

UNIVERSIDAD DE COSTA RICA
SISTEMA DE ESTUDIOS DE POSGRADO

PRACTICA I

**CALIBRACIÓN DE SERIES DE TIEMPO REFERENTES A ESTADÍSTICAS EN SALUD DE LA
CAJA COSTARRICENSE DEL SEGURO SOCIAL UTILIZANDO MINERÍA DE DATOS EN R**

PRACTICA II

**DESARROLLO DE UNA HERRAMIENTA WEB PARA CALIBRACIÓN DE SERIES DE TIEMPO
BAJO UN ENFOQUE DE MINERÍA DE DATOS**

**Trabajo final de investigación aplicada sometido a la consideración de la Comisión del
Programa de Estudios de Posgrado en Estadística para optar por el grado y título de
Maestría Profesional en Estadística**

GABRIEL ARTURO CORDERO MORA

Ciudad Universitaria Rodrigo Facio, Costa Rica

2017

DEDICATORIA

A:

Dios, por darme la oportunidad de llegar a este momento de mi vida y poder disfrutarlo al máximo, darme salud y por estar conmigo en todo el proceso, por fortalecer mi corazón e iluminar mi mente y por haber puesto en mi camino a aquellas personas que han sido mi soporte y compañía durante todo el periodo de estudio.

Mi madre Maritza Mora, por darme la vida, apoyarme en todo momento, creer en mí y por ser un ejemplo de perseverancia y esfuerzo para salir adelante. Mita gracias por darme una carrera para mi futuro, todo esto se lo debo a usted.

Mi tía María, por ser esa persona amorosa y protectora, que nos acoge a todos con sus brazos abiertos, a usted tía le debo muchísimo, siempre voy a estar en deuda.

Mis hermanos, Berna, Lupe, Cruz y Turito, a cada uno les agradezco por formar parte de mi familia y apoyarme de una u otra forma en distintos momentos de mi vida, los quiero mucho.

Abuela Lourdes, ¡mujer excepcional!! Todo lo que somos hoy es gracias a usted y este título se lo dedico de todo corazón.

Mi esposa, Lupita Jiménez, mi compañera de vida y mi motor de vida, siempre tuvo las palabras justas para darme el impulso necesario para seguir adelante, esto también se lo dedico a usted y espero que sea bendición en nuestra familia.

A mis amigos, compañeros de trabajo, demás familiares, les agradezco compartir conmigo mis buenos y malos momentos y saber que siempre conté con su apoyo.

AGRADECIMIENTO

Mi principal agradecimiento es para Dios, que me ha dado la fortaleza para mantenerme enfocado en busca de este objetivo a pesar de todas las dificultades.

A toda mi familia, esposa, madre, padre, hermanos, mi abuela, en fin todos aquellos a los que he tenido que sacrificarles el tiempo que merecían, porque a pesar de todo, siempre me apoyaron y me impulsaron a seguir adelante.

A mi tutor, Ricardo Alvarado, por todo su apoyo, orientación y oportunidades que me abrió con el desarrollo de esta investigación. Solo me queda decir que como profesional lo admiro muchísimo y le agradezco me diera la oportunidad de tenerlo como tutor, es todo un orgullo para mí.

A uno de mis lectores, Gilbert Brenes, que siempre estuvo disponible para ayudarme, que siempre ha sido además de mi profesor, una excelente persona, muchas gracias profe.

A mis lectores Sandra Hernández y Juan Antonio Rodríguez, les agradezco todo lo que aportaron a mi investigación, ambos son grandes profesionales.

Por último, a esta gran institución que hace muchos años me abrió las puertas, cuando apenas era un adolescente y ahora me ve salir como todo un profesional. Me siento 100% orgulloso de ser UCR.

Este trabajo final de investigación aplicada I fue aceptado por la Comisión del Programa de Estudios de Posgrado en Estadística de la Universidad de Costa Rica, como requisito parcial para optar al grado y título de Maestría Profesional en Estadística.

M.Sc. Ricardo Alvarado Barrantes

Profesor Guía

Ph.D. Gilbert Brenes Camacho

Lector

M.Sc. Juan Antonio Rodríguez Álvarez

Lector

Gabriel Arturo Cordero Mora

Sustentante

Este trabajo final de investigación aplicada II fue aceptado por la Comisión del Programa de Estudios de Posgrado en Estadística de la Universidad de Costa Rica, como requisito parcial para optar al grado y título de Maestría Profesional en Estadística.

M.Sc. Ricardo Alvarado Barrantes

Profesor Guía

Ph.D. Gilbert Brenes Camacho

Lector

M.Sc. Sandra Hernández Rojas

Lectora

Gabriel Arturo Cordero Mora

Sustentante

TABLA DE CONTENIDO

UNIVERSIDAD DE COSTA RICA	i
DEDICATORIA	ii
AGRADECIMIENTO	iii
TABLA DE CONTENIDO	vi
LISTA DE TABLAS	vii
LISTA DE GRÁFICOS	viii
LISTA DE FIGURAS	viii
PRÁCTICA PROFESIONAL I	1
RESUMEN	1
I. INTRODUCCION	2
II. VARIABLES Y METODOS	3
2.1 Variables.....	3
2.1.1 Variables de Consulta Externa	3
2.1.2 Variables de hospitalizaciones	4
2.2 Métodos	4
2.2.1 Métodos de suavización exponencial	4
2.2.2 Modelos ARIMA (enfoque Box-Jenkins).....	6
III. METODOLOGIA	6
3.1 Creación de la base de datos	6
3.2 Calibración de los modelos	7
3.3 Selección del mejor modelo: Holt-Winters vrs ARIMA	8
IV. RESULTADOS	10
V. DISCUSION.....	16
VI. ANEXOS	17
6.1 Sintaxis en R	17
6.2 Paneles de estimaciones.....	23
VII. Bibliografía	30
PRÁCTICA PROFESIONAL II	31
RESUMEN	31

