 0.25             |
| 6                | 0.54                          | 0.56                          | 0.25      | 0.31                                       | 0.27             |
| 12               | 0.53                          | 0.55                          | 0.27      | -                                          | 0.38             |

Cuadro 5.1: RMSE de pronósticos entre 1h y 12hs (en mts)

Analizando la regresión lineal que asumió como mejor predicción la última hora observada, se puede ver que los errores de los distintos horizontes pronosticados fueron 2 veces los correspondientes de la mejor regresión lineal. Por otro lado, la regresión que asumió como mejor predicción el promedio anual dio como resultado un error estable de entre 0.55 y 0.57m, 2 veces el error de la mejor regresión en un pronóstico a 12hs.

Como se observa, la regresión lineal seleccionada, SARIMAX y LSTM tienen precisión

similar en las primeras horas pronosticadas. Ya en un pronóstico a 12hs se puede ver que el modelo no lineal se aleja en términos de precisión, con un APE 40 % mayor.

## 5.2. Comparación con modelos numéricos preexistentes

Con el objeto de comparar los resultados obtenidos con modelos preexistentes productivos, se analizaron resultados de dos modelos numéricos:

- **Modelo de pronóstico de olas SMARA/WAM**

Modelo de pronóstico de olas de tercera generación SMARA/WAM. Al momento de la publicación citada, el mismo se incluyó en un sistema troncal y de pronóstico numérico de tiempo en el Río de la Plata. Según dicha trabajo, el modelo se conducía por modelos meteorológicos, y permitía la asimilación de datos de distintas fuentes de información. Fue utilizado con fines de pronóstico. El modelo numérico de olas SMARA/WAM se aplicó en el Atlántico Sudoccidental desde 30°S hasta la costa antártica y desde 20°O hasta 70°O. Los pronósticos y retroanálisis fueron conducidos por campos de viento a 10m producidos por el National Centers for Environmental Prediction (NCEP). Los campos de viento eran actualizados cada 6 horas en el retroanálisis pero sólo se realizaban dos pronósticos diarios [Etala et al., 2009]. Para la estación de Palermo en el Río de la Plata se dispone del pronóstico de la Central de Prevención de Crecidas del Río de la Plata del SHN, que está basado en observaciones recientes del nivel del agua.

- **Modelo de pronóstico de ondas de tormenta Eta SMN**

Modelo de pronóstico numérico de ondas de tormenta Eta SMN (versión 2005), parte integrante del sistema general de pronóstico numérico. La performance de los modelos de ondas de tormenta se monitorea en forma continua con los datos de nivel del agua que integran la red de mareógrafos del SHN en distintos puntos de la costa, incluyendo Palermo. Las predicciones de marea astronómica realizadas por el SHN para cada ubicación son adicionadas a la onda de tormenta modelada para obtener el nivel del agua total.

Para ambos modelos, el análisis comparativo entre estos y los modelos generados considera los resultados publicados en [Etala et al., 2009] entre Mayo de 2007 y Febrero de 2009, donde se muestra el MAE calculado en la estación Palermo.

Para la comparación con el modelo Eta SMN se incluyeron además los resultados arrojados por el modelo en el período Enero 2014 a Diciembre 2017, mismo período al incluido en este trabajo.

### 5.2.1. Error medio de pronóstico del nivel de agua >1m

La Fig.5.1 muestra el error medio de pronóstico de los modelos de olas y ondas de tormenta calculado en el período Mayo 2007 a Febrero 2009 en Palermo. Este pronóstico está basado en observaciones recientes del nivel del agua. El tiempo de propagación de la onda de tormenta a lo largo de la costa establece un límite de anticipación posible de este último pronóstico. Por eso, en la figura el plazo se sitúa por debajo de la primera valicación del

pronóstico numérico, que es a las 12hs de predicción. En ese lapso, también se realizaron actualizaciones de pronóstico, denominadas *nowcast*. Ambos pronósticos se utilizan como referencia para evaluar la calidad del producto del modelo Eta SMN.

En la figura se observa que los errores del pronóstico numérico Eta SMN se mantienen por debajo hasta poco más de las 48hs y a excepción del *nowcast*, que resulta ser la predicción de mejor performance. Esto podría entenderse como que el sistema numérico extiende el plazo posible de pronóstico a 48hs.

