ESPECIALIZACIÓN EN CIENCIA DE DATOS - ITBA

Análisis de Series de Tiempo

PRONÓSTICO DE DEMANDA DE USO DE AEROPUERTOS EN ARGENTINA AL 2022

Trabajo Final Integrador
Tutora: María Juliana Gambini

José Ignacio López Sáez
29-8-2018

Tabla de contenido

Análisis de Series de Tiempo

PRONÓSTICO DE DEMANDA DE USO DE AEROPUERTOS EN ARGENTINA AL 2022

Resumen

En la mayoría de los negocios, se desea ser capaz de estimar la demanda futura de un producto o servicio dado. El análisis sobre series temporales permite utilizar la información histórica para ofrecer un número aproximado de dicho valor, dentro de un rango de probabilidades determinado.

Este estudio surge a partir de la necesidad que tiene el Ministerio de Transporte de la Nación de conocer el número de pasajeros que utilizarán cada aeropuerto del país en el futuro para poder asignar de una manera más eficiente los recursos disponibles y orientar inversiones.

En este estudio se han evaluado en total 4 modelos de proyección de series de tiempo: suavizamiento exponencial (Holt-Winters), ARIMA, Prophet (desarrollado por Facebook) y redes neuronales (procedimientos de aprendizaje automático o *machine learning*), para las cuales se probaron tres implementaciones distintas en R. Se obtiene así el resultado de proyección de pasajeros domésticos e internacionales para cada una de estas seis implementaciones y para todos los aeropuertos de la Argentina para un horizonte de 5 años (60 meses a partir del último disponible), entregando también el error de ajuste de cada uno de los modelos.

Los resultados obtenidos para el año 2022 sobre el total país alcanzan un valor de 17,8 millones de pasajeros en vuelos de cabotaje, lo que constituye un incremento del 37% contra el año 2017, cuando viajaron 13,0 millones de personas. En vuelos regionales e internacionales, el número al 2022 es de 18,6 millones de pasajeros, que implica una mejora del 27% contra los 14,7 millones del año 2017. Estos valores no alcanzan las cifras previstas en las hipótesis: 19,5 y 21,0 millones, respectivamente.

En vista de los resultados, se podrán evaluar y llevar a cabo distintas políticas aerocomerciales que procuren incrementos mayores en el corto plazo. Si bien no se podrá modelar esto por adelantado, con cada nuevo mes de información, todos los resultados se ajustarán y el algoritmo entregará nuevos valores finales acordes.

Palabras clave

Modelo, Proyección, Series de Tiempo, Tendencia, Estacionalidad, Ciclo.

Introducción

El Gobierno Nacional, a través del Ministerio de Transporte de la Nación, avanza en lo que se denominó "La Revolución de los Aviones", plan que engloba una serie de acciones cuyo objetivo principal es la duplicación de pasajeros de cabotaje, haciendo que el volar sea accesible para todos y con servicios disponibles a lo ancho y largo del país. Esto lo sostiene en tres pilares: el crecimiento sostenido de la aerolínea de bandera (Aerolíneas Argentinas), el ingreso de nuevas empresas y la modernización de la infraestructura y rediseño del espacio aéreo.

Argentina es el octavo país del mundo en cuanto a su extensión territorial. Sin embargo, aún se encuentra muy por debajo del resto de los países de la región en cuanto a la cantidad de personas que utilizan el avión como medio de transporte. Por ejemplo, en Perú hay 35 personas por cada 100 habitantes que vuelan; en Brasil, 44; en Colombia, 51 y en Chile, 63. En Argentina, en cambio, el número no supera los 24 pasajeros por cada 100 habitantes. Por otro lado, un estudio realizado sobre las distintas regiones del país permitió estimar que existe un aproximado de 12 millones de personas que podrían remplazar su medio de movilidad actual (auto, micro, tren) por el avión en el mediano plazo. Todo ello pone en evidencia la potencialidad del sector aéreo comercial en el país.

A modo de referencia, en el año 2017 hubo en Argentina un total de 27,8 millones de pasajeros repartidos en poco más de 283 mil vuelos. El aeropuerto más utilizado fue el Aeroparque Jorge Newbery, por el cual circularon 12,8 millones de pasajeros (arribados y despegados) y se efectuaron 135 mil movimientos (despegues y aterrizajes). Esto último pone de manifiesto que la situación es muy desigual en las distintas provincias y el estado de los aeropuertos de las mismas, dado que el 60% de los pasajeros despegan desde el Aeroparque Jorge Newbery o el Aeropuerto Internacional Ministro Pistarini y el restante 40% se reparte entre más de 38 aeropuertos. A la vez, el equipamiento tecnológico de los mismos es obsoleto en muchos casos y requieren inversiones urgentes para poder seguir operando servicios de aviación comerciales en sus terminales. Es decir, si bien muchas regiones tienen mucho potencial de crecimiento, éste será imposible en la medida que no se acompañe dicha evolución con inversiones acordes.

En función de ello, el Ministerio de Transporte de la Nación y la Empresa Argentina de Navegación Aérea (EANA S.E.), necesitan conocer un estimado de crecimiento de cada terminal en el futuro. De esta manera se podrá priorizar las inversiones necesarias en función de un criterio unificado (número de pasajeros, por ejemplo). EANA es la empresa encargada de llevar adelante todas las erogaciones correspondientes al equipamiento en comunicaciones de las torres de control, centros de control de área, servicios de asistencia a la navegación aérea, entre otras. Además, es quien debe disponer del

personal para ocupar los puestos de control en las torres y centros regionales. Esto último siempre resulta proporcional al número de movimientos (aterrizajes y despegues) que maneja cada aeropuerto y zona del país.

Para poder acercar una solución a esto, se estudian las series históricas de pasajeros por aeropuerto (disponibles desde el año 2001 a la actualidad), y se determinan los parámetros que caracterizan a cada una de ellas, así como al total del país.

Los métodos de pronóstico y suavizamiento exponencial permiten un primer acercamiento al asunto, ofreciendo una solución relativamente rápida. Por otro lado, los métodos de análisis de correlaciones y ARIMA, hacen posible también la utilización de los patrones subyacentes, pero utilizando distintas funciones, que hacen que su automatización sea un poco más compleja. Por último, las técnicas de aprendizaje automático (*machine learning*) pueden ser aplicadas para el estudio de series históricas y la predicción de las variables de estudio para los períodos futuros, sea en un paso o realizando predicciones recursivas de largo plazo utilizando los propios valores estimados para las proyecciones.

Marco teórico

Serie de tiempo

Definición y componentes

Una serie de tiempo es una forma estructurada de presentar los datos, en donde un registro de fecha/hora lleva asociado un valor. Es decir, es una secuencia de observaciones sobre intervalos de tiempo regulares.

En el presente estudio, con la base de datos sin procesar, se tiene un movimiento (despegue o aterrizaje) y un valor de pasajeros y carga (en kilogramos) asociado a la fecha y hora en que se haya registrado dicha operación.

Actualmente, el volumen de datos mensual está en el orden de los 45 a 50 mil registros, lo que lleva que la base de datos con varios años de historia sea difícil de manejar correctamente. De todas maneras, según el análisis, se pueden consolidar los datos sobre una base mensual para cada aeropuerto y tipo de operación, disminuyendo significativamente el volumen final del *dataset*. De todas maneras, con las mismas técnicas uno podría trabajar al nivel de detalle que se requiera: puede existir un negocio donde los patrones de demanda presenten cierta tendencia y ciclo a nivel diario u horario y en esos casos también se podrá realizar la proyección correspondiente.

Una serie temporal se puede caracterizar de acuerdo a sus componentes:

1) Tendencia: es la componente de largo plazo que determina la base de crecimiento (o decrecimiento) de la serie. Si la serie es estacionaria, su media y varianza son invariantes.
2) Estacionalidad: es el comportamiento de una serie dentro de un período dado. Las series temporales pueden formar patrones que se repiten de un período al siguiente.
3) Ciclos: son desviaciones de la tendencia subyacente debido a distintos factores (generalmente externos), diferentes de la estacionalidad. El tiempo y duración de los ciclos no necesariamente es regular.
4) Aleatoriedad: fluctuaciones impredecibles o no periódicas que subyacen en la serie.

Los pasajeros en los distintos aeropuertos suelen tener un comportamiento estacional que se repite año a año. Por ejemplo, aquellos en centros turísticos en la costa (Aeropuerto de Mar del Plata, por ejemplo), tienen un fuerte afluente de pasajeros en los meses de verano y una baja demanda el resto del año.

Con el correr de los años, gracias a una mayor oferta de asientos y baja de precios (sobre todo, precios relativos contra otros medios de transporte, compitiendo principalmente contra los micros de larga distancia), el sector está en constante crecimiento, mostrando una tendencia de fondo creciente en la mayoría de los aeropuertos. Por supuesto, el sector aerocomercial no será ajeno a los vaivenes económicos y políticos del país, y sufrirán o se beneficiarán de acuerdo a en qué momento se encuentre la Argentina en su ciclo económico. En efecto, es un sector muy sensible a algunas variables, tales como el tipo de cambio, que hará que pueda existir sustitución de pasajeros en vuelos de cabotaje por internacionales o regionales y viceversa. Por último, siempre hay observaciones que no se explican por la tendencia de fondo, la estacionalidad propia del aeropuerto o el ciclo. Ejemplo de ello son las cuestiones climáticas, paros de actividad o cierres por obras.

Pronóstico de series de tiempo

El pronóstico de la serie temporal implica extender los valores históricos hacia el futuro. Las dos variables que lo definen son: el período, dado por el nivel de agregación (días, horas, meses, etc.), y el horizonte, dado por la cantidad de períodos a proyectar.

Métodos de series de tiempo

Los métodos de análisis de las series tienen en consideración que los datos históricos pueden estar correlacionados, así como la tendencia subyacente y la estacionalidad propia que la define.

Los métodos se pueden clasificar según:

1) Métodos de pronóstico y suavizamiento simple: de acuerdo a los patrones existentes (dados por la tendencia, estacionalidad y ciclo) se deberá elegir el método adecuado dependiendo de si los mismos son estáticos (es decir, se observa siempre el mismo comportamiento) o dinámicos (éstos van cambiando con el tiempo). A partir de allí, se realiza la extrapolación hacia el futuro.
2) Métodos de análisis de correlaciones y modelo ARIMA (promedio móvil integrado autorregresivo): también utilizan los patrones subyacentes en la serie, pero utilizando funciones de diferenciación, autocorrelación y autocorrelación parcial. Si bien pueden ser más flexibles a los datos que en el caso de los métodos simples, su automatización no es sencilla.

A estos modelos se le suman las técnicas de aprendizaje automático (*machine learning*), que refieren a sistemas que aprenden automáticamente. Es decir, que identifican patrones complejos entre los datos y en función de ellos, son capaces de predecir el comportamiento futuro. El término "automático", implica que estos modelos continuarán aprendiendo en forma autónoma con el tiempo. Concretamente, los pesos relativos de los factores (el modelo resulta equivalente a una regresión no lineal), irán ajustándose con el tiempo gracias a los nuevos datos y valores de salida con los cuales el modelo se reentrena.

Existen dos tipos de pronósticos:

1) Predicción de un paso: utiliza la información del pasado y pronostica el valor siguiente.
2) Predicción recursiva de largo plazo: requerirá utilizar valores calculados por el propio sistema para la estimación.

En este trabajo de investigación, se busca hacer una proyección a un horizonte de 60 meses y, por lo tanto, nos encontramos en el segundo caso. El problema que puede existir en tal situación usando una predicción recursiva es que, justamente porque utiliza otras estimaciones, el error se puede propagar rápidamente y el intervalo de confianza será mayor cuanto más lejos se encuentre el período a proyectar con respecto al último dato histórico disponible.

Descripción del Contexto

EANA es una sociedad del estado existente bajo la órbita del Ministerio de Transporte y fue creada para efectivizar el traspaso de los servicios de navegación aérea al ámbito civil. La empresa tiene bajo su responsabilidad y control un total de 56 aeropuertos, agrupados en 5 centros regionales de control de área (ACC): Ezeiza, Mendoza, Córdoba, Comodoro Rivadavia y Resistencia.

En línea con el Plan Integral del Estado implementado a través del Ministerio de Transporte, cuyo objetivo es hacer más eficientes las operaciones aéreas e impulsar el desarrollo del sector y de las distintas regiones del país, se comenzó a trabajar en estadísticas de manera de poder contar con la información necesaria y adecuada para la toma de decisiones y definición de inversiones.

El vacío estadístico

Durante muchos años, en Argentina no se han confeccionado ni publicado estadísticas relacionadas al sector aéreo de ningún tipo: pasajeros, movimientos, etc. Lo único que había eran datos que publicaban algunos organismos en forma esporádica y estrictamente relacionadas con su negocio particular.

Por ejemplo, hasta el año 2008, Aeropuertos Argentina 2000 (concesionaria de más de 30 aeropuertos en la Argentina) publicaba en su sitio varias planillas de cálculo con información estadística de cada uno de sus aeropuertos, pero luego no sólo se discontinuó dicha práctica, sino que se eliminó lo publicado. Por su parte, el ORSNA (Organismo Regulador del Sistema Nacional de Aeropuertos) comenzó a publicar desde 2014 información estadística con datos desde el 2001 con informes anuales. Sin embargo, como su actividad se centra en los aeropuertos, existe doble conteo de pasajeros (un pasajero despegado desde A y arribado a B, se contabiliza tanto en A como en B), lo cual lo hace sus números pocos representativos de la actividad en su conjunto. Además, la nueva información de cada año demora varios meses en ser publicada. Por último, el INDEC divulgaba, en forma trimestral, algunos datos de pasajeros (con información de Migraciones), movimientos y carga transportada, pero se discontinuó la actualización de la información en octubre 2015.

Sistema Integrado de Aviación Civil

La base de datos “Sistema Integrado de Aviación Civil” (SIAC) es, actualmente, la fuente primaria de información. La carga de datos es manual por parte de operadores EANA en cada aeródromo bajo responsabilidad de la empresa (56 en el país), pero es administrado por la autoridad aeronáutica (Administración Nacional de Aviación Civil, ANAC). En consecuencia, sólo se cuenta con un permiso de acceso limitado para poder extraer información de movimientos por aeródromo. El archivo descargado (uno o varios aeródromos para un rango de fechas determinado), tiene formato .txt donde cada aeródromo registra un movimiento (salida o arribo) y se genera un nuevo registro. Si se trata de un vuelo de cabotaje, habrá entonces 2 registros que identifiquen ambos movimientos en forma individual; si es internacional, sólo 1. No se registran sobrevuelos.

Por otro lado, un análisis minucioso del transporte argentino, realizado con metodología de aplicación internacional, mostró que existen 12 millones de pasajeros potenciales, en vuelos de cabotaje únicamente, considerando la eventual sustitución de medios de transporte alternativos al avión, como

ser los servicios de micros interurbanos, autos particulares y, en menor medida, el tren. En consecuencia, para poder acercar a dichas personas a la posibilidad de efectuar sus viajes por vía aérea, se avanza en la modernización de la infraestructura de los servicios que dan apoyo a los vuelos.

Definición del problema

Personal de EANA en cada aeropuerto registra los aterrizajes y despegues que se llevan a cabo, detallando pasajeros transportados, hora de salida/llegada, origen y destino, carga transportada, aerolínea u operador. Con estos datos se construye una base de datos, consolidando la información de cada uno de los aeropuertos, con una actualización mensual.

Al momento no existe un modelo que permita, en función de esta información histórica, generar un pronóstico de mediano/largo plazo (al menos 5 años) para cada aeropuerto.