VIII.	INTRODUCCIÓN.....	32
IX.	VARIABLES Y MÉTODOS	33
9.1	Variables.....	33
9.1.1	Variables de Consulta Externa	33
9.1.2	Variables de hospitalizaciones	34
9.2	Métodos de estimación	34
9.2.1	Métodos de suavización exponencial	35
9.2.2	Modelos ARIMA (enfoque Box-Jenkins).....	36
9.3	Desarrollo de aplicación con Shiny	38
9.3.1	Calibración de los modelos	41
9.3.2	Selección del mejor modelo: Holt-Winters vs. ARIMA.....	42
9.3.3	Construcción de una interfaz para el usuario final	43
X.	RESULTADOS	45
10.1	Análisis descriptivo de las series de tiempo.....	45
10.2	Exploración de resultados: Holt-Winters -ARIMA.....	46
10.3	Generación de pronósticos	48
XI.	DISCUSION.....	50
XII.	ANEXOS	52
12.1	Manual de uso.....	52
12.2	Sintaxis en R	60
12.2.1	Archivo ui.R	60
12.2.2	Archivo server.R	63
12.3	Costo de licencias Shiny	70
XIII.	BIBLIOGRAFÍA.....	71

LISTA DE TABLAS

Tabla 1. Muestra del archivo de datos de las series temporales bajo análisis.	7
Tabla 2: Comparativo de las medidas de ajuste y estabilidad de las predicciones.	13
Tabla 3. Muestra del archivo de datos de las series temporales bajo análisis.	41
Tabla 4. Costos de implementación de las distintas modalidades de licencias de Shiny.	70

LISTA DE GRÁFICOS

Gráfico 1. Consultas brindadas según tipo de consulta, CCSS 2007-2012.....	11
Gráfico 2. Horas programadas y horas utilizadas, CCSS 2007-2012.	11
Gráfico 3. Atenciones hospitalarias dadas según servicio, CCSS 2007-2012.	12
Gráfico 4. Análisis de la serie de tiempo y corrección de valores extremos y valores faltantes.	45
Gráfico 5. Prueba de la normalidad de los residuos.	46

LISTA DE FIGURAS

Figura 1. Ventanas de observación y de desempeño.	8
Figura 2. Panel de estimación para la cantidad de consultas subsecuentes. CCSS (2007-2012).	14
Figura 3. Panel comparativo de la estimación de la cantidad de consultas de primera vez en el año. CCSS (2007-2012).....	23
Figura 4. Panel comparativo de la estimación de la cantidad de consultas de primera vez en la especialidad. CCSS (2007-2012).....	23
Figura 5. Panel comparativo de la estimación de la cantidad de consultas de primera vez en la vida. CCSS (2007-2012).....	24
Figura 6. Panel comparativo de la estimación de la cantidad de consultas subsecuentes. CCSS (2007-2012).....	24
Figura 7. Panel comparativo de la estimación de la cantidad de horas programadas. CCSS (2007-2012).	25
Figura 8. Panel comparativo de la estimación de la cantidad de horas utilizadas. CCSS (2007-2012).	25
Figura 9. Panel comparativo de la estimación de la cantidad total de consultas. CCSS (2007-2012).	26
Figura 10. Panel comparativo de la estimación de la cantidad de hospitalizaciones en el servicio de cirugía. CCSS (2007-2012).	26
Figura 11. Panel comparativo de la estimación de la cantidad de hospitalizaciones en el servicio de ginecología. CCSS (2007-2012).....	27
Figura 12. Panel comparativo de la estimación de la cantidad de hospitalizaciones en el servicio de medicina. CCSS (2007-2012)	27
Figura 13. Panel comparativo de la estimación de la cantidad de hospitalizaciones en el servicio de neonatología. CCSS (2007-2012).....	28
Figura 14. Panel comparativo de la estimación de la cantidad de hospitalizaciones en el servicio de obstetricia. CCSS (2007-2012)....	28
Figura 15. Panel comparativo de la estimación de la cantidad de hospitalizaciones en el servicio de pediatría. CCSS (2007-2012).....	29
Figura 16. Panel comparativo de la estimación de la cantidad de hospitalizaciones en el servicio de psiquiatría. CCSS (2007-2012).....	29
Figura 17. Ejemplo de una interfaz creada con la librería de Shiny.....	40
Figura 18. Ventana de observación y de desempeño para el presente estudio.....	42

Figura 19. Panel principal de la herramienta.....	44
Figura 20. Panel de estimación utilizando Suavizado Exponencial.....	47
Figura 21. Panel de estimación utilizando ARIMA.....	48
Figura 22. Barra indicativa de avance.....	49
Figura 23. Panel de generación de los pronósticos finales.....	49

PRÁCTICA PROFESIONAL I

CALIBRACIÓN DE SERIES DE TIEMPO REFERENTES A ESTADÍSTICAS EN SALUD DE LA CAJA COSTARRICENSE DEL SEGURO SOCIAL UTILIZANDO MINERÍA DE DATOS EN R

Gabriel Cordero Mora

RESUMEN

La presente investigación tiene como objetivo generar una herramienta de modelación automática de series de tiempo, que sea capaz de generar pronósticos de las atenciones en salud brindadas en la Caja Costarricense del Seguro Social (CCSS). En total se analizan 14 variables que provienen del sitio web de esta institución, por lo que son datos de carácter público y todas hacen referencia a las atenciones médicas que se dan a nivel general en los distintos centros de salud.