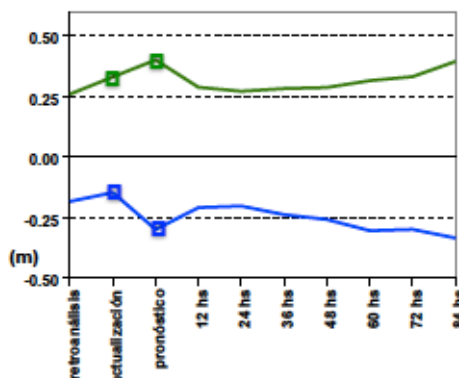


Figura 5.1: Error medio (azul) y desvío estándar del error (verde) de los pronósticos de nivel del agua de los modelos de olas y de ondas de tormenta en Palermo. El *nowcast* oficial (SHN) o actualización del pronóstico presenta el menor error sistemático. Los cuadrados representan los pronósticos de plazo más corto del SHN, basados en observaciones de mareógrafos e independientes del sistema numérico.

Al comparar los resultados de los modelos generados en un período similar, el error de la regresión lineal y LSTM para el pronóstico a 6hs (0.20m y 0.22m respectivamente) no mejora la performance del *nowcast* oficial ( $\sim -0.17$ m). En cambio, al comparar un horizonte de 12hs, mientras el pronóstico oficial alcanza  $\sim -0.30$ m la regresión muestra un error aproximado al modelo Eta SMN, mejorando la performance de la red neuronal (Cuadro 5.2).

|            | May/2007 - Feb/2009 | May/2015 - Feb/2017 |              |         |      |
|------------|---------------------|---------------------|--------------|---------|------|
| Pronóstico | Pron. oficial       | Eta SMN             | Regr. lineal | SARIMAX | LSTM |
| 6hs        | $\sim -0.17$        | -                   | 0.20         | 0.24    | 0.22 |
| 12hs       | $\sim -0.30$        | 0.17                | 0.21         | -       | 0.28 |

Cuadro 5.2: Error medio de pronóstico oficial, Eta SMN, regresión lineal, SARIMAX y LSTM (período 2 años)

### 5.2.2. Error medio de pronóstico en períodos de sudestada y bajante

Este análisis comparativo evaluó los resultados del modelo Eta SMN y de los obtenidos por los modelos generados durante eventos de sudestadas y bajantes en el Río de la Plata.

Para la selección de eventos de sudestada, se adoptó el criterio del trabajo de [Escobar G, 2004], tomando registros de nivel de agua  $\geq 1.60$ m persistentes por al menos 24hs. En el caso de eventos de bajante, la selección consideró aquellos registros de onda de tormenta negativa que sobrepasaron un umbral de  $-1.80$ m [Cerne et al., 2005].

Bajo estas definiciones se analizaron los pronósticos a 6hs y 12hs para eventos ocurridos

entre 2014 y 2017 (Cuadro 5.3 y Fig.5.2).

| # | Fecha     | Tipo | 12:00hs (+6) |            |         |       | 18:00hs (+12) |              |              |             |
|---|-----------|------|--------------|------------|---------|-------|---------------|--------------|--------------|-------------|
|   |           |      | Nivel(m)     | Regr. lin. | SARIMAX | LSTM  | Nivel(m)      | Eta SMN      | Regr. lin.   | LSTM        |
| 1 | 18/Abr/16 | △    | 2.14         | 0.88       | 0.90    | 0.38  | <b>3.14</b>   | <b>0.16</b>  | 1.04         | <b>1.46</b> |
| 2 | 18/Oct/16 | △    | 1.81         | -0.24      | 0.21    | -0.06 | 2.14          | -0.21        | 0.14         | 0.30        |
| 3 | 25/Ago/14 | ▽    | -0.46        | -0.41      | 0.47    | -0.68 | <b>-1.27</b>  | <b>-0.85</b> | <b>-1.65</b> | -1.63       |
| 4 | 11/Sep/17 | ▽    | -1.07        | -0.47      | -1.00   | -0.79 | -1.00         | -0.48        | -0.40        | -1.38       |

Cuadro 5.3: Nivel del agua observado en Palermo y error medio de pronósticos a 6hs y 12hs durante sudestadas y bajantes (Palermo, 2015-2017)