En el año 2016 se contrató una firma de consultoría que entregó un modelo utilizando software IBM SPSS. Sin embargo, esto no significó una solución definitiva, debido a que sólo es capaz de estimar un trimestre hacia adelante (y no desglosado mensualmente) y por grupos de aeropuertos. Por otro lado, sólo se realizó para vuelos de cabotaje. Asimismo, se intentó en varias oportunidades ofrecer un número estimativo de la demanda futura por aeropuerto en función de la capacidad ofrecida (oferta). Se entiende que el simple dato del número de asientos que las aerolíneas tienen planificado ofrecer no es capaz de explicar la totalidad de la demanda para cada aeropuerto o región.

En conclusión, al momento no se ha encontrado una forma práctica y sistemática de explotar la base de datos para armar un modelo predictivo de pasajeros futuros.

Justificación

Tanto EANA como el Ministerio de Transporte tienen la necesidad de estimar el número de pasajeros (mediano/largo plazo) para cada aeropuerto.

En el caso de EANA, servirá para poder mejorar la planificación de las inversiones a realizar en infraestructura de comunicaciones, así como en las operaciones diarias, dado que el trabajo de los controladores está directamente relacionado con la cantidad de aeronaves que atraviesan el espacio aéreo argentino.

En el caso del Ministerio de Transporte, servirá para poder orientar las inversiones a realizar en infraestructura tecnológica y edilicia de los distintos aeropuertos del país, coordinando así el trabajo con las distintas direcciones que de él dependen, así como para planificar y llevar adelante las distintas políticas comerciales que permitan desarrollar el sector.

En ambos casos, lo que se busca es una asignación más eficiente de los recursos financieros. Los datos a proveer, entonces, serán necesarios para facilitar la toma de decisiones y evitar que las mismas no cuenten con datos de respaldo. Dado que las distintas provincias, buscarán orientar las inversiones hacia sus propios aeropuertos, es un tema de suma importancia para el Ministerio de Transporte el poder contar con los números apropiados para justificar sus asignaciones.

Alcances

1. El presente estudio analizará los datos históricos recolectados por el personal de EANA S.E. en cada uno de los aeropuertos bajo su jurisdicción y que se encuentran volcados en el Sistema Integrado de Aviación Civil.
2. La finalidad del estudio es proporcionar una herramienta que le permita a EANA S.E. pronosticar los pasajeros futuros (horizonte de 5 años) para cada uno de los aeropuertos.
3. Otro usuario de la herramienta será el Ministerio de Transporte de la Nación, que la podrá utilizar para una mejor asignación de los recursos.

Limitaciones

1. La carga de datos manual hace que no se puedan considerar consolidados los registros con una antigüedad menor a los 3 meses: muchos operadores demoran la publicación de la información en el sistema y lo hacen a mes vencido. Es por ello que cada mes se vuelve a descargar la información del último trimestre, sobrescribiendo los registros.
2. Quedan fuera del alcance de este estudio el análisis de la carga transportada.
3. Se omite el modelado de aquellos aeropuertos que presenten menos de 1.000 pasajeros en total dentro del set de entrenamiento de los modelos (70% de los registros), en forma separada para vuelos de cabotaje e internacionales. Es decir, los datos de un aeropuerto con pasajeros de cabotaje, pero con escasos (o nulos) internacionales, sólo se modelan los valores de los vuelos domésticos.
4. Se consideran series de tiempo univariadas: el único elemento a modelar es el número de pasajeros en función de la fecha. No se incorporan otros regresores en los modelos. Se entiende que los diversos efectos (cierre de un aeropuerto, sustitución de pasajeros de cabotaje e internacionales entre sí, nuevas frecuencias de alguna aerolínea en algún destino, etc.) se irán incorporando a medida que la nueva información disponible mes a mes así lo haga. Lo mismo es válido para la información de los años pasados: no se incorpora ningún tipo de dato exógeno.
5. Sólo se podrá remplazar el valor de pasajeros en los datos de entrada por “NA” (*Not Available*) si se quisiera omitir el modelado de dicho período. Esto permitirá evitar que los distintos

métodos incorporen en su entrenamiento información que no es relevante o representativa de la actividad de tal aeropuerto. Por ejemplo, meses en los que la terminal permanece cerrada por obras.

Hipótesis

- En pasajeros de cabotaje, se alcanzará un nivel cercano a los 19,5 millones de pasajeros al 2022, valor un cincuenta por ciento superior al de 2017.
- En pasajeros regionales/internacionales, se alcanzará un nivel cercano a los 21 millones de pasajeros al 2022, cifra un cuarenta por ciento superior a la del 2017.

Objetivos

Objetivo general

Entregar a EANA S.E. un modelo informático que tome la información histórica de pasajeros para cada aeropuerto hasta el último mes cerrado y sea capaz de realizar una estimación de esta variable a un horizonte de 5 años con desglose mensual.

Objetivos específicos

- Analizar métodos de suavizado exponencial y ARIMA.
- Analizar técnicas de aprendizaje automático: implementación de redes neuronales.
- Consolidar la fuente de datos actual para cada aeropuerto y agregarla a nivel mensual.
- Generar un algoritmo genérico que itere sobre los distintos aeropuertos, tanto en sus valores de pasajeros domésticos como internacionales y, para cada modelo:
    - Sobre un subgrupo de entrenamiento, determinar los parámetros que describan cada serie.
    - Aplicar el modelo al subgrupo de prueba.
    - Entrenar el modelo sobre todo el set de datos y realizar la proyección con un horizonte de 5 años.
- Entregar los resultados de todos los modelos indicando el error de ajuste en cada caso para cada aeropuerto.
- Entregar un tablero de control (*dashboard*) donde se puedan observar las proyecciones de cada modelo por aeropuerto y los porcentajes de crecimiento resultantes.

Análisis de los modelos

Holt-Winters

Los métodos de suavizamiento exponencial surgen como una evolución de aquellos de promedio móvil simple y ponderado. En ellos, el valor pronóstico para una determinada variable está dado por el promedio del valor de los períodos anteriores (más precisamente, de una cantidad de datos que se debe definir y no de la serie completa). La diferencia es que, cuando es ponderado, se les asigna un peso relativo mayor a aquellos temporalmente más próximos al momento de la evaluación.

Esta técnica tiene en cuenta el error de pronóstico actual en los siguientes. Esto hace que no necesite un gran volumen de datos históricos de la variable a pronosticar: para un período *n*, sólo requiere el valor real y pronosticado para el instante *n-1*, así como la constante de suavización:

$$Pronóstico_n = Pronóstico_{n-1} + \alpha \times (Valor\ Real_{n-1} - Pronóstico_{n-1})$$

La constante de suavización (α), con un valor comprendido entre 0 y 1, es el factor de ponderación: ante la necesidad de darle un mayor peso relativo a los valores más recientes, éste es más elevado.

Con todo ello, el suavizamiento exponencial simple resulta muy efectivo para series temporales de demanda en donde exista una tendencia (aunque ésta sea local en un período) y un patrón estacional constante, donde con un peso mayor sobre los últimos períodos, se evita el efecto de los elementos irregulares del pasado.

Al igual que los métodos de promedio móvil, la desventaja que tienen estos métodos es la respuesta a la tendencia: cambios importantes de la demanda de un período al siguiente hacen más grande el error de pronóstico.

En estos casos, se puede recurrir a un suavizado exponencial doble (o “Método de Holt”), en donde se agrega una segunda constante (generalmente denominada *β*), cuya función es la de reducir este error. Es decir, a un valor de constante *β* más elevado (también acotado entre 0 y 1), el índice de respuesta frente a los cambios de tendencia es mayor. Explicado de otra manera, el suavizamiento exponencial doble consiste en la conjunción de dos métodos simples: uno para pronosticar el nivel del valor de período *n+1* y otro para la tendencia.

El modelo Holt-Winters (o suavizamiento exponencial triple) engloba varios procedimientos que conforman la base de las series temporales de suavizamiento exponencial y puede adaptarse en forma sencilla a cambios, tendencias y patrones estacionales.

En el método de suavizamiento exponencial doble existen dos componentes: nivel y tendencia. En el caso del modelo Holt-Winters, se incorpora un tercer elemento: la estacionalidad. Dicho esto, el mismo no funciona correctamente para el pronóstico de series de tiempo que no presenten patrones de variación más o menos regulares a lo largo de los meses del año. La estacionalidad, a la vez, se explica en dos componentes: su largo, o número de períodos que dura el mismo patrón que se repite; y su componente estacional, es decir, para un período, cuánto se desvía positiva o negativamente, del valor medio.

En síntesis, los componentes de los distintos modelos exponenciales son los siguientes:

1) Exponencial simple: nivel
2) Exponencial doble: nivel y tendencia
3) Exponencial triple: nivel, tendencia y estacionalidad

Una ventaja importante de los modelos de suavizamiento y, en particular, de Holt Winters, es que son univariados. Es decir, que implican una única variable exploratoria o, en otras palabras, no necesitan más información que la serie temporal, prescindiendo de la utilización de variables de regresión exógenas.

Estacionalidad

Aplicado en nuestro caso de estudio, los vuelos y pasajeros que arriban o despegan de un aeropuerto suelen estar fuertemente correlacionados con la época de año. Más aún en aquellos que sirven a destinos turísticos, tales como los aeropuertos del sur del país, donde hay un marcado crecimiento en los meses estivales y otro pico en las semanas de vacaciones de invierno y temporada de ski. A modo de ejemplo, el gráfico 1 muestra la cantidad de pasajeros (en miles, sobre el eje derecho) que viajaron desde y hacia el aeropuerto de Bariloche durante el 2017 y los movimientos (en cantidad, sobre el eje izquierdo):

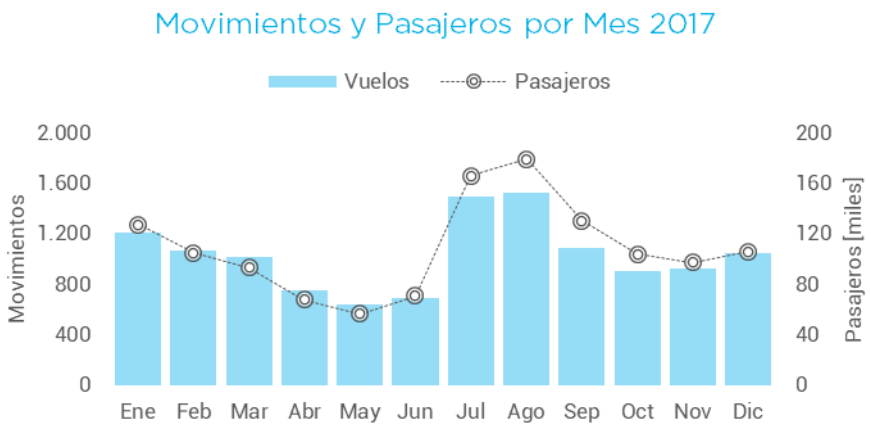


Gráfico 1. Movimientos y pasajeros por mes (2017) - Bariloche

Adicionalmente, en el gráfico 2 se ve la foto de los últimos cinco años completos, y se observa el mismo comportamiento marcadamente estacional:

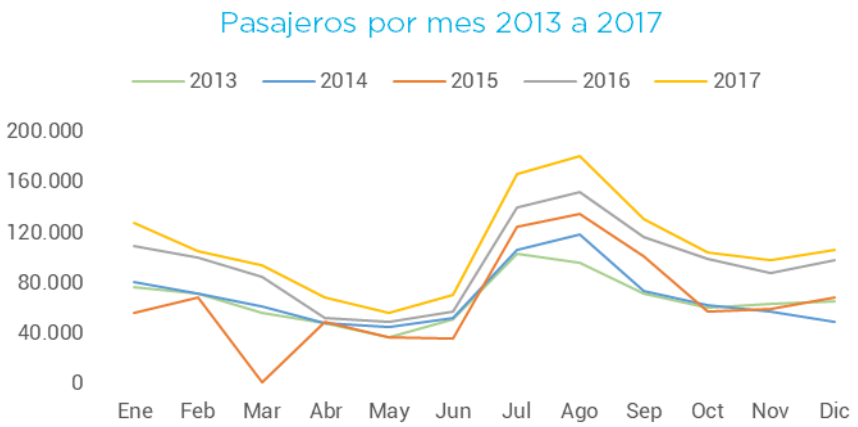


Gráfico 2. Pasajeros por mes 2013 a 2017 - Bariloche

Por lo tanto, algo conocido de antemano **al trabajar con las series temporales de pasajeros por aeropuerto es que la longitud de los períodos es de 12 meses**.

Tendencia

Por otro lado, el gráfico 3 permite observar estos 5 años en forma continua, dejando en evidencia la tendencia de fondo, representada con una línea punteada:

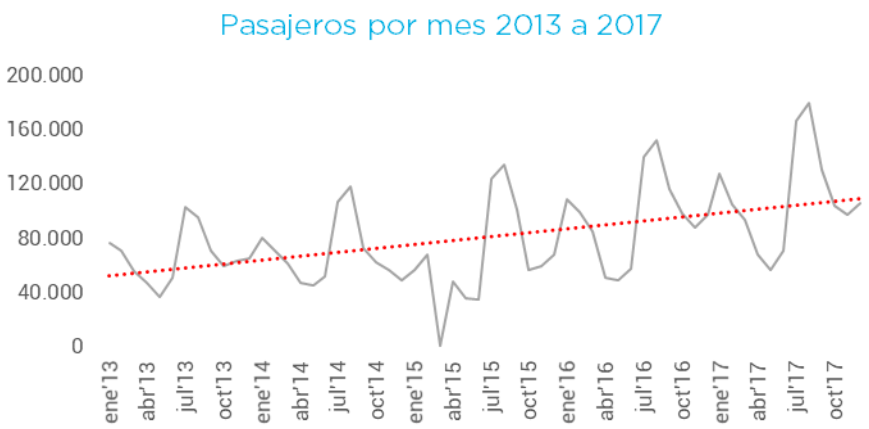


Gráfico 3. Pasajeros por mes 2013 a 2017 – Bariloche.

En este caso, una primera aproximación rápida a la tendencia se puede hacer con una estimación lineal, que refleja un crecimiento continuo a lo largo de los años para este aeropuerto. La pendiente de la tendencia en un período dado depende del incremento (o decremento) del número de pasajeros a lo largo de los meses y se debe, generalmente, a cuestiones exógenas tales como la instalación de una nueva aerolínea, mayores frecuencias de las existentes, apertura de nuevas rutas desde y hacia este aeropuerto, campañas turísticas, promociones particulares que pueden impulsar fuertemente un período dado, etc. Por supuesto, el crecimiento demográfico y la situación económica, tanto nacional como propia de la región, también están vinculados, dado que, por ejemplo, un determinado valor del tipo de cambio hace que los argentinos hagan más turismo interno que viajes al exterior, o viceversa, y de igual manera para los visitantes extranjeros.