Las técnicas que se evalúan son los modelos de Holt-Winters y ARIMA, para ello se calibra un modelo de cada uno y se compara uno contra el otro para determinar cuál de los dos genera estimaciones con un menor error. Para cada una de las 14 variables se genera un modelo con parámetros específicos que minimizan el error.

Para tomar la decisión de qué modelo es mejor se calculan varios índices de ajuste entre los cuales se puede mencionar el error relativo (ER), error cuadrático medio (ECM) y, por último, las medidas de error absoluto porcentual promedio, desviación media absoluta y la desviación media al cuadrado, MAPE, MAD y MSD respectivamente por sus siglas en inglés, y a partir de estas se crea un nuevo índice que evalúa estas medidas en conjunto para tomar una decisión.

Una vez seleccionado el mejor modelo se procede a generar los pronósticos y sus respectivos intervalos de confianza para cada una de las 14 variables.

I. INTRODUCCION

La presente investigación tiene como objetivo generar una herramienta de modelación automática de series de tiempo, para la estimación de las atenciones en salud brindadas en la Caja Costarricense del Seguro Social (CCSS). Las variables bajo análisis serán 14, 7 de ellas correspondientes a la consulta externa: total de consultas, consultas subsecuentes, consultas de primera vez en el año, consultas de primera vez en la especialidad, consultas de primera vez en la vida, horas programadas y horas utilizadas; y 7 variables correspondientes a la cantidad de hospitalizaciones según servicio: cirugía, ginecología, medicina, neonatología, obstetricia y pediatría. Estas variables corresponden a una agregación de la información a nivel de toda la institución, pero la herramienta tendrá la capacidad de estimar estas mismas variables a niveles más específicos, como por ejemplo a nivel de un hospital, área de salud o EBAIS.

Hoy en día los análisis de series de tiempo son de gran importancia para las distintas empresas, organizaciones o instituciones, cualquiera de ellas requiere conocer el comportamiento futuro de ciertos fenómenos con el fin de planificar y prevenir ciertas situaciones, para ello lo que se debe hacer es predecir lo que ocurrirá con una variable en el futuro a partir del comportamiento de esa variable en el pasado. En este caso conocer el volumen de consultas que se dará en un periodo, permitirá a las autoridades de la CCSS contar con el personal necesario para abastecer toda la demanda de servicios en salud y a su vez presupuestar los costos asociados a estas consultas.

Una de las problemáticas que enfrentan muchas empresas en la actualidad, entre ellas la CCSS, es que no cuentan con el suficiente personal capacitado ni con el tiempo para generar este tipo de análisis, o bien no se cuenta con los recursos para adquirir una herramienta de modelado automático de series temporales que pueda generar rápidamente estas estimaciones con una precisión aceptable.

Es por esta razón, que desarrollar una herramienta en un código abierto y gratuito como R puede ser de gran utilidad, pues este es capaz de realizar estimaciones automáticas de manera que ocupe una intervención mínima de un usuario, el cual no precisará que cuente con conocimientos especializados en el tema de series de tiempo ni programación, si no únicamente en el campo de la información en salud, para que sea capaz de validar la consistencia de los resultados.

Hoy en día, los algoritmos de predicción automática más populares se basan en modelos ARIMA o de suavizado exponencial (Holt-Winters), es por esta razón que esta investigación se enfocará principalmente en estas dos opciones.

Actualmente ya existen programas que incluyen paquetes de modelación automática, sin embargo, por el elevado costo que suelen tener estas herramientas, no todas las compañías

pueden contar con ellas, de ahí que se justifique aún más la utilización de software libres como R.

II. VARIABLES Y METODOS

2.1 Variables

Los datos utilizados para esta investigación provienen del sitio web de la CCSS, por lo que son datos de carácter público.

Esta información hace referencia principalmente a las atenciones médicas que se dan a nivel general en los distintos centros de salud, o sea es un dato acumulado, pero como se mencionó antes, la herramienta está en capacidad de estimar modelos a niveles más desagregados como hospitalares, áreas de salud o EBAIS.

En total se estarán analizando 14 variables y son las siguientes:

2.1.1 Variables de Consulta Externa

- **Consultas de primera vez en la vida (CPVV):** Atención brindada en la consulta externa a un paciente que nunca ha sido atendido dentro de la CCSS en esta modalidad.
- **Consultas de primera vez en el año (CPVA):** Atención brindada en la consulta externa a un paciente que no ha sido atendido en esta modalidad durante el año en curso, pero si en años anteriores.
- **Consultas de primera vez en la especialidad (CPVE):** Atención brindada en la consulta externa a un paciente que ya ha sido atendido en esta modalidad durante el año en curso y que es su primera vez en ese año para la especialidad en específico en la que fue atendido.
- **Consultas subsecuentes (CS):** Atención brindada en la consulta externa y que no cumple ninguna de las condiciones de los tipos de consultas anteriores.
- **Total de consultas:** Total de atenciones brindadas en la consulta externa y se calcula como la suma de CPVV, CPVA, CPVE y CS.
- **Horas programadas:** Total de horas programadas de los médicos para atender la consulta externa.
- **Horas utilizadas:** Total de horas utilizadas por los médicos para atender la consulta externa.