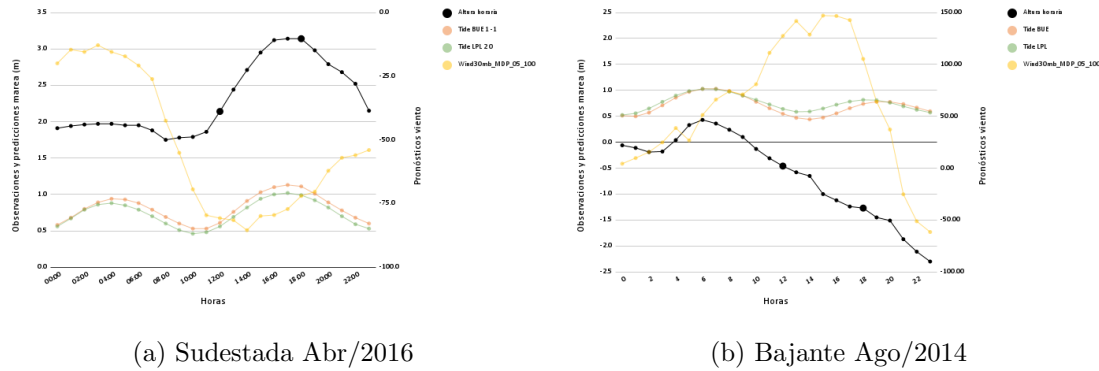


Figura 5.2: Valores de las series durante eventos de niveles de agua extremos en el Río de la Plata

El error medio del pronóstico del nivel del agua en los períodos de sudestada para Palermo se muestra en las Fig. 5.3a y 5.3b. En ambas figuras se observa que el error del modelo Eta SMN a 12hs se mantiene en el orden de los 0.20m, mientras que en la regresión lineal y LSTM obtiene un orden de error similar en la predicción de la bajante de Octubre, pero superando 1m de error para el evento más pronunciado durante Abril del mismo año.

En la práctica se podría decir que el sistema numérico mantiene la calidad en pronósticos a 12hs para eventos de sudestada en el Río de la Plata. Por el contrario, el modelo lineal y la red neuronal no evidencian tener un comportamiento estable frente a crecientes generadas en el corto plazo de algunas horas.

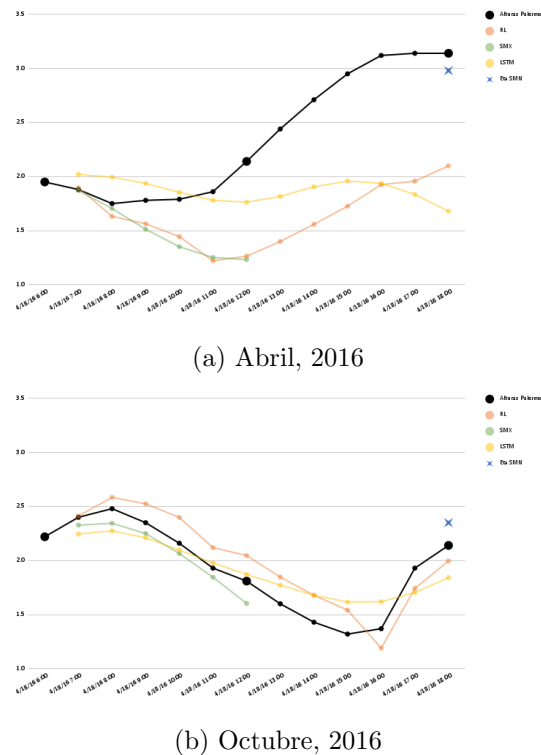


Figura 5.3: Sudestadas Palermo, 2015-2016

Analizando las ondas de tormenta negativas durante Agosto de 2014 y Septiembre de 2017, se observa que si bien la regresión lineal tiene un error medio menor en el primer evento que los restantes modelos, en el segundo evento con una bajante pronunciada de 1.70m en 12hs su error medio aumenta considerablemente (Fig. 5.4a y 5.4b).

LSTM registró el mayor error medio en ambos casos, tanto en la predicción a 6hs como en la de 12hs. Por otro lado, contrario a los resultados obtenidos durante las sudestadas, el error medio de los pronósticos del modelo numérico a 12hs se incrementó entre 3 y 4 veces.

Observando el error medio de los modelos lineales y LSTM en los pronósticos de sudestada y bajante extremos con variaciones mayores a 1.2m en una ventana de 12hs (Cuadro 5.3 #1 #3), se vio que la capacidad predictora disminuyó considerablemente en cuanto a la precisión, registrándose errores medios  $>1m$ .

### 5.2.3. Error de predicción acumulado entre 2014 y 2017

Utilizando la misma métrica que en la comparación de los modelos generados, se calculó el error de predicción acumulado (APE) sobre el modelo numérico con un horizonte de pronósticos a 12hs para el período completo considerado en este trabajo. El modelo numérico no muestra el error calculado a 6hs, ya que el solamente se contó con pronósticos a 12hs.