Nivel

El volumen de pasajeros de cada aeropuerto depende de muchas variables: las opciones turísticas de la región, la cantidad de aeropuertos cercanos, la población del lugar, la cantidad de destinos que se sirven y si son únicamente de cabotaje o hay rutas internacionales, la actividad económica de la zona (actividades como la industria del petróleo, por ejemplo, mueven mucha gente en forma periódica), la cercanía a los núcleos urbanos más importantes del país (por ejemplo, Rosario tiene relativamente pocos pasajeros de cabotaje, que se explica, en parte, por su cercanía con las ciudades de Córdoba y Buenos Aires), entre otros. En el gráfico 4 observamos los niveles correspondientes a cada aeropuerto (los que tienen vuelos comerciales regulares) como un promedio de todos los meses del 2017:

Gráfico 4. Pasajeros por aeropuerto - Promedio mensual 2017 (en miles)

Los aeropuertos de Buenos Aires predominan fuertemente por sobre el resto y son los que más influyen sobre las variaciones (crecimiento) que se ven a nivel país. La diferencia entre ellos es que, en Aeroparque, el mayor porcentaje de los vuelos son de cabotaje y en Ezeiza, internacionales. Córdoba se ubicaría en un segundo escalón y está experimentando un fuerte impulso a partir de la creación de un *hub* (centro de conexiones) a partir de abril del 2017. Luego Mendoza, Salta, Bariloche e Iguazú, se afianzan al ser grandes centros turísticos a nivel local.

La situación es bastante heterogénea a lo largo del país debido a que no todos los aeropuertos cuentan con instalaciones propias adecuadas para recibir cualquier tipo de aeronave comercial (sea por tamaño de pista, tecnología para aterrizar y despegar independientemente de las condiciones climáticas, tamaño de la terminal, disponibilidad de mangas, etc.).

Un argumento que se puede esgrimir es que, en aeropuertos de menor envergadura u operados por una única línea aérea, la demanda está fuertemente condicionada a la oferta de asientos existentes. Por ejemplo, el aeropuerto de El Calafate muestra un comportamiento marcadamente estacional y sólo es operado (al presente) por Aerolíneas Argentinas. Sin embargo, es claro que si hubiera una fuerte demanda de asientos desde/hacia este destino, por ejemplo, la oferta respondería siempre y cuando resulte un destino rentable. Si este fuera el caso y la aerolínea actual no lo hiciera, es probable que otra buscaría atender esa demanda remanente.

En conclusión, en este trabajo se mantiene la hipótesis de que, al menos en el mediano y largo plazo, la oferta se acomoda a la demanda en precios y asientos disponibles.

Adicionalmente, un comentario que se puede realizar y que aplica transversalmente a todos los modelos previstos, es si hay que definir un máximo de capacidad de pasajeros para un aeropuerto determinado. Es cierto que todos tienen un nivel de saturación dado y no pueden albergar más vuelos/pasajeros que

para lo que están diseñados, o no al menos en forma constante. Es decir, es probable que una terminal pueda manejar un pico (como consecuencia de demoras, cancelaciones, etc.) pero no que, en forma sostenida, el nivel de vuelos diarios se encuentre por encima de su capacidad máxima (limitado por tamaño de la terminal, capacidad de uso de la pista, posiciones de estacionamiento de aeronaves, cantidad de puertas, etc.). De todas maneras, a excepción de Aeroparque, ningún aeropuerto argentino en la actualidad se encuentra cerca de su punto de saturación y esto no se alterará en el mediano plazo. Por otro lado, justamente, las proyecciones sirven para tener una estimación del número de pasajeros futuros y así llevar adelante las obras que permitan albergarlos. **Por lo tanto, no se establece en este trabajo un máximo teórico de pasajeros mensuales por aeropuerto que limite los modelos de pronóstico.**

Valores Atípicos

En el gráfico 2, es evidentemente como hay algunos meses (uno o varios continuos) que pueden resultar atípicos y donde el volumen de movimientos y pasajeros es prácticamente nulo (en el ejemplo, marzo 2015). Esto se debe, generalmente, a que se llevan adelante obras de remodelación en el mismo. En contrapartida, otros aeropuertos suelen servir de alternativa y ven su afluencia incrementada. El gráfico 5 muestra los pasajeros por mes (en miles) del 2016 para los aeropuertos de Mendoza (donde se desarrollaron obras de remodelación entre septiembre y diciembre) y San Juan:

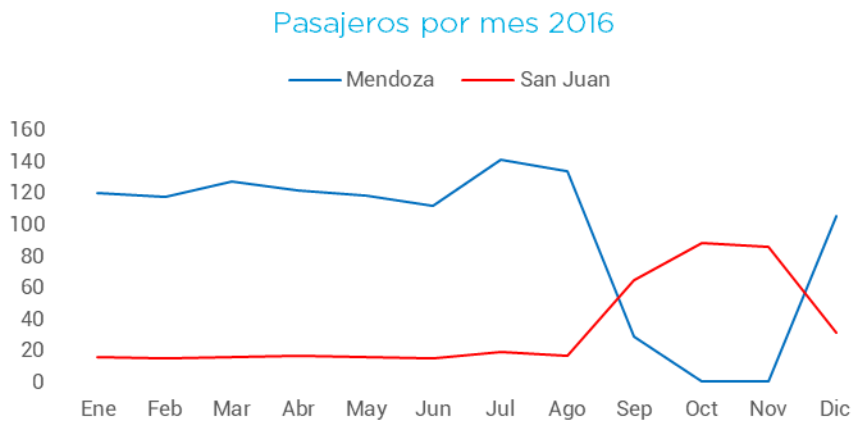


Gráfico 5. Pasajeros por Mes 2016 - Mendoza y San Juan (en miles)

Las obras en el aeropuerto tuvieron lugar entre el 7 de septiembre y el 7 de diciembre del 2016. Durante ese período, todos los pasajeros con destino final Mendoza, fueron derivados en vuelos hacia San Juan, principalmente, y San Rafael y San Luis, en menor medida. Como se puede apreciar en el gráfico, la demanda a lo largo del año para el aeropuerto sanjuanino es relativamente estable en torno a los 15 a 20 mil pasajeros al mes. Sin embargo, durante la obra en Mendoza, este número alcanzó los 88 mil. En Mendoza, en octubre y noviembre, el volumen fue de 0 pasajeros. El problema con este tipo de acontecimientos es que se trata de situaciones atípicas y los modelos, al tomar como fuente la

información histórica de pasajeros no deberían considerar los valores tal como están, dado que no son representativos de la operación y demanda genuina del aeropuerto.

Es por ello, que se puede tratar las series originales con un algoritmo de suavización: filtro Hampel. Éste consiste en una ventana móvil de longitud *2k+1* que recorre la serie temporal calculando, para cada punto, la mediana a lo largo de la ventana (*k* valores a cada lado del punto en cuestión). Adicionalmente, estima la desviación estándar del mismo contra la mediana de la ventana calculando la desviación absoluta mediana (MAD). Si esta diferencia (entre el valor y la mediana) es superior a 3 veces este desvío estándar, se remplaza con la mediana.

$$MAD = mediana(|X_i - mediana(X)|)$$

Donde X_i es el valor y X la serie de valores comprendidos a *k* posiciones a derecha e izquierda de X_i.

Luego, se estima el desvío estándar:

$$\sigma \approx 1{,}4826 \times MAD$$

Con lo cual, si $|X_i - mediana(X)| > 3\sigma$, se lo remplaza por la mediana para esa ventana.

Algunos ejemplos de aplicación sobre los datos (con k=3):

1) Pasajeros Aeroparque 2009 a 2017: el filtro elimina el fuerte pico negativo de Nov'10 (cierre del aeropuerto) y Jul'11 (cancelaciones de vuelos por presencia de cenizas volcánicas). Sin embargo, se pierden dos picos positivos en Ene'16 y Mar'17. El gráfico 6 muestra los valores de la serie original y la resultante de aplicar el filtro para Aeroparque entre el 2009 y el 2017:

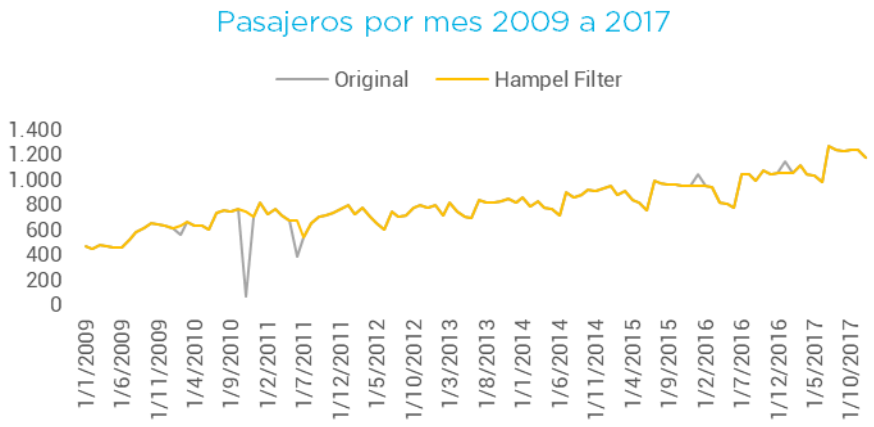


Gráfico 6. Pasajeros Aeroparque 2009 a 2017 - Comparación Original vs Filtro Hampel

2) Pasajeros Mendoza: con este valor de k no se logra solucionar los valores atípicos de fin del 2016. El gráfico 7 muestra ambas series correspondientes al aeropuerto de Mendoza entre el 2009 y 2017 para k=3:

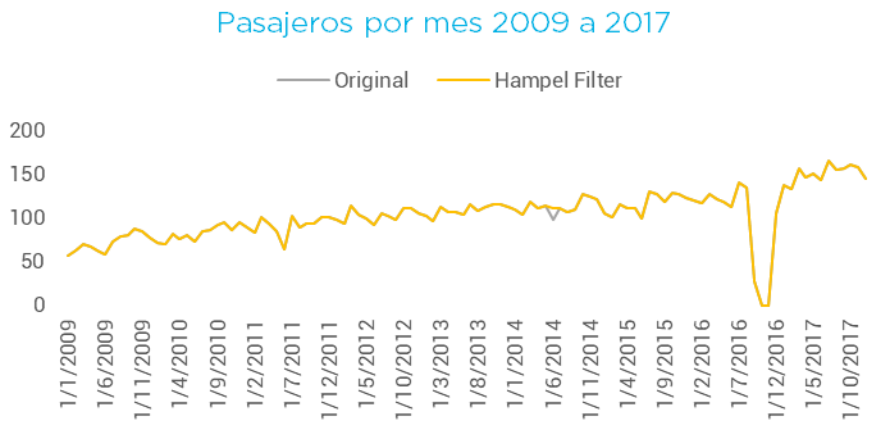


Gráfico 7. Pasajeros Mendoza 2009 a 2017 - Comparación Original vs Filtro Hampel con k=3

3) Puerto Madryn: logra filtrar 2 picos positivos muy fuertes que tuvo el aeropuerto en Mar'13 (cierre aeropuerto Trelew) y Mar'17 (ídem). El gráfico 8 permite apreciar las series original y filtrada para Trelew entre 2009 y 2017:

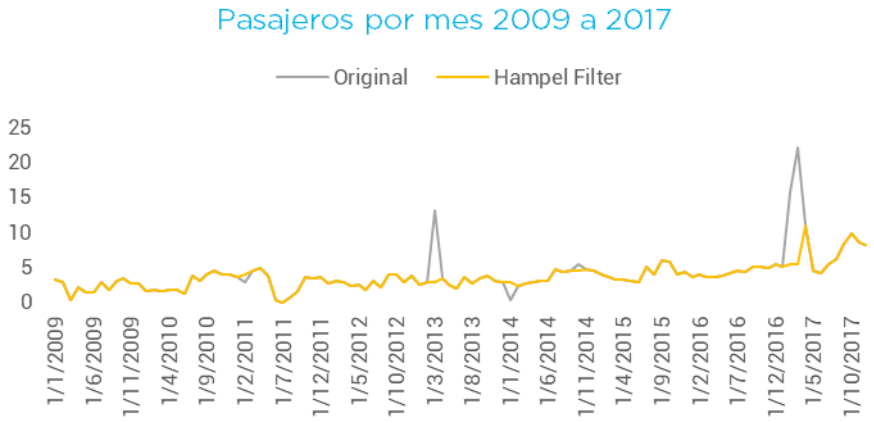


Gráfico 8. Pasajeros Trelew 2009 a 2017 - Comparación Original vs Filtro Hampel con k=3

El valor del parámetro k es clave a la hora de definir el filtro, dado que determina la "sensibilidad" del mismo frente a cambios abruptos en la tendencia de la serie y su duración. Por ejemplo, el gráfico 9 muestra el resultado de rehacer los cálculos con los datos del aeropuerto de Mendoza, para k=6 (es decir, una ventana de 13 valores de la serie):

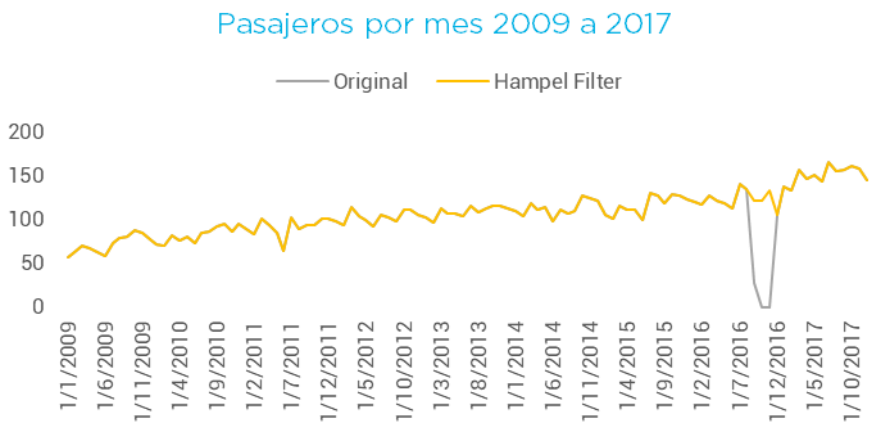


Gráfico 9. Pasajeros Mendoza 2009 a 2017 - Comparación Original vs Filtro Hampel con k=6

El problema está en aquellos aeropuertos cuyo comportamiento responde a picos de demanda muy marcados en algún momento del año. Suele pasar en algunos con los pasajeros internacionales. El gráfico 10, a continuación, muestra las series de pasajeros internacionales en Bariloche entre 2009 y 2017 para k=6:

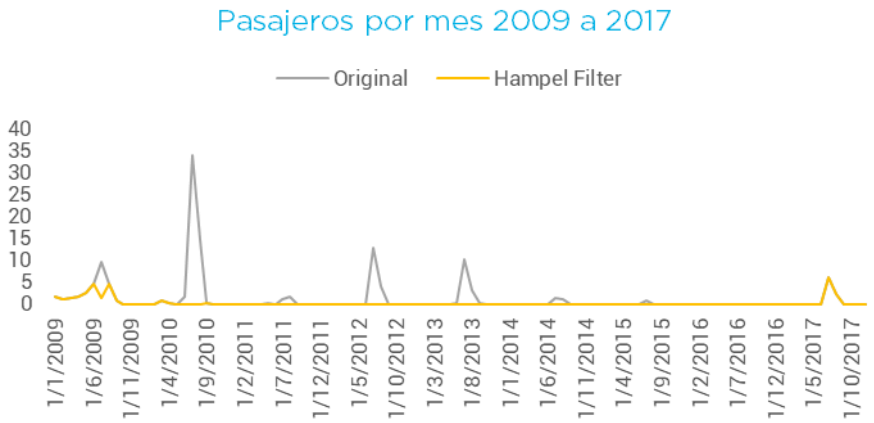


Gráfico 10. Pasajeros Internacionales Bariloche 2009 a 2017 - Comparación Original vs Filtro Hampel

Se hace evidente que el filtro, con ese valor de k, hace que se pierda la (poca) información relevante que tiene ese aeropuerto para ese tipo de pasajeros. En consecuencia, cualquier análisis que se haga sobre la serie filtrada para poder obtener un pronóstico, tiene un error significativo.