2.1.2 Variables de hospitalizaciones

- **Atenciones hospitalarias:** Son 7 variables y cada una hace referencia a las atenciones dadas a personas hospitalizadas en los servicios de cirugía, ginecología, medicina general, neonatología, obstetricia, pediatría y psiquiatría.

Todas estas variables tienen una periodicidad mensual, corresponden a una medición acumulada a fin de mes y los datos disponibles son de enero del 2007 a diciembre del 2012 (6 años, 72 observaciones).

2.2 Métodos

A continuación, se mencionan las técnicas que se estarán utilizando para realizar las estimaciones. Es importante recalcar que no se profundizará en cada método, únicamente se indican algunas características y parte del planteamiento matemático, pues el objetivo principal de esta investigación no es profundizar en la parte teórica de las mismas, de ahí que se omite en este documento la explicación de algunos conceptos básicos en el análisis de series de tiempo como por ejemplo los componentes de tendencia, ciclicidad, estacionalidad y por último, el de aleatoriedad o bien aspectos aún más específicos de cada técnica.

2.2.1 Métodos de suavización exponencial

Estos métodos tienen como ventaja un bajo costo y sencillez al momento de aplicarlos, además puede ser utilizado a pesar de que se disponga de una pequeña cantidad de observaciones.

Entre los métodos más comunes, se puede mencionar el de **Suavización exponencial simple**, que como menciona Hernández “es apropiado para series que no tienen patrones estacionales ni de tendencia y cuya media o nivel cambia lentamente”.

Para este caso el pronóstico se puede denotar de la siguiente forma:

$$P_{t+1} = P_t + \alpha(Z_t - P_t)$$

Donde Z, es la variable que se desea predecir, P es el pronóstico y α es una constante.

Ahora bien, si la serie presenta un patrón lineal de tendencia, el método anterior pierde validez y se hace necesario optar por métodos más complejos, en este caso el **método lineal de Holt** viene a solventar en cierta medida la limitante del método de Suavización exponencial simple.

Este método utiliza 3 ecuaciones y 2 parámetros, α y β que toman valores entre 0 y 1. Las ecuaciones son las siguientes:

$$\begin{aligned} a_t &= \alpha Z_t + (1 - \alpha)(a_{t-1} + b_{t-1}) \\ b_t &= \beta(a_t - a_{t-1}) + (1 - \beta)b_{t-1} \\ P_{t+m} &= a_t + b_t m \end{aligned}$$

El valor de a_t es una estimación del nivel promedio de la serie en el tiempo t, mientras que b_t es una estimación de la pendiente de la serie Z en el tiempo t.

En el caso de P_{t+m} es la ecuación que permite pronosticar el valor de Z, m periodos más adelante del tiempo t, sumando m veces el valor de la pendiente estimada en el tiempo t al nivel de la serie en el tiempo t.

Por su parte, los valores de α y β se obtienen al minimizar la suma de cuadrados de los errores de pronóstico para todos los valores posibles de α y β en el intervalo (0,1).

Un tercer método que viene a mejorar el anterior es el desarrollado por Winters (1960) para tomar en cuenta además del componente de tendencia, la estacionalidad, naciendo así el **método de Holt-Winters** en el cual se pueden encontrar dos variantes, el multiplicativo y el aditivo. Para la presente investigación se estará utilizando este método en su variante aditiva.

La diferencia básica entre uno y otro es que en el modelo aditivo se considera que la variable observada se puede descomponer en la suma de los factores (tendencia, estacionalidad, ciclicidad y aleatoriedad), mientras en el multiplicativo, el comportamiento de la variable observada se expresa como el producto de los componentes anteriores.

El método multiplicativo se denota de la siguiente manera:

$$\begin{aligned} a_t &= \alpha \frac{Z_t}{S_{t-s}} + (1 - \alpha)(a_{t-1} + b_{t-1}) & S_t &= \gamma \frac{Z_t}{a_t} + (1 - \gamma)S_{t-s} \\ b_t &= \beta(a_t - a_{t-1}) + (1 - \beta)b_{t-1} & P_{t+m} &= (a_t + b_t m)S_{t-s+m} \end{aligned}$$

El método aditivo se denota de la siguiente manera:

$$\begin{aligned} a_t &= \alpha(Z_t - S_{t-s}) + (1 - \alpha)(a_{t-1} + b_{t-1}) \\ b_t &= \beta(a_t - a_{t-1}) + (1 - \beta)b_{t-1} \\ S_t &= \gamma(Z_t - a_t) + (1 - \gamma)S_{t-s} \\ P_{t+m} &= a_t + b_t m + S_{t-s+m} \end{aligned}$$

2.2.2 Modelos ARIMA (enfoque Box-Jenkins)

Este modelo usa la notación $ARIMA(p, d, q)$ para representar un modelo autoregresivo integrado de promedios móviles para Z_t , donde p indica el orden de la parte autorregresiva del modelo, d indica el número de veces que se ha diferenciado Z_t para hacer estacionaria su media y q especifica el orden de la parte de promedios móviles del modelo.

En este tipo de modelos, el parámetro d permite obtener una serie estacionaria cuando esta no lo es, dado que se aplica las diferenciaciones consecutivas de orden d para la serie Z_t . En cuanto al parámetro p , este representa la cantidad de rezagos, mientras que q representa la cantidad de datos utilizados para calcular los promedios móviles.

Estos modelos son adecuados para series temporales que son estacionarias. Se dice que una serie es estacionaria si cumple las siguientes condiciones:

- La serie Z_t debe tener una media que toma el mismo valor sea cual sea el valor de t .
- Debe tener una varianza constante para cada valor de t .
- La correlación entre valores de la serie en t y $t-k$, Z_t y Z_{t-1} , vale lo mismo para k dada toda t .