Los resultados del error calculado se muestran en el Cuadro 5.4.

| Hs | Eta SMN | Regr. lin. | SARIMAX | LSTM |
|----|---------|------------|---------|------|
| 6  | -       | 0.25       | 0.31    | 0.27 |
| 12 | 0.22    | 0.27       | -       | 0.37 |

Cuadro 5.4: APE calculado en predicciones a 6hs y 12hs entre 2014 y 2017

En el cuadro se observa que para los modelos generados el error en pronósticos a 6hs es levemente menor en la regresión lineal, aunque el error de SARIMAX y LSTM se ubica en el mismo orden.

Ya en pronósticos a 12hs, el modelo numérico mejora en términos del error calculado, con un APE final de 0.22m, en el mismo orden que el modelo lineal, con 0.27m, ambos mejorando considerablemente el error de 0.37m obtenido con LSTM.

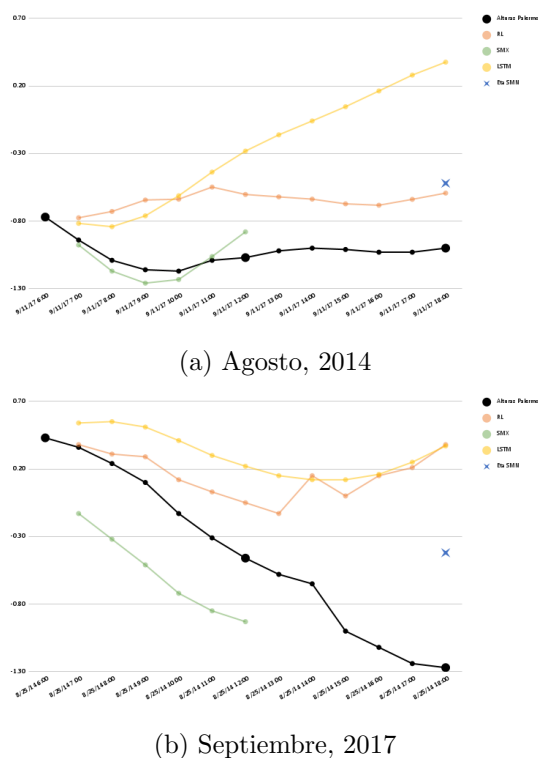


Figura 5.4: Bajantes Palermo, 2014-2017

## Capítulo 6

# Discusión y conclusiones

Se han analizado los factores que hacen de la predicción de alturas en el Río de la Plata un problema complejo a predecir con cierto rango de horas de anticipación. Por otro lado, el análisis de patrones y la generación de modelos estuvieron condicionados por la falta de datos en períodos consecutivos en determinados puntos de medición, lo que forzó en casos como vientos y alturas del agua a reemplazos por pronósticos aproximados.

Si bien los modelos lineales demostraron buenos resultados en cuanto a su capacidad predictiva, en el caso de SARIMAX esta quedó limitada a un horizonte de 6hs. Esta limitación, dada por la implementación de este modelo en la incorporación de variables exógenas, fue un condicionante al momento de comparar los resultados con otros modelos.

Siguiendo las correlaciones halladas de los pronósticos con el error residual de los modelos, las implementaciones incorporaron un ensamble de 4 variables generadas a partir de medias móviles ( $mm$ ) y desplazamientos temporales ( $dt$ ). La incidencia de los pronósticos de vientos provenientes de Montevideo ( $3_{mm} + 1_{dt}$ ) y Mar del Plata ( $5_{mm}$ ) sobre las predicciones del nivel del agua en el Río de la Plata fueron significativas, así como los pronósticos de marea astronómica estimados en puntos cercanos como Buenos Aires ( $1_{mm} + 1_{dt}$ ) y La Plata ( $2_{mm}$ ).

Los modelos lineales obtuvieron buenos resultados, comparables en rendimiento a los obtenidos por los modelos preexistentes oficial y Eta SMN. Evaluando los resultados en períodos similares con crecientes  $>1m$ , el error medio de los modelos generados para pronósticos a 6hs se mantuvo en el mismo orden que el pronóstico de olas oficial, mejorándolo un 30 % en pronósticos a 12hs. En este último escenario, el modelo numérico Eta SMN obtuvo el menor error medio.

Sin embargo, al analizar las predicciones de eventos extremos específicos la precisión obtenida no se mantuvo. Al comparar los resultados con las observaciones tanto en la sudestada como en la bajante máximas de estudio, se hizo evidente que los modelos generados subestiman los picos de altura. Los modelos generados para pronósticos a 6hs incrementaron el error medio entre 2 y 4 veces con respecto a los calculados en el período antes evaluado. El modelo numérico de ondas de tormenta mantuvo su precisión en ambos eventos de sudestada, pero no así en las bajantes de Agosto 2014 y Septiembre 2017 donde su error medio llegó a ser hasta 4 veces mayor. Los resultados en estos eventos no fueron concluyentes.