En definitiva, hay que revisar la elección del filtro dado que, en algunos casos, cumple con el objetivo planteado (omitir algún evento extraordinario, como el cierre de un aeropuerto por obras) y en otros, filtra información de la serie relevante para la regresión. Una solución para ello es dejar el valor del tamaño de la ventana de filtro como parámetro a definir por el usuario o que se itere sobre el mismo obteniendo distintos resultados posibles. Luego, sobre el set de prueba se puede estimar el error (con

una suma de los cuadrados de los mismos) y compararlo contra el que se obtiene con el modelado de la serie temporal original y allí decidir.

Modelo ARIMA

ARIMA hace referencia a un modelo autorregresivo integrado de promedio móvil, que utiliza las variaciones y regresiones existentes entre los datos para determinar los patrones intrínsecos en la serie y, a partir de ellos, puede generar un pronóstico de los mismos. Al igual que en el caso del modelo Holt-Winters, es un modelo univariado, donde los eventos futuros se proyectan únicamente en función de los datos históricos y no por variables independientes.

Sin embargo, a diferencia del anterior, que se basa en los valores de tendencia y estacionalidad de la serie temporal, un modelo ARIMA intenta explicar las correlaciones que existen entre los distintos puntos dentro de la propia serie.

Consta de las siguientes componentes (que, de hecho, le dan su nombre):

1. Autorregresiva (AR): asume que el valor de la serie en un determinado instante se corresponde con la combinación lineal de la función en instantes anteriores (hasta un número máximo determinado de ellos, llamado *"p"*), a lo que se adiciona un componente de error aleatorio. Es decir, la información presente de un evento está relacionada con los valores pasados.
2. Integración (I): se aplicarán sucesivas diferenciaciones en los casos en que las series muestren evidencia de no-estacionalidad.
3. Promedio Móvil (MA, del inglés *moving average*): asume que el valor observado en un instante se corresponde con un término de error aleatorio a lo que le adiciona una combinación lineal de errores aleatorios previos (hasta un número máximo de ellos, llamado *"q"*).

En un modelo MA, la correlación entre el objeto de un instante determinado en la serie de tiempo y los valores que se encuentran más allá de *p* períodos en el pasado (siendo *p* el orden del modelo) es siempre cero, mientras que en uno de tipo AR de igual orden, la correlación decrece progresivamente cuanto más en el pasado uno se ubica. En otras palabras, en un modelo MA, una variación extraordinaria (en nuestro caso podría ser la llegada de muchos nuevos pasajeros a un aeropuerto en un mes determinado como consecuencia del cierre de otro cercano), se disipa rápidamente en el tiempo; mientras que en otro AR, el efecto tiene una inercia mayor.

La componente autorregresiva (AR) y la parte del promedio móvil (MA) combinadas dan origen a lo que se conoce como modelo ARMA(p,q). En otras palabras, el valor presente depende de la información pasada (componente AR) y lo restante se modela gracias a un proceso de medias móviles (componente

MA). De todas maneras, para analizar una serie de tiempo y que ajuste un modelo de este tipo, se deben realizar las transformaciones necesarias de manera que la serie resulte estacionaria. Esto significa que se requiere que las observaciones no estén correlacionadas entre sí y que tengan media igual a cero, así como que la varianza y covarianza no dependan del paso del tiempo. En consecuencia, toda componente estacional y de tendencia de nuestra serie temporal debe ser omitida y únicamente se modela el ruido aleatorio.

Entonces, dada una serie de tiempo, hay que determinar en primer lugar si se trata de un modelo MA o AR y luego, el orden del mismo. Lo primero se resuelve con un gráfico denominado *ACF* (*Auto-Correlation Function*). En él se aprecia la correlación entre los valores para distintos grados de periodos pasados. Dado que en una función MA de orden *k* no existe correlación entre el elemento X_t y X_{t-k-1}, el gráfico corta el eje de las abscisas en ese valor de *k*. A modo de ejemplo, en el gráfico 11 se muestran los correspondientes para una serie MA(2):

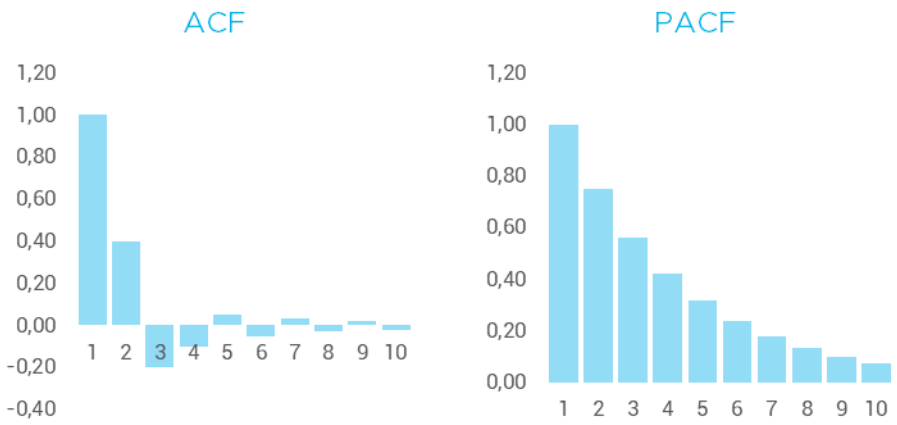


Gráfico 11. Ejemplo de gráficos ACF y PACF para una serie de tipo MA(2)

Si fuera una serie de tipo AR, la correlación disminuye progresivamente sin pasar a terreno negativo. Luego, si se tratara entonces de una serie AR y quisiéramos determinar su orden, deberíamos observar una gráfica de correlaciones parciales (PACF). Así, si se tratara de una serie AR de orden k, la correlación entre X_t y X_{t-k-1} es nula y allí se corta el eje de las abscisas. En el gráfico 12 se observa cómo sería el ACF y PACF para una serie de tipo AR(2):

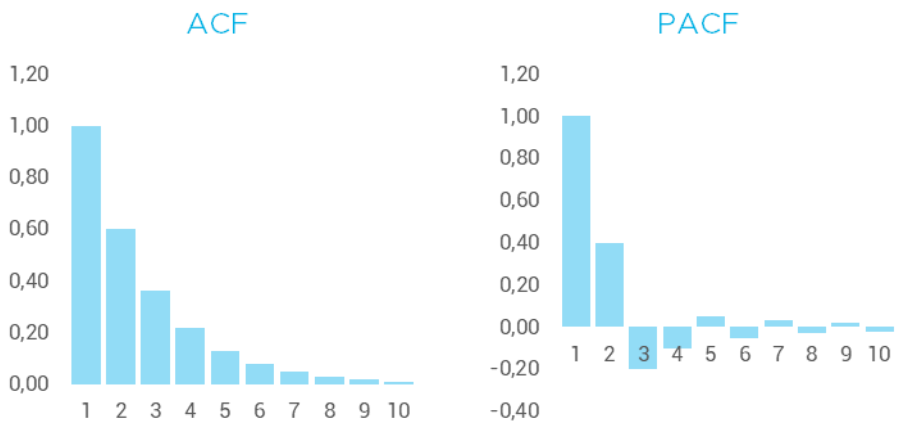


Gráfico 12. Ejemplo de gráficos ACF y PACF para una serie de tipo AR(2)

Cuando la serie no es estacionaria en varianza (dispersión no constante en el tiempo, también denominada *heteroscedasticidad*), se requiere una transformación logarítmica o de Box-Cox; mientras que para eliminar la componente de tendencia (serie no estacionaria en su media) se suelen considerar las diferencias entre los valores a lo largo de la serie en lugar de la original. En otras palabras, si la tendencia es lineal, el tomar la serie transformada $Y = X_t - X_{t-1}$ subsana esta cuestión. De cumplirse esto último, se puede afirmar entonces que estamos tratando con una serie temporal integrada de primer orden, que se denomina I(1).

En caso de tendencias de mayor orden, se hacen diferenciaciones sucesivas hasta determinar el coeficiente correspondiente y conseguir una serie estacionaria. El número de operaciones a llevar a cabo se denomina *"d"* y se identifica como *I(d)*.

Con todo ello, queda definido el modelo ARIMA(p,d,q).

A modo de ejemplo, podemos ver dos casos de series de pasajeros por aeropuerto que presentan un comportamiento marcadamente estacional. En primer lugar, el aeropuerto de Ushuaia, con una afluencia mayor en los meses de verano para luego tener un flujo inferior el resto del año. Para poner de manifiesto el componente estacional y la tendencia, el gráfico 13 muestra los pasajeros por mes del 2013 a 2017 para el aeropuerto de Ushuaia superpuestos, mientras que en el 14 se observan los valores mensuales en forma continua para ese período:

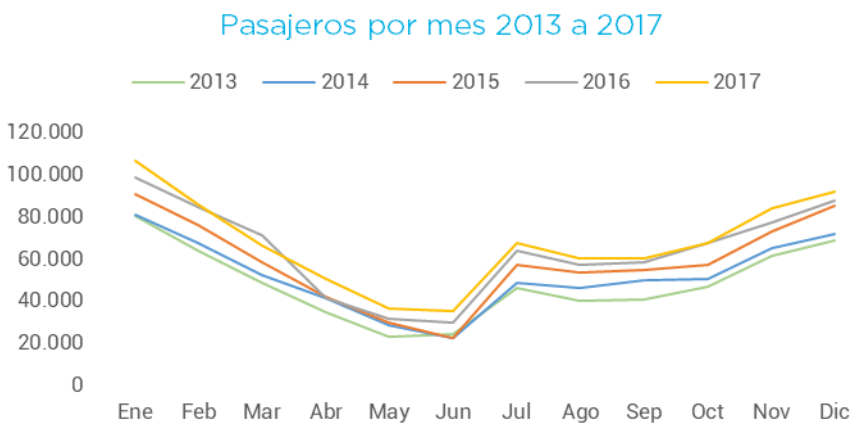


Gráfico 13. Pasajeros por mes 2013 a 2017 – Ushuaia

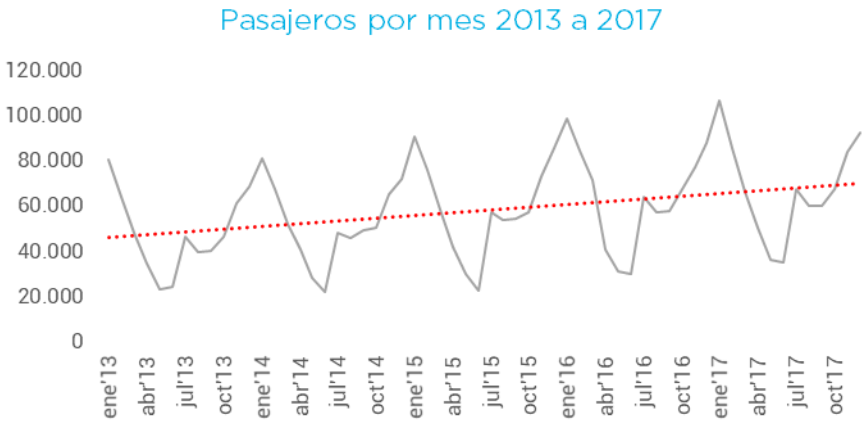


Gráfico 14. Pasajeros por mes 2013 a 2017 – Ushuaia

Estos gráficos permiten apreciar la importante componente estacional que presenta este aeropuerto: fuerte afluencia en diciembre y enero, a lo que se le suma un flujo un poco menor en julio, coincidente con las vacaciones de invierno en la mayoría de los casos. El segundo, deja en evidencia la tendencia creciente con los años y también se puede pensar que no hay un aumento de la varianza con los años: la diferencia entre los picos superiores e inferiores con respecto a la tendencia lineal no aumenta notoriamente. Haciendo en forma sucesiva la transformación logarítmica y de diferencias, se obtienen una serie que, a primera vista, parecer ser estacional. En el gráfico 15 se tienen las transformaciones logarítmica y diferencial de primer orden para los pasajeros de Ushuaia entre 2009 a 2017:

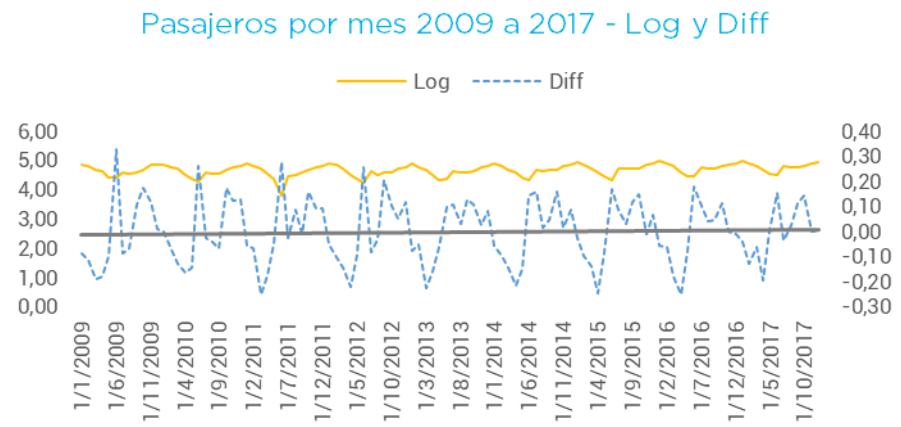


Gráfico 15. Transformaciones Log y Diff 2009 a 2017 – Ushuaia

La línea horizontal evidencia que ya con la segunda transformación (diferenciación aplicada sobre la serie logarítmica), se obtiene una serie con media 0 y estacional. En el gráfico 16 se observan los gráficos ACF y PACF para la serie (2009 a 2017) para el mismo aeropuerto:

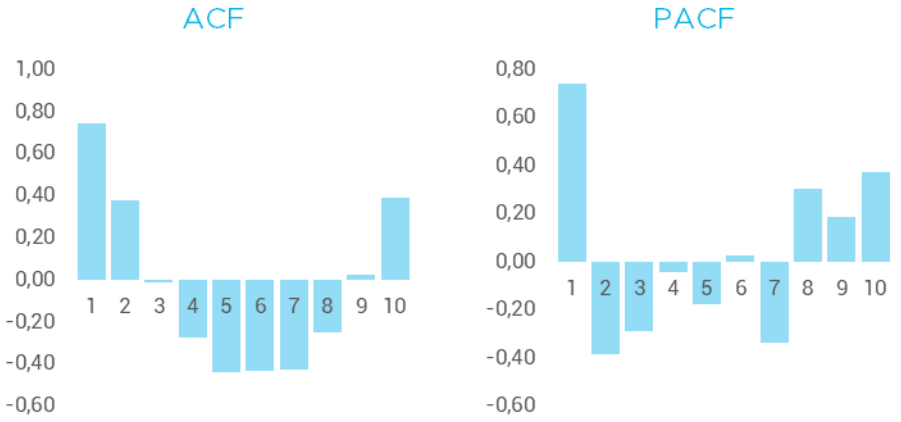


Gráfico 16. Gráficos ACF y PACF 2009 a 2017 – Ushuaia

Dado que el gráfico ACF tiende hacia cero en 4 períodos (*lags*), mientras que el PACF corta abruptamente en el segundo, podemos pensar que se trata de una serie a ajustar con un modelo de tipo AR(2).