Es importante en este punto mencionar que la mayoría de las series temporales no son estacionarias, sin embargo, se pueden convertir fácilmente en series estacionarias, calculando la diferenciación consecutiva de sus valores y de ser necesario, se aplica una segunda diferenciación en caso de que la primera no haya logrado hacer de la serie, una serie estacionaria.

III. METODOLOGIA

El fin práctico de este trabajo es la estimación automática de pronósticos para las series de tiempo de la CCSS, y para poder lograrlo fue necesario seguir una serie de pasos que se detallan a continuación.

3.1 Creación de la base de datos

Para el análisis de las series de tiempo, se decidió estructurar un archivo de datos donde cada registro corresponde al dato mensual y en las columnas se van a encontrar las distintas variables (ver Figura 1). En este caso se tiene un archivo de datos con 14 variables, pero la herramienta está diseñada para aceptar una matriz con n variables a estimar.

Tabla 1. Muestra del archivo de datos de las series temporales bajo análisis.

Fecha	PVA	PVE	PPV	CS	HP	HU	TC	Cirugía	Ginecología	Medicina	Neonatología	Obstetricia	Pediatria	Psiquiatría
31/01/2007	664794	25458	21433	100559	266752	242874	812244	4897	2010	4041	1053	7426	2503	392
28/02/2007	449829	51957	22137	254286	257733	234775	778209	4929	1942	3929	982	6542	2388	405
31/03/2007	395182	63628	23444	387668	282457	258041	869922	5733	2321	4284	1137	7318	2771	465
30/04/2007	241198	50096	19302	404983	233790	211703	715579	4971	2026	3814	1052	6915	2545	377
31/05/2007	249955	63763	25301	542136	283601	258782	881155	5699	2229	4204	1093	7255	2627	468
30/06/2007	206362	62294	22769	552398	268676	244524	843823	5656	2102	4422	1087	7261	2632	445
31/07/2007	171654	54571	24173	579470	262320	238086	829868	5388	2088	4925	1139	7517	2541	415
31/08/2007	150536	59812	23502	616021	272125	247408	849871	5732	2051	5140	1105	7848	2688	438
30/09/2007	134509	56503	21842	601282	258556	234693	814136	5879	2135	4720	1121	8178	3027	371
31/10/2007	135031	63663	23991	664997	286230	258579	887682	5758	2255	4726	1170	8070	3324	406
30/11/2007	117614	59319	22644	652125	280794	252175	851702	5832	2245	4461	1222	7861	3165	418
31/12/2007	83840	39767	16133	528247	228773	203575	667987	4993	1964	4284	1182	7874	2614	392
31/01/2008	660591	23329	20847	96068	266006	241457	800835	5093	1965	3843	1037	7368	2409	379
29/02/2008	460135	50772	26119	267841	271204	245815	804867	5983	2208	3810	1094	6919	2458	367
31/03/2008	350878	55167	19732	342939	253534	230744	768716	5783	2077	3813	1073	7197	2525	413
30/04/2008	294263	67586	22579	476124	280805	255634	860552	5774	2124	3805	1095	7263	2562	464
31/05/2008	242291	66557	23225	522561	279457	255032	854634	5788	2085	3989	1149	7541	2661	418
30/06/2008	217935	67377	23107	566336	279305	254489	874755	5494	2102	3891	1084	7251	2492	369
31/07/2008	185301	60337	22553	607192	283562	258754	875383	5386	2144	4306	1111	7615	2519	442
31/08/2008	149145	59501	21224	593089	266341	242061	822959	5372	2168	4166	1150	7979	2678	452
30/09/2008	145779	59895	24603	645941	283262	257861	876218	5576	2020	4077	1228	8321	2933	417

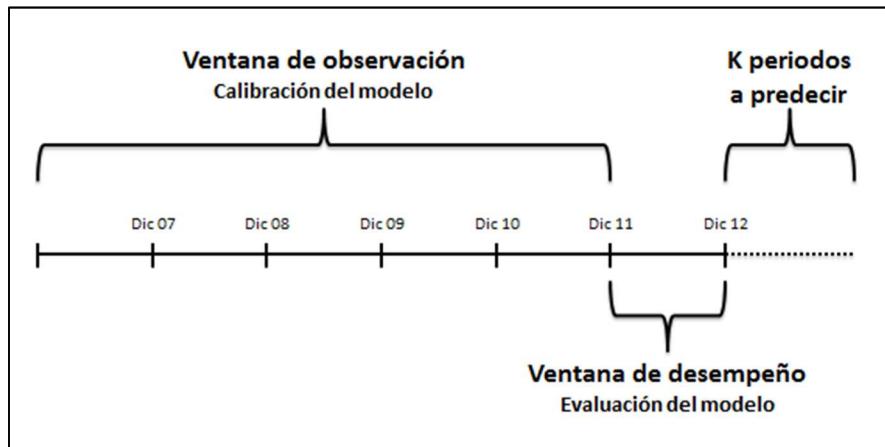
En este caso no era estrictamente necesario que la base de datos tuviera esta estructura, cada fila pudo representar una variable y las columnas las fechas, sin embargo, esto hubiera requerido que la herramienta realizara un trabajo adicional para transponer la base y lograr la estructura anterior, que es la que R va a reconocer para manejar los datos como una serie de tiempo.

3.2 Calibración de los modelos

Para calibrar el mejor modelo de Holt-Winters y ARIMA, se hace uso de la herramienta de programación R donde se crea un proceso capaz de detectar el mejor modelo para cada una de las técnicas.