Desde un punto de vista más genérico, el cálculo del APE en el período 2014 a 2017 refleja que la regresión lineal y SARIMAX tienen error acumulado de mismo orden en

predicciones a 6hs, pero no mejoran el resultado a 12hs del modelo de ondas de tormenta, que mostró un error 0.05m y 0.15m mejor que la regresión y el LSTM respectivamente.

La aplicación de modelos en la predicción de alturas del Río de la Plata presenta sin dudas resultados positivos. Entre las redes neuronales, LSTM fue capaz de predecir tendencias no lineales de la serie. En pronósticos a corto plazo el error acumulado calculado entre 2014 y 2017 se mantuvo apenas por debajo del error obtenido con modelo lineal SARIMAX (0.04m), pero tuvo una tendencia creciente en pronósticos de un horizonte mayor con respecto a la regresión lineal. La regresión por otro lado mantuvo un error estable en pronósticos a 12hs, y fue el que más aproximó a la mejor precisión obtenida con el modelo numérico (Fig.6.1).

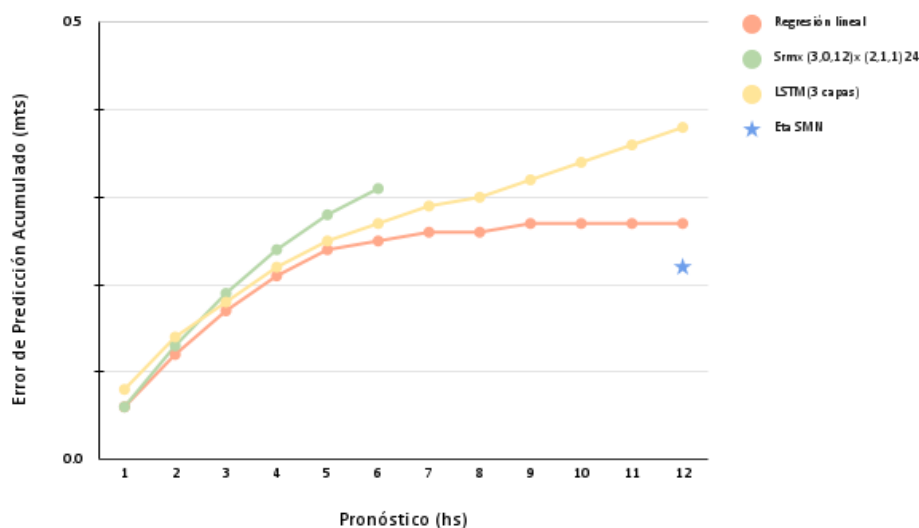


Figura 6.1: Error de Predicción Acumulado (APE) 2014-2017

Los resultados indican que LSTM parece mostrar un mejor rendimiento comparado con el modelo autorregresivo de media móvil SARIMAX. En principio, esto muestra una diferencia con respecto a resultados publicados en trabajos previos sobre series de tiempo, donde los modelos autorregresivos de media móvil presentaron resultados mejores ([Fashae et al., 2019]).

Tradicionalmente, las redes neuronales puras en predicciones de series de tiempo no han obtenido los resultados más óptimos en términos de precisión. Sin embargo, en los últimos años han habido importantes avances de modelos de deep learning en este área, muchos de ellos surgidos a partir de competencias internacionales abiertas donde si bien los modelos estadísticos eran los ganadores, a partir del 2018 lo son modelos estadísticos combinados con modelos de deep learning.

Si bien nuevos modelos de redes neuronales puros como N-BEATS (red neuronal univariada [Oreshkin et al., 2019]), o N-BEATSx (que extiende a la anterior incorporando variables exógenas [Olivares et al., 2021]), podrían ser una futura línea de trabajo, es posible que no podamos alejarnos completamente de la necesidad de un conocimiento previo o pre-procesamiento de la serie de tiempo que incluya técnicas estadísticas. En este sentido los modelos híbridos plantean un camino con potenciales buenos resultados, como se muestra en la publicación de [Choi, 2018] sobre el rendimiento de un modelo ARIMA-LSTM en la predicción del precio de acciones.