Un segundo ejemplo, lo tenemos con el aeropuerto de Mar del Plata. En este caso, hay un pico de pasajeros muy importante entre diciembre y enero para luego mantener un caudal muy bajo a lo largo del año. Tal como se hizo con Ushuaia, en los gráficos 17 y 18 se muestran los valores de personas que arribaron y despegaron del aeropuerto de Mar del Plata, en forma superpuesta y continua entre los años 2013 a 2017:

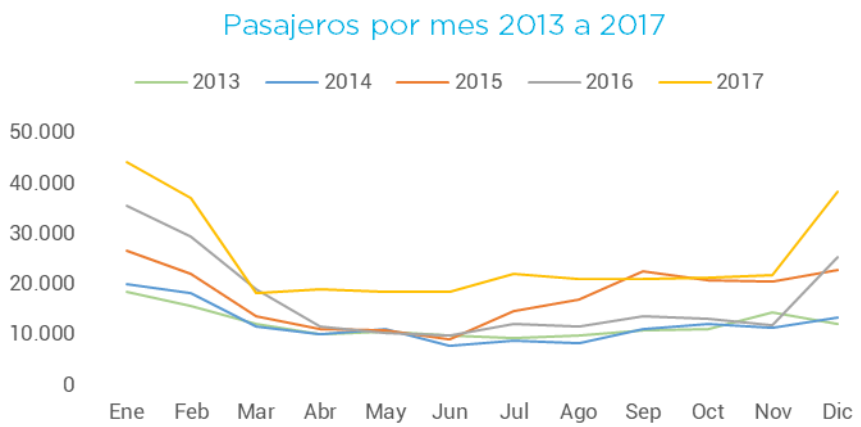


Gráfico 17. Pasajeros por mes 2013 a 2017 – Mar del Plata

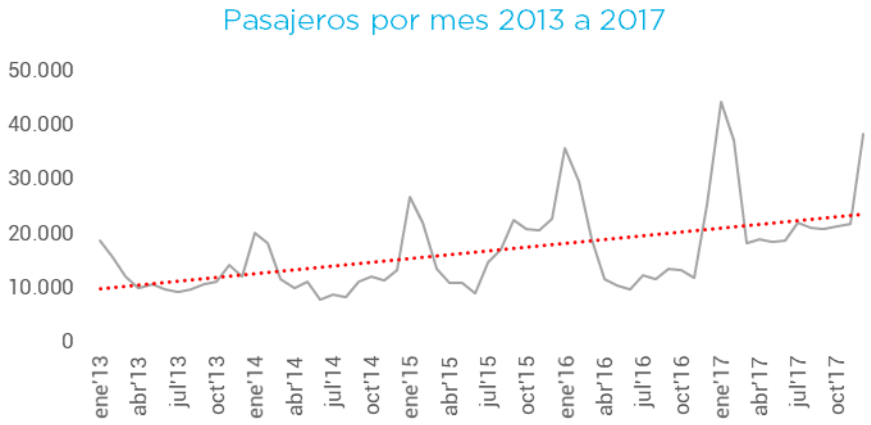


Gráfico 18. Pasajeros por mes 2013 a 2017 – Mar del Plata

Aquí, a diferencia de caso de Ushuaia, se aprecia cómo, si bien la tendencia lineal es creciente, los picos son cada vez mayores, lo que da idea de una varianza que no es constante en el tiempo. En el gráfico 19, observamos el efecto de las transformaciones sucesivas de logaritmo y diferencia de primer orden sobre esta serie:

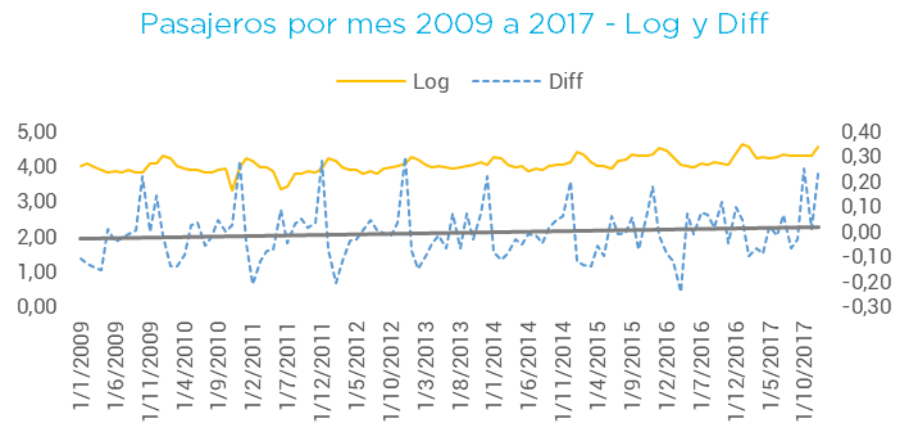


Gráfico 19. Transformaciones Log y Diff 2009 a 2017 – Mar del Plata

También conseguimos en este caso una serie final (Diff) que resulta, a priori, estacionaria (media en torno a cero) y cuya varianza (amplitud de los picos) no varía con el paso de los meses. En el gráfico 20, observamos los gráficos ACF y PACF para estos datos, de manera de estudiar qué modelo AR o MA ajusta:

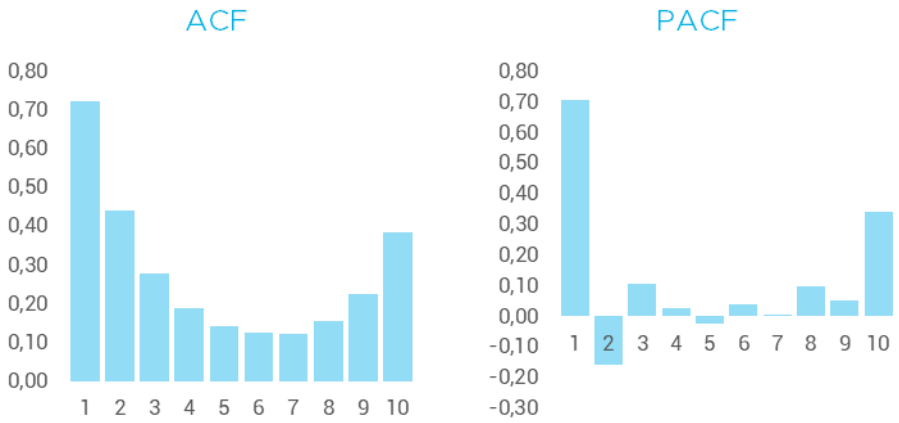


Gráfico 20. Gráficos ACF y PACF 2009 a 2017 – Mar del Plata

Aquí, a diferencia de Ushuaia, es bastante más evidente que un modelo de tipo AR(2) ajusta bien, dado que el gráfico ACF, al menos para los primeros 10 *lags* ni siquiera corta el eje de las abscisas.

El problema con esta implementación es que hay que ir viendo cada serie en forma individual y llevar a cabo las transformaciones pertinentes para luego poder determinar los parámetros que definen el modelo (p, q, d) y a partir de allí, poder hacer las predicciones correspondientes. Una generalización de todas las series a un modelo con los parámetros (p, q, d) fijos no funciona adecuadamente para todos los aeropuertos.

El proceso de ajuste con un modelo ARIMA se podría resumir en los siguientes pasos:

1. Observar la serie de tiempo (*plot*) y determinar si hay valores inusuales o *outliers* que hay que dejar fuera de la muestra o suavizar la serie.
2. Llevar a cabo las transformaciones necesarias para estabilizar la varianza (por ejemplo, tomando el logaritmo de la serie).
3. Si la serie no es estacional (la original y, en consecuencia, la logarítmica), ir tomando las primeras diferencias hasta que así resulte (en nuestro caso de pasajeros por aeropuerto es esperable que haya cierto componente estacional en la demanda).
4. Observar los gráficos ACF y PACF y ver si podemos determinar el valor de *p* o *q* que ajustaría un modelo AR o MA, como se mostró en los ejemplos anteriores.
5. Ir probando los modelos y determinar el mejor (con algún criterio o test, como puede ser la sumatoria de los cuadrados de los errores u otros más sofisticados como el criterio AIC (*Akaike Information Criteria*).
6. Verificar que los residuos del modelo elegido correspondan a una serie aleatoria ("ruido blanco"), es decir, con media centrada en cero, varianza constante y valores no correlacionados. Esto significaría que ya hemos podido extraer el máximo posible de información de nuestros datos.
7. Realizar el pronóstico con el modelo elegido.

Una generalización de los modelos ARIMA(p,q,d) es cuando los residuos de la función también presentan un comportamiento que puede ser modelado de esta misma manera. Esto se corresponde con series con un componente estacional marcado (se repite en todas las observaciones). En estos casos, estamos hablando de ajuste con modelo SARIMA(p,q,d)(P,Q,D)$_m$, haciendo referencia a ambas partes individuales con sus componentes autorregresivo (p,P), de medias móviles (q,Q) y diferencial (d,D). El valor de *m* corresponde con el número de períodos por estación (es decir, 12 si es anual):

$$ARIMA\ (p,q,d)\ (P,Q,D)_m$$

$$\begin{cases} (p,q,d) \rightarrow Parte\ no\ estacional\ del\ modelo \\ (P,Q,D)_m \rightarrow Parte\ estacional\ del\ modelo \end{cases}$$

Facebook Prophet

En febrero 2017, *Facebook* liberó una herramienta (de código abierto) disponible para Python y R, llamada *Prophet*. En grandes líneas, es una librería que permite construir modelos de ajuste y pronósticos de series. Para ello no utiliza los métodos más tradicionales (como ARIMA, por ejemplo) sino que lo que ellos denominaron *curve fitting*.

Permite manejar sets de datos en donde haya observaciones nulas, faltantes, grandes *outliers*, cambios de tendencia importantes (en nuestro caso podría ser la llegada de una nueva aerolínea a un aeropuerto o un aumento significativo de las frecuencias, sobre todo en aeropuertos chicos, por ejemplo) o mismo, tendencias no lineales. Es decir, resulta extremadamente versátil para utilizarlo sobre todo tipo de serie de tiempo sobre la cual se quieran realizar pronósticos.

De acuerdo a los desarrolladores, los resultados se obtienen de manera muy sencilla y son comparables con aquellos pronósticos llevados a cabo por analistas especializados o utilizando métodos más sofisticados.

En la parte más técnica, *Prophet* es un modelo regresivo aditivo con cuatro componentes:

- Tendencia: detecta automáticamente cambios en la misma seleccionando los distintos quiebres de tendencia dentro del set de datos y así arma la función (definida por partes) de tendencia linear o de crecimiento logístico (que alcanza nivel de saturación).
- Estacionalidad anual: la modela utilizando series de Fourier
- Estacionalidad semanal: la modela con variables de tipo *dummy*.
- Fechas importantes, feriados, etc: el usuario las puede definir de antemano si significan un quiebre a tener en cuenta por el modelo. En nuestro caso, le podríamos ingresar como parámetro las fechas de obras en los aeropuertos.

Con esto, el modelo puede ser formulado de la siguiente manera:

$$y_{(t)} = g_{(t)} + s_{(t)} + h_{(t)} + \varepsilon_t$$

Donde:

- $g_{(t)}$: función (definida por partes) que describe la tendencia lineal o de crecimiento logístico de largo plazo (modelado de los cambios no periódicos de la serie)
- $s_{(t)}$: función que define la estacionalidad (anual, mensual, semanal, etc.)
- $h_{(t)}$: hechos concretos que pueden alterar los valores de la serie (vacaciones, feriados, paros, etc.), a definirse por el usuario
- ε_t: término de error para aquellos cambios en los valores de la serie no ajustados por el modelo

Como se observa, el tiempo es utilizado como regresor y *Prophet* intenta ajustar las funciones (lineales o no) a los datos sumando los distintos efectos (igual concepto que el aplicado en el modelado de *Holt-Winters*, por ejemplo). Algo importante a destacar es que no busca encontrar la dependencia intrínseca entre los eventos y el tiempo, sino que es simplemente un ajuste a la curva.

El primer término se modela como una curva de crecimiento logístico (curva sigmoidea) si así se introduce el parámetro correspondiente o lo detecta en forma automática. Estos casos corresponden a situaciones en donde se da una saturación que no permite el crecimiento más allá de un límite determinado, cuestión que desechamos en nuestro análisis dado que se puede presuponer una capacidad ociosa importante en los aeropuertos a nivel nacional (y, donde no, se asume que se darán las obras tales que permitan albergar a los potenciales pasajeros futuros). Por otro lado, la curva de tendencia se determina por tramos, detectando automáticamente el algoritmo de *Facebook* los puntos de quiebre correspondientes (aunque esto también puede ser incorporado en forma manual por el usuario) y permitiendo así un ajuste que modele correctamente aquellos cambios rápidos de tendencia por distintos motivos.

En el caso del ajuste estacional, éste se hace utilizando series de Fourier. En su expresión genérica, una sumatoria infinita que converge a una función periódica y continua. En otras palabras, que toda función de este tipo puede ser expresada como la sumatoria de funciones seno y coseno:

$$f_{(t)} \sim \frac{a_0}{2} + \sum_{n=1}^{\infty} \left[a_n cos\left(\frac{2n\pi}{T} t\right) + b_n sen\left(\frac{2n\pi}{T} t\right) \right]$$

Particularmente, en esta aproximación, se define entonces la función $s_{(t)}$ como:

$$s_{(t)} = \sum_{n=1}^{N} \left[a_n cos\left(\frac{2n\pi}{T} t\right) + b_n sen\left(\frac{2n\pi}{T} t\right) \right]$$

Donde T es el período correspondiente (7 sí es información semanal, 365.25 si es anual, etc.) y los parámetros [a_1, a_2, …, a_N] y [b_1, b_2, …, b_N] deben ser estimados para un dado N. Este último, además, determina si los componentes de alta frecuencia deben ser considerados información relevante o ruido. Es decir, para un mayor N, mayores son las pequeñas variaciones que el modelo puede ajustar (cantidad de armónicos de la función a incorporar). De no especificarse manualmente, *Prophet* utiliza un N de 10.

Por último, el algoritmo permite el ingreso de eventos que deben ser analizados en forma particular. Si tratamos con información con detalle diario de pasajeros por aeropuerto, podríamos no sólo incorporar las fechas especiales o feriados del año (Semana Santa, Año Nuevo, Navidad, etc.) sino también todas aquellas fechas donde hubo, por ejemplo, un paro total o parcial de transporte que pudo haber afectado el normal tránsito aéreo. De todas maneras, al tratar con información con agregación mensual, podemos obviar el efecto de esas fechas particulares y sólo incorporar los meses en los cuales el movimiento de pasajeros en los aeropuertos se haya visto afectado en gran medida, tal como las obras (que afectan en unos en forma negativa y en los que funcionan de alternativa, positivamente), cenizas

volcánicas (tal como hubo entre 2010 y 2011 que afectó toda la operación a nivel nacional), entre otros. Si tuviéramos información futura, también se podría incorporar. Este *dataset* se puede preparar fácilmente en una hoja de cálculo para luego ser leída desde R y agregada al modelo.

Redes Neuronales

Las redes neuronales informáticas tratan de simular, en forma virtual, el comportamiento de un cerebro biológico. De igual manera, están compuestas por unidades simples interconectadas (donde se llevan adelante operaciones de suma) y cuyos enlaces pueden hacer que el estado de activación de las células vecinas se vea incrementado o inhibido.

A diferencia de otros algoritmos informáticos, pueden aprender por sí mismos a medida que va ingresando nueva información al sistema y el modelo se va adaptando, variando los pesos de las conexiones.