Para iniciar la exploración del mejor modelo, se divide la información en una ventana de observación (enero 2007 a diciembre 2011) y otra para medir el desempeño del modelo (enero 2012 a diciembre 2012). En la ventana de observación se calibra ambos métodos y en la ventana de desempeño se mide el ajuste y calidad de las estimaciones.

Figura 1. Ventanas de observación y de desempeño.



En el caso del modelo ARIMA, se hace uso de la función `auto.arima` que pertenece a R y fue desarrollada por Rob J. Hyndman en el 2008. Esta función tiene la capacidad de devolver el mejor modelo ARIMA según alguno de los criterios de información siguientes: AIC (default), BIC AICC.

Por otra parte, para el modelo de Holt-Winters se utilizó una función que genera un proceso iterativo en el cual se optimiza los valores para los parámetros alpha, beta y gamma, que minimizan el error cuadrático medio del modelo. En este proceso se calculan un total de 900 modelos diferentes para las posibles combinaciones de los parámetros ($\text{alpha}=0.1, 0.2, \dots, 1$ – alpha no puede ser cero--; $\text{beta}=0.0, 0.1, 0.2, \dots, 1$; $\text{gamma}=0.0, 0.1, 0.2, \dots, 1$).

Estas dos funciones permiten obtener el modelo que genera menor error tanto para Holt-Winters como para ARIMA.

3.3 Selección del mejor modelo: Holt-Winters vrs ARIMA

Una vez que se tienen los dos mejores modelos, se deben poner a competir entre sí para seleccionar el que tenga asociado un error más pequeño. Para ello se utiliza una serie de medidas de ajuste, entre los cuales se puede mencionar el error relativo (ER), error cuadrático medio (ECM) y por último, las medidas de error absoluto porcentual promedio, desviación media absoluta y la desviación media al cuadrado, MAPE, MAD y MSD respectivamente por sus siglas en inglés.

Adicionalmente, se obtuvo el porcentaje de fallos hacia arriba (PFAr), porcentaje de fallos hacia abajo (PFAb), porcentaje total de fallos hacia arriba (PTFAr), porcentaje total de fallos hacia abajo (PTFAb), que ayudan a tener una idea rápida de si las estimaciones están subestimando o sobreestimando los valores reales (PFAr y PFAb) y en qué magnitud (PTFAr y PTFAb).

Las fórmulas de cálculo para cada una de las medidas anteriores son las siguientes:

Medidas para evaluar el ajuste

$$ER = \frac{\sum_{i=1}^N |R_i - P_i|}{\sum_{i=1}^N |R_i|} * 100$$

$$MAD = \frac{\sum_{i=1}^N |R_i - P_i|}{N}$$

$$ECM = \frac{1}{N} \sum_{i=1}^N (R_i - P_i)^2$$

$$MSD = \frac{\sum_{i=1}^N |R_i - P_i|^2}{N}$$

$$MAPE = \frac{\sum_{i=1}^N \left| \frac{(R_i - P_i)}{P_i} \right|}{N} * 100$$

Donde

N : Número de datos pronosticados

R_i : Valor real en el tiempo i

P_i : Valor pronosticado en el tiempo i

Medidas para evaluar la estabilidad de las estimaciones

$$PFAr = \sum_{i=1}^A \frac{1}{N}$$

$$PTFAr = \frac{\sum_{i=1}^A (R_i - P_i)}{\sum_{i=1}^A |R_i|}$$

$$PFAb = \sum_{i=1}^B \frac{1}{N}$$

$$PTFAb = \frac{\sum_{i=1}^B (R_i - P_i)}{\sum_{i=1}^B |R_i|}$$

Donde

R_i : Valor real en el tiempo i

P_i : Valor pronosticado en el tiempo i

N : Número de datos pronosticados

A : Número de datos en los que lo real fue mayor al pronóstico

B : Número de datos en los que lo real fue menor al pronóstico

La selección final de método a recomendar se realizó mediante el cálculo de un índice para cada modelo, en el que el modelo que obtiene el valor más pequeño es el elegido.

$$\begin{aligned}
I_{HW} &= \frac{ER_{HW}}{\max(ER_A, ER_{HW})} + \frac{ECM_{HW}}{\max(ECM_A, ECM_{HW})} + \frac{MAPE_{HW}}{\max(MAPE_A, MAPE_{HW})} \\
&\quad + \frac{MAD_{HW}}{\max(MAD_A, MAD_{HW})} + \frac{MSD_{HW}}{\max(MSD_A, MSD_{HW})} + \\
I_{ARIMA} &= \frac{ER_A}{\max(ER_A, ER_{HW})} + \frac{ECM_A}{\max(ECM_A, ECM_{HW})} + \frac{MAPE_A}{\max(MAPE_A, MAPE_{HW})} \\
&\quad + \frac{MAD_A}{\max(MAD_A, MAD_{HW})} + \frac{MSD_A}{\max(MSD_A, MSD_{HW})} +
\end{aligned}$$