# Bibliografía

- [Agencia de Protección Ambiental, 2014] Agencia de Protección Ambiental, M. d. A. y. E. P. (2014). Informe anual ambiental 2014. pages 17–27.
- [Balay, 1961] Balay, M. A. (1961). *El Río de la Plata entre la atmósfera y el mar*. Servicio de Hidrografía Naval, Argentina.
- [Bergmeir and Benítez, 2012] Bergmeir, C. and Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213.
- [Bishop, 2006] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- [Box and Jenkins, 1994] Box, G. E. P. and Jenkins, G. M. (1994). *Time Series Analysis: Forecasting and Control*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 3rd edition.
- [Cerne et al., 2005] Cerne, S., Possia, N., Campetella, C., D’Onofrio, E., Fiore, M., Saurral, R., and Bertolotti, M. (2005). Casos extremos de las bajantes del río de la plata.
- [Changnon, 2000] Changnon, S. A. (2000). *El Niño 1997-1998: the climate Event of the Century*. Oxford University Press.
- [Choi, 2018] Choi, H. K. (2018). Stock price correlation coefficient prediction with arima-lstm hybrid model. *arXiv preprint arXiv:1808.01560*.
- [Couharde et al., 2019] Couharde, C., Damette, O., Generoso, R., and Mohaddes, K. (2019). The Growth Effects of El Nino and La Nina: Local Weather Conditions Matter. Cambridge Working Papers in Economics 1992, Faculty of Economics, University of Cambridge.
- [Cryer and Kellet, 1991] Cryer, J. D. and Kellet, N. (1991). *Time series analysis*. Springer.
- [D’Onofrio et al., 1999] D’Onofrio, E., Fiore, M., and Romero, S. (1999). Return periods of extreme water levels estimated for some vulnerable areas of buenos aires. *Continental Shelf Research*, 19:1681–1693.
- [Escobar G, 2004] Escobar G, Vargas W, B. S. (2004). Wind tides in the Río de la Plata estuary: meteorological conditions. *Int J Climatol* 24, pages 1159–1169.
- [Etala, 1993] Etala, M. (1993). El modelado en la meteorología marina. *Boletín del Centro Naval*, pages 61–67.
- [Etala, 2009] Etala, P. (2009). On the accuracy of atmospheric forcing for extra-tropical storm surge prediction. *Natural hazards*, 51(1):49–61.
- [Etala et al., 2009] Etala, P., Alonso, S. M., Souto, D., Romero, C., and Suaya, M. (2009). Pronóstico numérico marino en el shn/smn.



- [Fashae et al., 2019] Fashae, O. A., Olusola, A. O., Ndubuisi, I., and Udombos, C. G. (2019). Comparing ann and arima model in predicting the discharge of river opeki from 2010 to 2020. *River research and applications*, 35(2):169–177.
- [Fox, 2015] Fox, J. (2015). *Applied Regression Analysis and Generalized Linear Models*. SAGE Publications.
- [Hastie et al., 2009] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer.
- [Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.
- [Lee and Jeng, 2002] Lee, T. and Jeng, D. (2002). Application of artificial neural networks in tide-forecasting. *Ocean Engineering*, 29(9):1003–1022.
- [Lee, 2004] Lee, T.-L. (2004). Back-propagation neural network for long-term tidal predictions. *Ocean Engineering*, 31(2):225–238.
- [Lee, 2008] Lee, T.-L. (2008). Back-propagation neural network for the prediction of the short-term storm surge in taichung harbor, taiwan. *Engineering Applications of Artificial Intelligence*, 21(1):63–72.
- [Olivares et al., 2021] Olivares, K. G., Challu, C., Marcjasz, G., Weron, R., and Dubrawski, A. (2021). Neural basis expansion analysis with exogenous variables: Forecasting electricity prices with nbeatsx. *arXiv preprint arXiv:2104.05522*.
- [Oreshkin et al., 2019] Oreshkin, B. N., Carpo, D., Chapados, N., and Bengio, Y. (2019). N-beats: Neural basis expansion analysis for interpretable time series forecasting. *arXiv preprint arXiv:1905.10437*.
- [Perillo et al., 2012] Perillo, G. M., Piccolo, M. C., and Pino-Quivira, M. (2012). *Estuaries of South America: their geomorphology and dynamics*. Springer Science & Business Media.
- [Saha et al., 2011] Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., Behringer, D., Hou, Y.-T., ya Chuang, H., Iredell, M., Ek, M., Meng, J., Yang, R., Mendez, M. P., van den Dool, H., Zhang, Q., Wang, W., Chen, M., and Becker, E. (2011). Ncep climate forecast system version 2 (cfsv2) selected hourly time-series products.
- [Sivaramanan, 2014] Sivaramanan, S. (2014). Impacts of el niño and la niña on the environment.
- [Tsai and Lee, 1999] Tsai, C.-P. and Lee, T.-L. (1999). Back-propagation neural network in tidal-level forecasting. *Journal of Waterway, Port, Coastal, and Ocean Engineering*, 125(4):195–202.
- [Wagenmakers et al., 2006] Wagenmakers, E.-J., P., G., and M., S. (2006). Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology*, page 149–166.