La ilustración 1 es el esquema para una única neurona (denominada "*perceptrón simple*"):

$$n_f = n_1 \times \omega_q + n_2 \times \omega_2$$

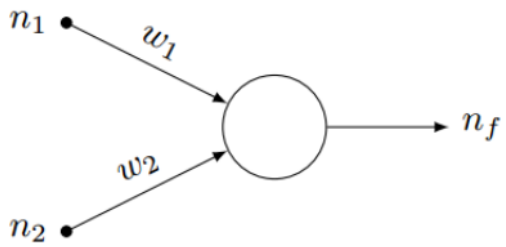


Ilustración 1. Perceptrón simple

El valor de n_f se puede comparar contra un determinado umbral, para el cual la función adopta un resultado binario si lo supera o no. La red entonces se entrena teniendo varios conjuntos con distintos valores de n_1, n_2 y n_f, de manera de poder ir ajustando los valores de los pesos (ω_1 y ω_2) hasta que obtenemos los valores de n_f adecuados (en un comienzo, son todos aleatorios).

Conceptualmente esto no cambia cuando tenemos varias neuronas conectadas entre sí y en distintas capas sucesivas. Cuantas más capas intermedias tengamos, más información podremos incorporar al modelo. Un modelo sin capas intermedias es equivalente a una regresión lineal: los valores de salida se corresponden a una combinación de los valores de entrada de los predictores. A partir de allí, las capas intermedias hacen que el modelo deje de ser lineal.

Si bien la idea general se mantiene, luego los algoritmos se van haciendo más complejos. Por ejemplo, en lugar de utilizar una función escalón en cada neurona (obtener un resultado binario luego de

procesar los pesos y valores de entrada y ver si supera o no el umbral), se puede utilizar una función sigmoidea para poder tener en salida variaciones más suaves en función de pequeñas oscilaciones en las entradas. Esto tiene como consecuencia un aprendizaje más fácil de la red ya que puede adaptar los pesos más finamente para obtener el mismo resultado en la salida.

Generalmente, las redes neuronales son utilizadas en problemas de clasificación, donde se busca predecir el estado de una variable (clase) que es función de varios parámetros de entrada y son muy útiles dado que permiten un análisis multivariado mucho más profundo que el que el ser humano podría llevar adelante. El entrenamiento de las mismas consiste, justamente, en encontrar los pesos de cada una de las conexiones de tal manera que se obtenga el resultado esperado dado un set de parámetros *input*. Luego, la red neuronal "entrenada" puede predecir el resultado en función de un nuevo juego de valores en sus entradas.

Al tratar con series de tiempo, los valores en las entradas se corresponden con el nivel de la función en los períodos pasados, de igual manera que lo veríamos en un modelo autorregresivo. Se suele utilizar la siguiente notación para especificar estos modelos:

$$NNAR(p, k)$$

Donde *p* corresponde al número de observaciones previas a considerar (y_{t-1}, y_{t-2}, y_{t-3}, ..., y_{t-p}) y *k*, el número de neuronas en la capa intermedia. De esta manera, un modelo NNAR(p, 0) es equivalente a un ARIMA (p, 0, 0).

Luego, se puede hacer una generalización adicional en el caso de tratar con series estacionales, de manera de poder incorporar no sólo una determinada cantidad de períodos anteriores a y_t, sino también el valor correspondiente al mismo elemento del período anterior. Es decir, si se tratara de una serie con estacionalidad anual (como es en nuestro caso de pasajeros por aeropuerto), es esperable que el valor para un determinado mes guarde estrecha relación con aquel de igual período del año previo. Así, la notación se extiende como sigue:

$$NNAR(p, P, k)_m$$

Donde *P* hace referencia, justamente, al número de elementos de períodos previos a considerar para el mismo elemento. Es decir, que las entradas son (y_{t-1}, y_{t-2}, y_{t-3}, ..., y_{t-p}, y_{t-m}, y_{t-2m}, ..., y_{t-Pm}), siendo *m* la longitud del período (12, si es anual). Luego, un modelo NNAR(p, P, 0)$_m$ es equivalente a un ARIMA (p, 0, 0)(P, 0, 0)$_m$.

Implementación

A continuación, se describen los principales pasos que sigue el algoritmo escrito en R para conseguir, a partir del set de datos de los pasajeros por aeropuerto, la proyección de esta variable para los próximos 5 años.

Archivo de entrada

En una hoja de cálculo, se construye el set de datos con el cual se trabajará para los 56 aeropuertos del país, así como una serie adicional para la sumatoria de los pasajeros a nivel nacional. Aquí hay una diferencia importante a mencionar: en el caso de cada aeropuerto, se trabaja con los pasajeros que llegan y salen del mismo, pero, al ver el país como un todo, se debe evitar el doble conteo de los pasajeros de cabotaje (el despegado desde uno será el arribado a otro). Es por ello, que la serie adicional correspondiente a los pasajeros a nivel nacional, no es equivalente matemáticamente a la suma lineal de todos los aeropuertos y se la proyecta en forma individual (en lugar de sumar los resultados obtenidos).

Los campos a incorporar al modelo son los siguientes:

- Año: formato yyyy
- Mes: formato m
- AñoMes: formato yyyymm
- Clasificación: Cabotaje / Internacional
- Aeropuerto: código de 4 caracteres asignado por OACI (*Organización de Aviación Civil Internacional*) que identifica en forma unívoca a cada uno de ellos
- Pasajeros: número de pasajeros por mes

A modo de ejemplo, en la tabla 1 se muestran los primeros 6 datos para Aeroparque (código "SABE"):

Año	Mes	AñoMes	Clasificacion	Aeropuerto	Pasajeros
2001	1	200101	Cabotaje	SABE	428973
2001	2	200102	Cabotaje	SABE	403086
2001	3	200103	Cabotaje	SABE	426197
2001	4	200104	Cabotaje	SABE	377537
2001	5	200105	Cabotaje	SABE	334841
2001	6	200106	Cabotaje	SABE	310274

Tabla 1. Encabezado del subset de Aeroparque

Algunos aeropuertos no tienen afluencia de pasajeros en forma constante cada mes o bien, no tienen vuelos internacionales. Otros, como Junín o Base Marambio, no tienen registros de pasajeros en absoluto (al menos desde enero 2001 en adelante). Sin embargo, para homogeneizar el juego de datos,

se deben ingresar todos los meses para ambos tipos de vuelos (cabotaje e internacionales) y se completa con 0 en el campo Pasajeros.

Esta base de datos se guarda en un archivo de texto, con el nombre "Base_SIAC.txt".

Por otro lado, se prepara un set de datos gemelo con el nombre "Base_SIAC_conNA.txt". La diferencia con el primero es que, en este, remplazaremos los valores de pasajeros por "NA" en todos aquellos casos que, a priori, buscamos dejar fuera del análisis. Ejemplo de ello son los meses en los que un aeropuerto permanece cerrado (y tiene menor o nula afluencia de pasajeros o sirve como alternativa y recibe más pasajeros de lo habitual) o algunos eventos puntuales, tales como los meses entre 2010 y 2011 en donde varios vuelos se vieron afectados por la presencia de cenizas volcánicas en el sur de nuestro país (aunque incluso se llegaron a cancelar vuelos salientes de Ezeiza y Aeroparque).

Si bien, como se analizó en el modelo de Holt-Winters, se puede usar un filtro como el Hampel para buscar suavizar la serie y evitar dichos valores atípicos, para procesar los períodos faltantes se utiliza una función de aproximación disponible en R: *na.approx()*. Ésta remplaza todos los valores faltantes de la serie temporal con una interpolación lineal entre los datos más cercanos disponibles. De esta manera, tenemos un manejo más homogéneo para todos los modelos de los datos que ya sabemos de entrada qué queremos dejar fuera de análisis. Como se mostrará luego en la implementación del modelo Holt-Winters, entonces, el efecto del filtrado se sumará al de dejar fuera estos datos y se evitarán, además, pequeños *outliers* locales a lo largo de la serie.

La única razón, entonces, por la que ingresamos ambos sets de datos es poder contar en un archivo de salida con los datos originales y los resultados de todos los modelos (que toman como *input* el set con los registros omitidos).

A partir del set de datos de entrada, leemos el valor del año y mes mínimo y máximo, para determinar cuál será la fecha máxima hasta donde llegará la proyección a realizar. Se toma un horizonte de 5 años (60 meses) a partir del último mes disponible, aunque este parámetro se puede modificar fácilmente.

Con los valores de año y mes, se incorpora el campo "*Fecha*", con formato *yyyy-mm-dd*, donde para el día, tomaremos el valor 01.

Una base auxiliar que se lee al comenzar el programa contiene la lista de los 56 aeropuertos a modelar y el valor número 57, dado para el valor "TODOS". Luego, todo el algoritmo se desarrolla dentro de un ciclo que aumenta desde el valor 1 al 57 (leyendo el número de filas de esta base) y se procesa aeropuerto por aeropuerto leyendo el valor del código OACI correspondiente.

Como último paso, colocamos la clasificación (cabotaje / internacional) como valor de columna y se selecciona únicamente el campo Fecha. En la tabla 2 se pueden observar, a modo de ejemplo, las primeras filas para Aeroparque con la estructura final:

Fecha	Cabotaje	Internacional
2001-01-01	428973	91467
2001-02-01	403086	71078
2001-03-01	426197	53346
2001-04-01	377537	46458
2001-05-01	334841	38146
2001-06-01	310274	33066

Tabla 2. Subset Aeroparque procesado

A excepción del modelo *Prophet*, que trabaja con estructuras del tipo *data-frame*, como la anterior mostrada, el resto toma como dato de entrada estructuras de R denominadas *time series*. Para ello, la función *ts()* realiza la transformación, tomando como parámetros los valores de año y mes iniciales y finales (que leímos anteriormente del dataset original) y la frecuencia (12, en nuestro caso, dado que son series mensuales).

A partir de la serie de tiempo, se hace una partición de la misma, de modo de obtener un conjunto de datos de entrenamiento (*training*) y otro de testeo (*testing*). Para ello, se utiliza un factor de partición de 70%: del total de meses que contiene el dataset original, tomamos el valor entero que corresponde al mencionado porcentaje de dicha cantidad y se separa el set de *training*, mientras que el resto de los meses (datos más actuales), sirven como set de prueba de los modelos. A modo de ejemplo, con datos entre enero del 2001 y junio de 2018, el set de entrenamiento contiene los datos desde el primer mes hasta marzo del 2013, mientras que el de testeo toma los siguientes 63 datos.

Algoritmo general

La separación en estos dos subconjuntos tiene únicamente como función el poder comparar los modelos. Con ello, el algoritmo general a ejecutar será el siguiente:

1. Se ajusta el modelo utilizando el subset *training*.
2. Se realiza la proyección de los datos con un horizonte dato por el número de registros del subset *testing*.
3. Se comparan los valores reales del subset *testing* con los valores proyectados por el modelo. Se toma como variable de comparación la media de los cuadrados de los errores (*RMSE – Root Mean Square Error*).
4. Se ajusta el modelo utilizando todo el set de datos.
5. Se realiza la proyección de los datos con el horizonte deseado (60 meses).

6. En caso de obtener en algún mes futuro un valor proyectado negativo, se remplaza por 0.

De manera de sortear ejecuciones del código innecesarias, se evalúan los pasajeros domésticos e internacionales de cada aeropuerto en forma separada. Si la suma de ellos a lo largo del set *training* no supera un valor de barrera (se toma 1000, omitiendo los meses faltantes), se evita modelar este aeropuerto para ese tipo de pasajeros (cabotaje / internacional) y la proyección queda simplemente completada con valores 0. Es decir, aeropuertos que sólo tienen (y tuvieron) pasajeros domésticos, se evitará el modelado de los datos (nulos o escasos) de vuelos internacionales, mientras que terminales que tuvieron escasa o nula afluencia de pasajeros totales (Base Marambio, Junín, Termas de Río Hondo, por citar algunos ejemplos) se omiten en su totalidad. Una objeción a ello podría ser el hecho de que un aeropuerto no haya tenido movimientos en todos los meses que abarca el subset *traning* pero sí en forma posterior. De todas maneras, en tales casos, una proyección realizada con tan pocos datos tendrá un error importante y tendrá más sentido realizar un análisis puntual en forma separada y no dentro de un algoritmo que buscar entregar un procedimiento homogéneo y general para todos los aeropuertos.

Dentro del ciclo que itera uno por uno los distintos aeropuertos, se separa en dos grandes bloques el tratamiento del *subset* de pasajeros de cabotaje e internacionales siguiendo la secuencia de 5 pasos mencionada anteriormente. La única excepción a ello es dentro del modelo de suavizado exponencial (Holt-Winters) donde, además, se compara el ajuste a la serie de datos original contra la que modela los datos filtrados con el suavizamiento *Hampel*:

1. Se ajusta el modelo utilizando el *subset training*.
2. Se ajusta el modelo utilizando el *subset training* filtrado:
    a. Se itera con distintos valores del parámetro *k* del filtro *Hampel* (amplitud de la ventana de datos sobre la cual se calcula la mediana y se compara cada punto para determinar si es *outiler* local o no), tomando valores entre 3 y 10.
    b. Para cada valor de *k*, se ajusta el modelo, se proyecta sobre el *subset testing* y se guarda el valor de *RMSE*.
    c. Se guarda el valor de *k* para cual el error medio cuadrático es el menor.
3. Se realiza la proyección de los datos con un horizonte dato por el número de registros del subset *testing*, tanto en su versión original como la filtrada.
4. Se comparan los valores reales del subset *testing* con los valores proyectados por el modelo. Se toma como variable de comparación la media de los cuadrados de los errores y determinamos si el ajuste con los datos originales es mejor (según este criterio) que el realizado con los datos filtrados.
5. Se ajusta el modelo utilizando todo el set de datos.

6. Se realiza la proyección de los datos con el horizonte deseado (60 meses).
7. En caso de obtener en algún mes futuro un valor proyectado negativo, se remplaza por 0.

ARIMA

Dado que el objetivo de este estudio es conseguir una función lo más automática posible, se emplea la función *auto.arima*, disponible dentro del paquete *forecast* de R. Ésta determina el mejor modelo ARIMA dentro de todas las posibilidades, habiendo llevado a cabo las transformaciones pertinentes. Para determinar cuál es el más adecuado, se puede elegir uno entre varios métodos de comparación: AIC, AICc (criterio de información de Akaike para muestras pequeñas) o BIC (criterio de información bayesiano). También se pueden ingresar como parámetros de la función los valores iniciales y máximos para las 6 componentes (p, P, q, Q, d, D) y el algoritmo prueba (si se definen los argumentos de esta manera) uno por uno los distintos modelos y se queda con aquel que compare mejor según el criterio elegido (AIC por ejemplo).

Claro está que un análisis particular individual de cada aeropuerto puede arrojar mejores o más precisos resultados, pero excede al alcance de este estudio.

Redes Neuronales

Se prueban tres implementaciones distintas de redes neuronales para la proyección de series temporales:

1. *nnetar()*: función presente dentro del paquete *forecast* de R, permite ajustar una serie de tiempo con una red neuronal pre-alimentada (*feed-forward neural net*) a partir de valores anteriores al último punto (*lagged values*), utilizando una única capa oculta y permitiendo realizar pronósticos para series univariadas. Un parámetro de la función permite establecer el número de modelos que entrenará. Se adoptó un valor fijo de 100 redes, dado que con números mayores (1000, por ejemplo), el algoritmo demanda significativamente más tiempo sin que los resultados sean mejores.
2. *mlp()*: esta función crea una red neuronal artificial denominada "perceptrón multicapa" (de ahí la sigla: "*multilayer perceptron*"). De no definirse ningún parámetro, la función intentará en forma automática especificar las entradas autorregresivas y cualquier tipo de pre-procesamiento que requieran los datos de la serie de tiempo. Entrenará en total 20 redes, cada una con una capa oculta de 5 nodos y el resultado será el promedio del *output* de todas ellas.
3. *elm()*: esta función crea lo que se denomina "máquinas de aprendizaje extremo" (de allí la sigla: "*extreme learning machine*"). La diferencia con la anterior es que, en el caso por defecto, entrena 20 redes neuronales con 100 nodos en la capa intermedia (*hidden layer*).