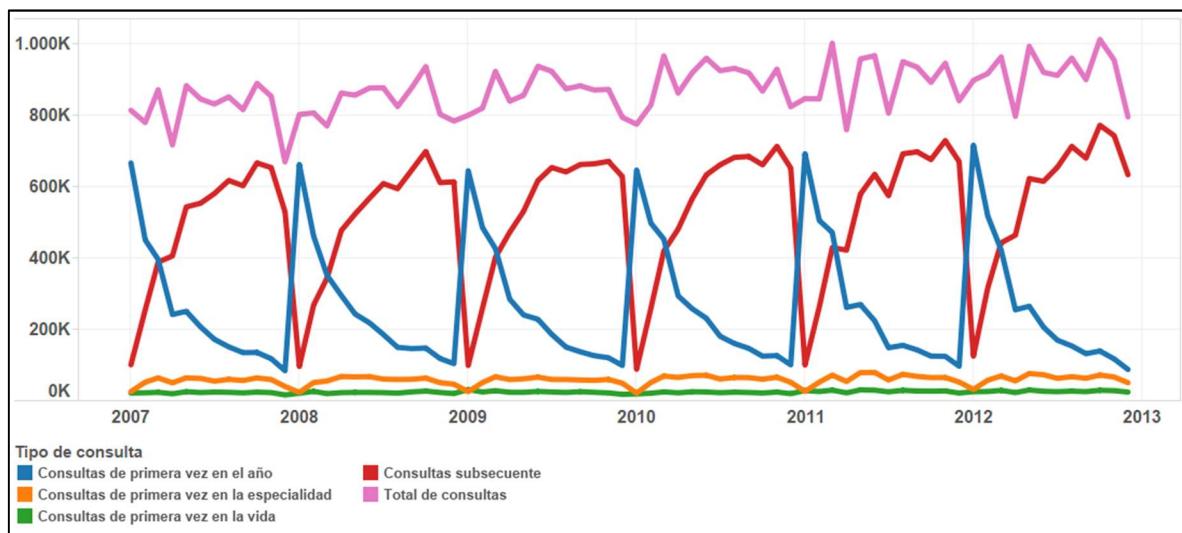
Debe quedar claro que el índice anterior sirve como criterio para dar una recomendación del mejor modelo, sin embargo, ambos modelos podrían ser satisfactorios, por lo que queda a criterio del usuario cual tiene mayor funcionalidad práctica y consistencia según lo observado en los pronósticos generados.

IV. RESULTADOS

Para iniciar se realizó un análisis previo de las variables, las cuales se clasificaron en 3 grupos según la naturaleza de estas, el primer grupo son las atenciones en Consulta Externa (gráfico 1), el segundo grupo incluye las horas programadas y utilizadas para realizar dichas atenciones (gráfico 2) y por último, las variables referentes a las hospitalizaciones (gráfico 3).

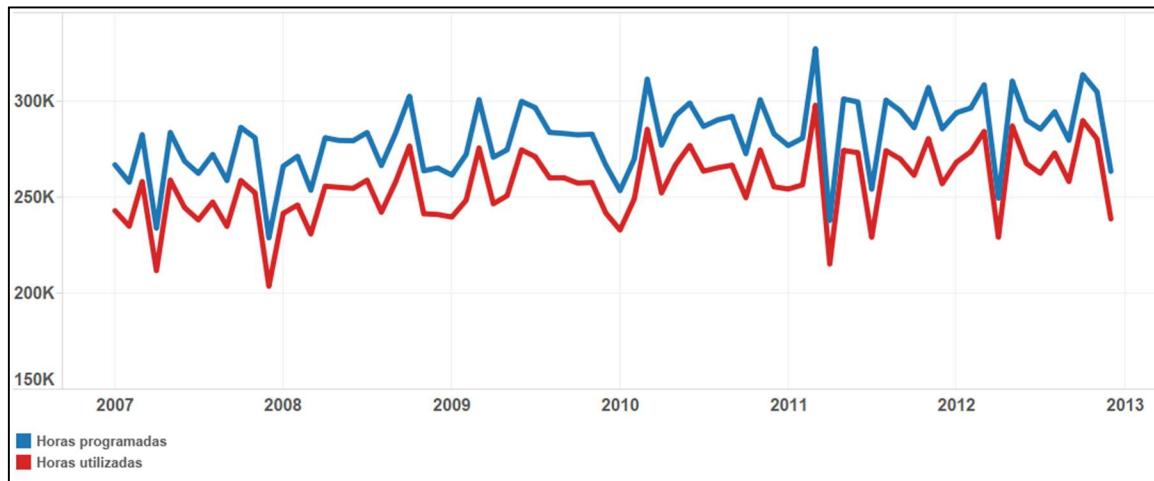
Este análisis permitió encontrar patrones bastante particulares, por ejemplo en el gráfico 1, se puede observar como las consultas de primera vez en el año (valores altos a inicios de año y decrece rápidamente conforme avanzan los meses), primera vez en la especialidad (valores bajos en enero, crece y se estabiliza) y consultas subsecuentes (valores bajos al inicio del año y crece rápidamente conforme avanzan los meses), tienen una estacionalidad bastante marcada, mientras que el total de consultas y las consultas de primera vez en la vida tienen comportamientos más estables a través del tiempo.

Gráfico 1. Consultas brindadas según tipo de consulta, CCSS 2007-2012.