# Índice de figuras

|                                                                                                       |    |
|-------------------------------------------------------------------------------------------------------|----|
| 1.1. Costanera Norte - Nov, 2014 . . . . .                                                            | 6  |
| 1.2. Costanera Norte - Jul, 2014 . . . . .                                                            | 7  |
| 2.1. Ubicación del Río de la Plata respecto a América de Sur y referencia de<br>coordenadas . . . . . | 10 |
| 2.2. Índice Oceánico Niño (ONI). Eventos de El Niño y La Niña entre 2014 - 2017                       | 11 |
| 2.3. Conjuntos de datos y organismos de procedencia . . . . .                                         | 12 |
| 2.4. Referencias geográficas del dominio . . . . .                                                    | 13 |
| 2.5. Alturas horarias Buenos Aires - 2014-2017 . . . . .                                              | 14 |
| 2.6. Alturas horarias Buenos Aires - 2017 . . . . .                                                   | 14 |
| 2.7. Alturas horarias Buenos Aires - Enero 2017 . . . . .                                             | 15 |
| 2.8. Pronósticos de marea Dársena F - Buenos Aires, Enero 2014 . . . . .                              | 16 |
| 2.9. Proceso iterativo de selección de modelo lineal . . . . .                                        | 22 |
| 2.10. Redes neuronales recurrentes - ( <i>Cortesía de imgur</i> ) . . . . .                           | 24 |
| 2.11. Bloque LSTM (Long Short-Term Memory) . . . . .                                                  | 25 |
| 2.12. Ejemplo de CV con retención . . . . .                                                           | 26 |
| 2.13. Validación cruzada anidada de origen móvil con recalibración . . . . .                          | 27 |
| 2.14. Cálculo del error de predicción acumulado (APE) . . . . .                                       | 30 |
| 3.1. Tabla mareas Buenos Aires - Enero 2014 . . . . .                                                 | 33 |
| 3.2. Alturas horarias y tabla mareas Buenos Aires - Enero 2014 . . . . .                              | 33 |
| 3.3. Valores extremos de alturas horarias Buenos Aires - Noviembre 2014 . . . .                       | 33 |
| 3.4. Test de Dickey-Fuller (Bs As 2014-2017) . . . . .                                                | 34 |
| 3.5. Autocorrelación de alturas horarias Buenos Aires (2014-2017) . . . . .                           | 35 |