Al igual que como se ha mencionado con el caso del ajuste ARIMA, bien valdría la pena intentar definir los parámetros a prueba y error para modelar cada aeropuerto en forma individual de manera de obtener los resultados más acordes al caso. Sin embargo, es válida la misma apreciación de que lo que en definitiva se busca es un proceso automático y lo más rápido posible que permita entregar resultados aceptables. Luego, se podrá indagar en algún caso en detalle, pero escapa al alcance de este estudio.

Archivos de salida

El algoritmo genera los siguientes archivos:

1. Resultados individuales por aeropuerto: dentro de cada ciclo de iteración, se genera un archivo cuyo nombre viene dado por el código OACI del mismo (ejemplos: Aeroparque = SABE, Mendoza = SAME, Ezeiza = SAEZ, etc.) y con una extensión .txt. Los campos que se muestran son:
    - Fecha
    - Cabotaje: serie original sin meses omitidos
    - Internacional: serie original sin meses omitidos
    - HW_Cabotaje: resultados del ajuste *Holt-Winters* para toda la serie de pasajeros de cabotaje y proyección a 60 meses.
    - HW_Internacional: ídem para pasajeros internacionales
    - AR_Cabotaje: resultados del ajuste *SARIMA* para toda la serie de pasajeros de cabotaje y proyección a 60 meses.
    - AR_Internacional: ídem para pasajeros internacionales
    - PR_Cabotaje: resultados del ajuste *Prophet* para toda la serie de pasajeros de cabotaje y proyección a 60 meses.
    - PR_Internacional: ídem para pasajeros internacionales
    - NN1_Cabotaje: resultados del ajuste con *nnetar()* para toda la serie de pasajeros de cabotaje y proyección a 60 meses.
    - NN1_Internacional: ídem para pasajeros internacionales.
    - NN2_Cabotaje: resultados del ajuste con *mlp()* para toda la serie de pasajeros de cabotaje y proyección a 60 meses.
    - NN2_Internacional: ídem para pasajeros internacionales.
    - NN3_Cabotaje: resultados del ajuste con *elm()* para toda la serie de pasajeros de cabotaje y proyección a 60 meses.
    - NN3_Internacional: ídem para pasajeros internacionales.

2. Resultados consolidados: se genera el archivo *tabla_consolidada.txt* donde se almacenan en forma progresiva los resultados (con los mismos campos que en el caso individual por aeropuerto) de cada terminal.
3. Resultados de ajuste de modelos: se genera el archivo *Resultados_RMSE.txt* donde se registran los valores del error cuadrático medio (*RMSE*) de cada modelo para cada tipo de pasajero y aeropuerto. En la tabla 3 se muestran las cifras para algunos aeropuertos a modo de ejemplo:

resultados_aeropuerto	resultados_hw_cab	resultados_ar_cab	resultados_pr_cab	resultados_nn1_cab	resultados_nn2_cab	resultados_nn3_cab	resultados_hw_int	resultados_ar_int	resultados_pr_int	resultados_nn1_int	resultados_nn2_int	resultados_nn3_int
SACO	52807	48492	43237	54946	54754	46945	27817	20710	12177	17572	14064	14899
SAEZ	18502	15894	27916	27269	44346	25044	42667	40139	51369	71147	61731	44095
SAME	19297	9303	7184	15119	10296	9540	9588	9411	10703	10762	11248	9452
SAZB	5077	5618	4619	9489	6623	6227	NA	NA	NA	NA	NA	NA
SAZS	23612	39995	38372	46709	54485	29452	1641	3681	5059	5732	7733	3404

Tabla 3. RMSE por modelo para Córdoba, Ezeiza, Bahía Blanca y Bariloche

Aquí se observa como en el caso de Bahía Blanca (SAZB) no se modelaron los pasajeros en vuelos internacionales.

Por otro lado, los valores no son comparativos de un aeropuerto a otro dado que no están normalizados y es esperable que el *RMSE* sea mayor cuanto así lo sea el número de pasajeros de cada uno. Es decir, los valores del error cuadrático medio me permiten evaluar cuál modelo ajustó mejor para cada aeropuerto, pero no si un mismo modelo ajustó mejor o peor entre dos terminales distintas.

Todos los archivos se almacenan dentro de una carpeta cuyo nombre corresponde con la fecha del día, de manera de que el usuario pueda almacenar distintas corridas a lo largo del tiempo. De hecho, eso es lo esperable, dado que cada mes habrá que reajustar los modelos con la nueva información del nuevo período cerrado y extenderlo un mes más en el horizonte deseado.

Tablero de control

Con la tabla consolidada generada en el algoritmo se arma un *dashboard* en Microsoft Excel que servirá al usuario final para poder observar los resultados en forma individual para cada aeropuerto, tanto en sus pasajeros de cabotaje como internacionales.

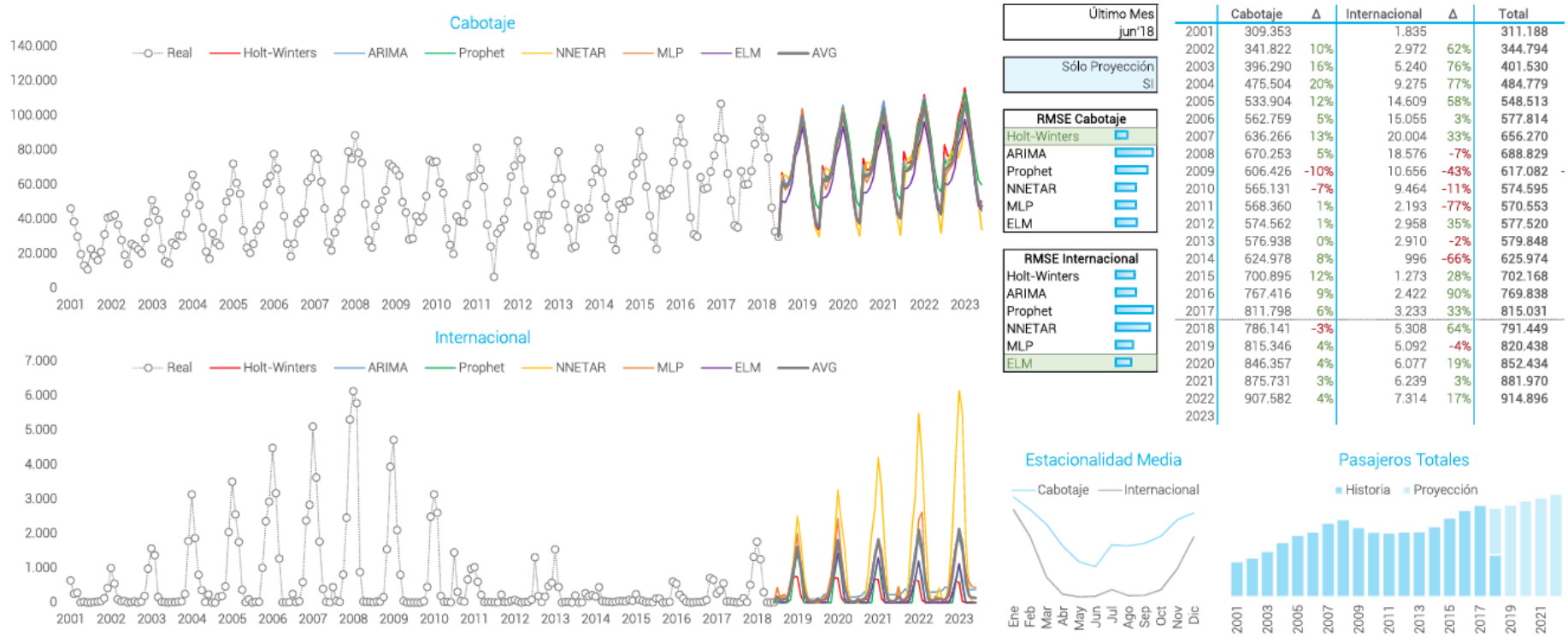
Se permite visualizar la información para algún modelo en forma individual o para el conjunto de ellos que se desee. En todos los casos, además, se mostrará una proyección dada por los valores promedio de todos los modelos. Utilizando dicho valor promedio, en una tabla se presentan los valores finales de pasajeros para cada año, así como el porcentaje de variación interanual.

Una tabla adicional compara los valores de *RMSE* de los 6 modelos estudiados y resalta cuál(es) de ellos ajusta mejor, en forma individual, para pasajeros de cabotaje e internacionales.

En otros dos gráficos, se observa el total por año y la estacionalidad promedio. En ambos casos, no se muestran valores, sino que se busca que sean simplemente de carácter orientativo, de manera de permitir observar el funcionamiento del aeropuerto en cuestión.

Por último, existe la opción de, para los datos históricos, mostrar únicamente los datos reales o bien, superponer a ellos el resultado de cada uno de los modelos (y así, observar cómo ajustaron al ser entrenados).

A modo de ejemplo, en la ilustración 2 se muestra el tablero de control para el aeropuerto de Ushuaia:



	Cabotaje	Δ	Internacional	Δ	Total
2001	309.353		1.835		311.188
2002	341.822	10%	2.972	62%	344.794
2003	396.290	16%	5.240	76%	401.530
2004	475.504	20%	9.275	77%	484.779
2005	533.904	12%	14.609	58%	548.513
2006	562.759	5%	15.055	3%	577.814
2007	636.266	13%	20.004	33%	656.270
2008	670.253	5%	18.576	-7%	688.829
2009	606.426	-10%	10.656	-43%	617.082
2010	565.131	-7%	9.464	-11%	574.595
2011	568.360	1%	2.193	-77%	570.553
2012	574.562	1%	2.958	35%	577.520
2013	576.938	0%	2.910	-2%	579.848
2014	624.978	8%	996	-66%	625.974
2015	700.895	12%	1.273	28%	702.168
2016	767.416	9%	2.422	90%	769.838
2017	811.798	6%	3.233	33%	815.031
2018	786.141	-3%	5.308	64%	791.449
2019	815.346	4%	5.092	-4%	820.438
2020	846.357	4%	6.077	19%	852.434
2021	875.731	3%	6.239	3%	881.970
2022	907.582	4%	7.314	17%	914.896
2023					

Ilustración 2. Dashboard para Ushuaia

Resultados

En cuanto al tiempo de ejecución del algoritmo, utilizando una notebook conectada a fuente de alimentación, procesador Intel i7 @2,70GHz y 16gb RAM, es de 3 a 4 minutos en promedio para cada aeropuerto. En total, se obtienen los archivos de salida para los 57 *inputs* luego de poco más de dos horas de ejecución (dado que, en algunos casos, se omite el modelado por falta de datos suficientes y la ejecución es prácticamente instantánea).

De todas maneras, es muy sencillo llevar adelante la ejecución del código omitiendo el ciclo de iteración y determinando un único aeropuerto para el cual se quiera obtener resultados.

En lo que refiere a los valores en sí, utilizando el tablero de control se puede ir navegando aeropuerto por aeropuerto comparando el ajuste y observar los resultados de cada uno de los modelos en los pasajeros de cabotaje e internacionales.

Como se mencionó previamente, la variable de comparación elegida para comparar el ajuste de los distintos modelos, es el error cuadrático medio, cuyo inconveniente es que no permite evaluar en forma transversal el ajuste en distintos aeropuertos dado que no es una variable normalizada.

En la tabla 4 se evalúa cuántas veces cada uno de los modelos ha realizado el mejor ajuste:

	Cabotaje	Internacional	Total
Holt Winters	15	5	**20**
ARIMA	15	5	**20**
Prophet	9	3	**12**
NNETAR	6	8	**14**
ELM	6	2	**8**
MLP	2	4	**6**
	53	**27**	**80**

Tabla 4. Cantidad de mejores ajustes por modelo

El número no se corresponde con el total de 56 aeropuertos (57 considerando la serie para la suma de todos ellos), dado que hay casos en donde la afluencia de pasajeros no es significativa y se omite su modelado.

Los modelos Holt-Winters y ARIMA son los que ajustan con un menor error sobre el total de los casos, situación que también se repite al evaluar únicamente las series temporales de pasajeros de cabotaje. Sin embargo, el método de red neuronal utilizando la función *nnetar()* es el que lo hace de mejor manera al comparar las series de pasajeros en vuelos internacionales. A priori, no se puede determinar el porqué de esta situación. Además, todas las funciones se ejecutaron prácticamente con sus parámetros en valores por defecto. Esto quiere decir que aún queda bastante margen para conseguir mejores y más

precisos ajustes. En tal caso, una recomendación es la de trabajar en forma individual por aeropuerto, descartando los parámetros *default* e iterar sobre los distintos valores de entrada de cada una de las funciones de manera de minimizar el error de ajuste.

Una situación que se observa en forma frecuente en el caso de las redes neuronales es la de alcanzar un valor estacionario a los pocos meses de proyección. En el gráfico 21 se tiene el resultado de ajuste y proyección dados por la función *nnetar()* para los pasajeros de cabotaje del Aeropuerto de Córdoba (SACO):

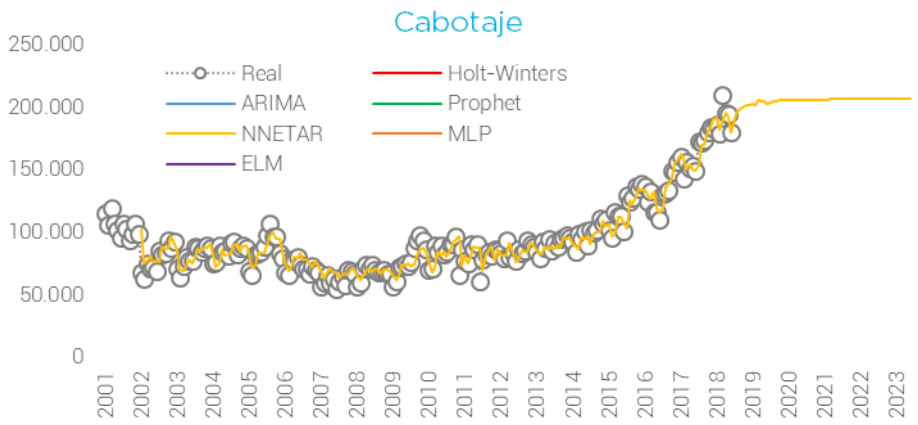


Gráfico 21. Ajuste función nnetar() para pasajeros de cabotaje en Córdoba

Esto es evidencia de que la red neuronal "deja de aprender" a medida que nos alejamos del último dato histórico disponible y nos adentramos en el horizonte de proyección. Es decir, no es capaz de encontrar, dado el set de datos de entrada, los patrones tales que le permitan inferir los resultados futuros.

En otros casos, los resultados que entregan los distintos modelos difieren radicalmente uno del otro. En el gráfico 22, observamos los pasajeros en vuelos regionales/internacionales en Aeroparque:

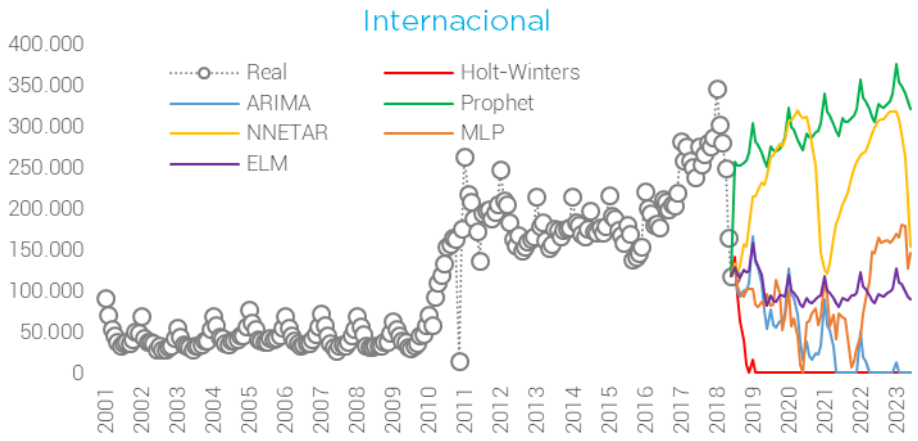


Gráfico 22. Ajuste de las distintas funciones para pasajeros internacionales en Aeroparque

El cambio de tendencia tan abrupto que sufrieron los vuelos internacionales en Aeroparque entre abril y junio del 2018 (dado que el 50% de los mismos se pasaron a Ezeiza), hace que todos los modelos propuestos evalúen en forma diferente la serie y entreguen proyecciones totalmente desiguales para los próximos 5 años. Si se observa, por ejemplo, Prophet, de Facebook, parece interpretar esta última información como "ruido" y regresa a la tendencia subyacente desde el año 2015 en adelante; mientras que el modelo de Holt-Winters asume que se continuará con el marcado descenso hasta anular por completo el valor de los pasajeros internacionales.

Claramente, los distintos algoritmos no tienen por qué considerar la realidad que subyace a la serie temporal de cada aeropuerto. En este último caso, por ejemplo, se tomó la decisión de migrar los vuelos regionales de Aeroparque hacia Ezeiza en dos etapas (mayo 2018 y 2019), dejando únicamente las conexiones con Uruguay desde esta terminal. Eso hará que los pasajeros en vuelos internacionales sean prácticamente nulos a partir de mediados del 2019. Es probable que a medida que se ingresen datos de julio 2018 en adelante, los modelos "entiendan" más lo que está ocurriendo en Aeroparque y mejoren sus proyecciones.

Dicho esto, se vuelve a mencionar que escapa al alcance del presente estudio en análisis pormenorizado de cada aeropuerto: la idea es generar un algoritmo que procese toda la información y entregue resultados que expliquen lo mejor posible a cada uno de ellos, sin considerar variables exógenas y en forma relativamente rápida y económica (en cuanto a procesamiento).

Por otro lado, el usuario final cuenta con los resultados para cada uno de los modelos, pudiendo elegir con cuál de ellos quedarse en forma individual para pasajeros de cabotaje e internacionales (o bien, tomando el promedio).

Por último, comparamos los valores obtenidos para la serie que involucra a todos los aeropuertos contra las hipótesis planteadas: los resultados obtenidos a diciembre 2022 alcanzan los 17,8 millones de pasajeros en vuelos domésticos (+37% vs 2017) y 18,6 millones en el caso de los pasajeros en vuelos regionales e internacionales (+27%). Estos números se encuentran por debajo de los planteado: 19,5 (-8,9%) y 21,0 millones, respectivamente (-11,4%). Sin embargo, los valores proyectados (considerando el promedio de los 6 modelos) corresponden a cifras obtenidas a partir únicamente de los datos históricos. Es decir, es un escenario que podría ocurrir de forma inercial. La introducción de medidas y políticas adecuadas podrá dinamizar el sector y llevarlo a los niveles deseados.

Conclusiones

El sector aerocomercial no disponía, hasta comienzos del año 2016, de ningún tipo de cifra sobre las cuales pararse para estudiar su desempeño. Más allá de las cifras que cada aerolínea manejaba por su cuenta (en función de sus tickets vendidos, por ejemplo), no existía a nivel gubernamental una base estadística que permitiera alinear todos los organismos de la órbita del Ministerio de Transporte en función de los objetivos de mediano y largo plazo.

Una primera respuesta ante tal vacío estadístico fue el desarrollo de las bases de datos que permitieron procesar toda la información disponible en el Sistema Integrado de Aviación Civil para, a partir de allí, y aun conociendo todas sus limitaciones, poder ofrecer un marco de referencia a quienes debían llevar adelante la gestión.

Una vez superada esta primera barrera, la pregunta que surge naturalmente se relaciona con el futuro: ¿qué actividad podemos esperar en los aeropuertos del sistema nacional en los próximos meses y años? La respuesta a dicha cuestión serviría para poder pensar las políticas a llevar adelante: priorizar inversiones tecnológicas, desarrollar la infraestructura necesaria, generar políticas de precios e incentivos adecuadas, pensar la red comercial de la aerolínea de bandera, establecer *hubs* o centros de conexiones regionales, etc.

En función de ello, se tomó el desafío de desarrollar un algoritmo que pueda generar, en forma rápida y sencilla, proyecciones de pasajeros tanto de vuelos de cabotaje como regionales e internacionales, por aeropuerto y con desglose mensual para un horizonte medio de 5 años.

A lo largo de estudio se analizaron distintos modelos que permiten ajustar las series de tiempo y a partir de ellas, inferir resultados: Holt-Winters y ARIMA son modelos estadísticos que trabajan en función de los patrones estacionales y de tendencia pasada; Prophet, de Facebook, ajusta los datos utilizando armónicos de funciones trigonométricas (series de Fourier) y, por último, las redes neuronales estudian las respuestas obtenidas en función de distintos valores de entrada, de manera de ajustar las ponderaciones de las conexiones y así recalcular la salida.

Para la implementación de cada una de las funciones, se utilizó un código escrito íntegramente en R y se optó por dejar, en la mayoría de los casos, los parámetros por defecto de cada una de ellas, dado que lo que se buscó fue una solución que trabajara en forma general.

Los valores de pasajeros por aeropuerto son heterogéneos a lo largo y ancho del país y también así lo son los resultados obtenidos para cada uno de ellos y en cada uno de los modelos, teniendo casos en los que la divergencia entre ellos es notoria y otros en los que, *a priori*, uno podría descartar dado que

se alejan de la realidad propia del aeropuerto. Sin embargo, el usuario puede hacer uso de un tablero de control para observar fácilmente cómo se han ajustado los modelos a sus datos y así determinar cuál de ellos le parece el más apropiado para cada aeropuerto.

El hecho de que los resultados obtenidos para el global del país se encuentren por debajo de los planteados en la hipótesis no resultan alarmantes: la diferencia no supera al 12%, considerando el promedio de las proyecciones de los modelos. La realidad es que lo que se observa a partir de este trabajo se podría considerar como un escenario de mínima: lo que es probable que pasa si se mantiene la situación actual. Sin embargo, la realidad es otra: desde comienzos del año 2016 el sector aerocomercial se ha dinamizado enormemente y lo está haciendo en forma homogénea a lo largo del país. Por ejemplo, tomando datos al primer semestre del 2018, los pasajeros se han incrementado, tanto en vuelos domésticos como internacionales, un 37% contra igual período del 2015; aeropuertos como Mar del Plata, Puerto Madryn, San Martín de los Andes, Bariloche, entre otros, han duplicado o triplicado su afluencia de pasajeros; los pasajeros en vuelos internacionales desde los aeropuertos del interior se han incrementado en un 170% y aquellos en vuelos domésticos que conectan ciudades del interior sin pasar por Buenos Aires han crecido en torno al 85%.

En definitiva, se puede esperar que este dinamismo continúe en el corto y mediano plazo e, independientemente de las cuestiones que puedan imperar en la macroeconomía del país, el avión ha comenzado a cambiar radicalmente la forma en que la Argentina está conectada. Lo desarrollado en este estudio permitirá ir siguiendo mes a mes la evolución del sector y el algoritmo irá incorporando esta nueva información y ajustará sus pronósticos, esperando que esa brecha entre lo obtenido a junio 2018 y lo deseado y planteado en las hipótesis del trabajo se achique cada vez más.

Como desarrollo ulterior, se puede buscar la manera de incorporar series temporales con información exógena que sirva como vector de regresión para cada aeropuerto en forma individual. Todos los modelos, a excepción de la función Holt-Winters, permiten la incorporación de un vector o matriz de regresores adicionales, aunque lo implementan en forma diferente. Por otro lado, se puede plantear la posibilidad de realizar pronósticos para otras variables que caracterizan al sector como, por ejemplo, la cantidad de despegues y aterrizajes en cada aeropuerto. Si bien en la gran mayoría de los casos, está directamente relacionado con el número de pasajeros, hay aeropuertos (el Aeródromo de Morón es el caso más significativo) donde las operaciones de aviación general con aviones de pequeño porte tienen una importancia mucho mayor que los vuelos comerciales (que incluso, pueden ser nulos).

Referencias

Alaya Solares, J. R. (17 de Abril de 2017). *Playing with Prophet on Bike Sharing Demand in Washington, D.C.* Obtenido de Towards Data Science: https://towardsdatascience.com/playing-with-prophet-on-bike-sharing-demand-time-series-1f14255f7ff0

Bontempi, G. (2013). *Machine Learning Strategies for Time.* Obtenido de Université libre de Bruxelles: http://www.ulb.ac.be/di/map/gbonte/ftp/time_ser.pdf

Bontempi, G. S.-A. (2013, Enero). *Machine Learning Strategies for Time Series Forecasting.* Retrieved from ResearchGate: https://www.researchgate.net/publication/236941795_Machine_Learning_Strategies_for_Time_Series_Forecasting

Brownlee, J. (2 de Diciembre de 2016). *What Is Time Series Forecasting?* Obtenido de Machine Learning Mastery: https://machinelearningmastery.com/time-series-forecasting/

Casares, F. (14 de Mayo de 2017). *La función auto.arima de R: una opción rápida para pronósticar*. Obtenido de Casares Félix: https://casaresfelix.com/2017/05/14/autoarima-r/

Chatterjee, S. (5 de Febrero de 2018). *Time Series Analysis Using ARIMA Model In R*. Obtenido de DataScience+: https://datascienceplus.com/time-series-analysis-using-arima-model-in-r/

Choudhary, A. (10 de Mayo de 2018). *Generate Quick and Accurate Time Series Forecasts using Facebook's Prophet (with Python & R codes)*. Obtenido de Analytics Vidhya: https://www.analyticsvidhya.com/blog/2018/05/generate-accurate-forecasts-facebook-prophet-python-r/

Ciflikli, G. (22 de Octubre de 2017). *Automatic Time-Series Forecasting with Prophet* . Obtenido de Computational Social Science: https://www.gokhan.io/post/prophet/#fnref2

Coglhan, A. (n.d.). *Using R for Time Series Analysis*. Retrieved from A Little Book of R for Time Series: http://a-little-book-of-r-for-time-series.readthedocs.io/en/latest/index.html

Dalinina, R. (1 de Octubre de 2017). *Introduction to Forecasting with ARIMA in R*. Obtenido de DataScience.com: https://www.datascience.com/blog/introduction-to-forecasting-with-arima-in-r-learn-data-science-tutorials

Glander, S. (13 de Junio de 2017). *Time Series Forecasting Part 3: Forecasting with Facebook's Prophet* . Obtenido de Shirin's playgRound: https://shiring.github.io/forecasting/2017/06/13/retail_forcasting_part3

Hayashi, H. (17 de Octubre de 2017). *Is Prophet Really Better than ARIMA for Forecasting Time Series Data?* Obtenido de Exploratory Blog: https://blog.exploratory.io/is-prophet-better-than-arima-for-forecasting-time-series-fa9ae08a5851

Hyndman, R. J. (27 de Octubre de 2011). Obtenido de Forecasting time series using R: https://robjhyndman.com/talks/MelbourneRUG.pdf

Hyndman, R. J. (Abril de 2017). *ARIMA modelling in R*. Obtenido de Forecasting: Principles and Practice: https://otexts.org/fpp2/arima-r.html

Hyndman, R. J. (Abril de 2018). *Neural network models*. Obtenido de Forecasting: Principles and Practice: https://otexts.org/fpp2/nnetar.html

Hyndman, R. J. (Abril de 2018). *Seasonal ARIMA models*. Obtenido de Forecasting: Principles and Practice: https://otexts.org/fpp2/seasonal-arima.html

Jinka, P. (20 de Marzo de 2018). *Holt-Winters Forecasting Simplified*. Obtenido de Vivid Cortex: https://www.vividcortex.com/blog/holt-winters-forecasting-simplified

Julián, G. (21 de Enero de 2016). *Las redes neuronales: qué son y por qué están volviendo*. Obtenido de Xataka: https://www.xataka.com/robotica-e-ia/las-redes-neuronales-que-son-y-por-que-estan-volviendo

Kourentzes, N. (10 de Febrero de 2017). *Forecasting time series with neural networks in R*. Obtenido de Forecasting Research: http://kourentzes.com/forecasting/2017/02/10/forecasting-time-series-with-neural-networks-in-r/

Legorreta, D. (27 de Abril de 2015). *Construyendo un modelo ARMA para series de tiempo*. Obtenido de DLEGORRETA: https://dlegorreta.wordpress.com/2015/04/27/construyendo-un-modelo-arma-para-series-de-tiempo/

Nishida, K. (12 de Abril de 2017). *An Introduction to Time Series Forecasting with Prophet Package in Exploratory*. Obtenido de Exploratory Blog: https://blog.exploratory.io/an-introduction-to-time-series-forecasting-with-prophet-package-in-exploratory-129ed0c12112

Pant, N. (7 de Septiembre de 2017). *A Guide For Time Series Prediction Using Recurrent Neural Networks (LSTMs)*. Obtenido de Stats and Bots: https://blog.statsbot.co/time-series-prediction-using-recurrent-neural-networks-lstms-807fa6ca7f

Salazar López, B. (s.f.). *Suavización Exponencial Simple*. Obtenido de Ingeniería Industrial Online: https://www.ingenieriaindustrialonline.com/herramientas-para-el-ingeniero-industrial/pron%C3%B3stico-de-ventas/suavizaci%C3%B3n-exponencial-simple/

Santana, A. &. (s.f.). *Objetos en R: Series temporales*. Obtenido de R4ULPGC: Introducción a R: http://www.dma.ulpgc.es/profesores/personal/stat/cursoR4ULPGC/14-seriesTemporales.html

Singh, G. (8 de Febrero de 2018). *7 methods to perform Time Series forecasting (with Python codes)*. Obtenido de Analytics Vidhya: https://www.analyticsvidhya.com/blog/2018/02/time-series-forecasting-methods/

Srivastava, T. (2015, Diciembre 16). *A Complete Tutorial on Time Series Modeling in R*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling/

Taylor, S. J. (23 de Abril de 2017). *Prophet: forecasting at scale*. Obtenido de Facebook Research: https://research.fb.com/prophet-forecasting-at-scale/

Tornero, J. (27 de Abril de 2017). *Introducción al Forecasting con R Statistics*. Obtenido de Doctor Metrics: http://www.doctormetrics.com/2017/04/27/introduccion-al-forecasting-con-r-statistics/#.W23M-LgnaUl

Trubetskoy, G. (17 de Febrero de 2016). *Holt-Winters Forecasting for Dummies*. Obtenido de Notes to self: https://grisha.org/blog/2016/02/17/triple-exponential-smoothing-forecasting-part-iii/