|                                                                                                                                             |    |
|---------------------------------------------------------------------------------------------------------------------------------------------|----|
| 3.6. Boxplot de serie original y transformada de alturas Buenos Aires (2011-2018)                                                           | 36 |
| 3.7. Alturas horarias y media móvil Buenos Aires . . . . .                                                                                  | 36 |
| 3.8. Boxplot de tablas de marea (AR y UR) . . . . .                                                                                         | 37 |
| 3.9. Comparación de la amplitud de mareas en el comienzo (a) y la salida (b)<br>del Río de la Plata - Enero 2016 . . . . .                  | 38 |
| 3.10. Tablas de marea entre Oyarvide y La Plata - Julio 2016 . . . . .                                                                      | 38 |
| 3.11. Intensidad y dirección mensual de vientos - Buenos Aires, 2014-2017 . . . .                                                           | 39 |
| 3.12. Correlación de vientos y pronósticos CFS Bs As y Mar del Plata, 2014-2017                                                             | 41 |
| 3.13. Generación variable sobre pronóstico CFS - Mar del Plata, 2014-2017 . . . .                                                           | 42 |
| 4.1. Observación y predicciones a 6h - Buenos Aires, 1ª semana Enero 2018 . . .                                                             | 44 |
| 4.2. Observación y pronósticos a 1h, 3hs, 6hs y 12hs - Buenos Aires, 1ª semana<br>Enero 2018 . . . . .                                      | 45 |
| 4.3. PACF y ACF de nivel de alturas Buenos Aires (2014-2017) . . . . .                                                                      | 47 |
| 4.4. PACF y ACF de la serie diferenciada a 24hs . . . . .                                                                                   | 47 |
| 4.5. Salida SARIMA (3,0,12)x(2,1,1) <sub>24</sub> . . . . .                                                                                 | 50 |
| 4.6. SARIMA (3,0,12)x(2,1,1) <sub>24</sub> - Residuos vs valores ajustados . . . . .                                                        | 50 |
| 4.7. Correlación de residuos SARIMA(3,0,12)x(2,1,1) <sub>24</sub> vs. pronósticos de vientos                                                | 51 |
| 4.8. Salida SARIMA (3,0,12)x(2,1,1) <sub>24</sub> con exóg=Wind30mb_MVD_3_1 . . . . .                                                       | 52 |
| 4.9. Correlación de residuos SARIMAX(3,0,12)x(2,1,1) <sub>24</sub> con exóg. vs pronósti-<br>cos de marea . . . . .                         | 53 |
| 4.10. Salida SARIMA (3,0,12)x(2,1,1) <sub>24</sub> y exóg=Wind30mb_v_MVD_3_1, Tide_BUE_1-<br>1 . . . . .                                    | 54 |
| 4.11. Salida SARIMA (3,0,12)x(2,1,1) <sub>24</sub> y exóg=Wind30mb_v_MVD_3_1, Tide_BUE_1-<br>1 y Wind30mb_MDP_05_100 . . . . .              | 55 |
| 4.12. Correlación de residuos SARIMAX(3,0,12)x(2,1,1) <sub>24</sub> con 3 exóg. vs tabla<br>de mareas . . . . .                             | 56 |
| 4.13. Salida SARIMA (3,0,12)x(2,1,1) <sub>24</sub> y exóg=Wind30mb_v_MVD_3_1, Tide_BUE_1-<br>1, Wind30mb_MDP_05_100, Tide_LPL_2_0 . . . . . | 57 |
| 4.14. Configuración del modelo LSTM - Métrica rendimiento (MSE) . . . . .                                                                   | 59 |

|                                                                                                                                                                                                                                                                                                                                                                                                                      |    |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 5.1. Error medio (azul) y desvío estándar del error (verde) de los pronósticos de nivel del agua de los modelos de olas y de ondas de tormenta en Palermo. El nowcast oficial (SHN) o actualización del pronóstico presenta el menor error sistemático. Los cuadrados representan los pronósticos de plazo más corto del SHN, basados en observaciones de mareógrafos e independientes del sistema numérico. . . . . | 62 |
| 5.2. Valores de las series durante eventos de niveles de agua extremos en el Río de la Plata . . . . .                                                                                                                                                                                                                                                                                                               | 63 |
| 5.3. Sudestadas Palermo, 2015-2016 . . . . .                                                                                                                                                                                                                                                                                                                                                                         | 63 |
| 5.4. Bajantes Palermo, 2014-2017 . . . . .                                                                                                                                                                                                                                                                                                                                                                           | 64 |
| 6.1. Error de Predicción Acumulado (APE) 2014-2017 . . . . .                                                                                                                                                                                                                                                                                                                                                         | 66 |

# Índice de cuadros

|                                                                                                                                                 |    |
|-------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 3.1. Estadísticos de tablas de marea de los puertos referenciados . . . . .                                                                     | 37 |
| 3.2. Frecuencia mensual de la dirección de vientos - Buenos Aires, 2014-2017 . .                                                                | 38 |
| 4.1. Detalle de conjuntos de datos generados . . . . .                                                                                          | 43 |
| 4.2. RMSE, coeficiente de determinación y AIC de regresiones lineales (predicciones a 6h) . . . . .                                             | 44 |
| 4.3. RMSE y coeficiente de determinación (1h, 3hs, 6hs, 12hs) . . . . .                                                                         | 44 |
| 4.4. Búsqueda hiperparámetros y comparación de modelos SARIMA . . . . .                                                                         | 49 |
| 4.5. Criterio de información y RMSE de modelos . . . . .                                                                                        | 57 |
| 5.1. RMSE de pronósticos entre 1h y 12hs (en mts) . . . . .                                                                                     | 60 |
| 5.2. Error medio de pronóstico oficial, Eta SMN, regresión lineal, SARIMAX y LSTM (período 2 años) . . . . .                                    | 62 |
| 5.3. Nivel del agua observado en Palermo y error medio de pronósticos a 6hs y 12hs durante sudestadas y bajantes (Palermo, 2015-2017) . . . . . | 63 |
| 5.4. APE calculado en predicciones a 6hs y 12hs entre 2014 y 2017 . . . . .                                                                     | 64 |