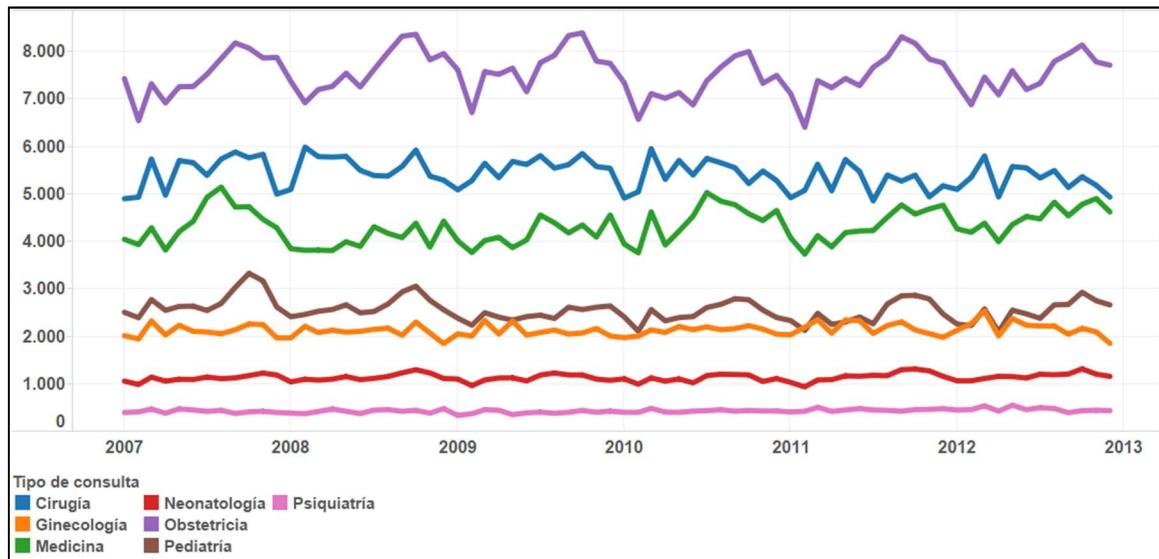
Similar es el caso de las horas programadas y las horas utilizadas, que como se puede observar en el gráfico 2 parecen ser bastante estables, solamente existe una pequeña estacionalidad en ciertos meses y también una leve tendencia creciente.

Gráfico 2. Horas programadas y horas utilizadas, CCSS 2007-2012.



En cuanto a las atenciones hospitalarias, se puede notar como aparentemente no hay una tendencia de crecimiento, lo que seguramente obedezca a la falta de ampliación de los servicios hospitalarios de la CCSS. Con respecto a la estacionalidad, se puede observar como febrero suele ser el mes más bajo en la mayoría de las series (se debe en cierta medida a que tiene menos días), mientras que octubre y setiembre todo lo contrario, son la época de mayor concurrencia en estos servicios. Esto se nota muy claramente en el servicio de obstetricia.

Gráfico 3. Atenciones hospitalarias dadas según servicio, CCSS 2007-2012.



Como se puede notar en las series anteriores, ninguna de ellas tuvo cambios abruptos que haga pensar en la necesidad de utilizar variables de intervención, ni tampoco se observan valores atípicos, que puedan afectar las estimaciones, esto se debe principalmente a que los datos utilizados en este caso corresponden a la totalidad de la información de la CCSS, por lo que posibles problemas a nivel de un centro de salud en específico, podrían estar siendo disimulados dentro de toda la vista general.

Por esta razón, los resultados obtenidos por la herramienta, a pesar de ser totalmente válidos a niveles de mayor desagregación como por ejemplo un EBAIS, deben validarse para comprobar la consistencia entre datos-modelo y modelo-realidad, dado que el presente modelo no cuenta con métodos que controlen el efecto de posibles valores extremos o bien algún cambio brusco en el comportamiento de las series.

Terminado este análisis inicial, se procede a realizar la corrida para obtener el mejor modelo posible para cada técnica y posteriormente hacer la comparación entre ambos para recomendar el modelo más idóneo.

Los resultados obtenidos se resumen en la siguiente tabla, como se puede observar en todas las variables, los resultados obtenidos por el modelo de Holt-Winters superaron a los modelos ARIMA. Este resultado es muy interesante y podría justificar en próximos estudios, intentar construir una función que genere resultados más satisfactorios que auto.arima, para darle mayor peso a esta técnica.

Tabla 2: Comparativo de las medidas de ajuste y estabilidad de las predicciones.

VARIABLE	METODO	Medidas de ajuste					Medidas de estabilidad				DECISIÓN
		ECM	ER	MAPE	MAD	MSD ('000)	PFAr	PFab	PTFAr	PTFab	
PVA	Holt-Winters	17,578.3	5.4%	0.1	14,166.3	308,997.2	50.0%	50.0%	6.2%	4.8%	Holt-Winters
	ARIMA	19,097.2	5.6%	0.1	14,871.8	364,704.3	66.7%	33.3%	6.5%	4.7%	
PVE	Holt-Winters	3,315.2	4.2%	0.0	2,602.6	10,990.3	58.3%	41.7%	3.6%	5.1%	Holt-Winters
	ARIMA	3,893.7	5.4%	0.1	3,350.9	15,161.2	50.0%	50.0%	4.4%	6.6%	
PVV	Holt-Winters	1,272.3	4.2%	0.0	1,110.0	1,618.7	58.3%	41.7%	3.8%	4.7%	Holt-Winters
	ARIMA	2,764.1	8.8%	0.1	2,338.3	7,640.5	0.0%	100.0%	Nan	8.8%	
CS	Holt-Winters	25,516.8	3.6%	0.0	20,564.6	651,105.6	50.0%	50.0%	4.1%	3.3%	Holt-Winters
	ARIMA	46,026.7	6.8%	0.1	38,441.3	2,118,458.7	25.0%	75.0%	3.8%	8.0%	
HP	Holt-Winters	13,377.8	3.8%	0.0	11,067.1	178,966.7	41.7%	58.3%	5.4%	2.8%	Holt-Winters
	ARIMA	17,644.2	5.4%	0.1	15,610.8	311,318.6	50.0%	50.0%	4.3%	6.5%	
HU	Holt-Winters	12,385.7	4.0%	0.0	10,822.0	153,404.3	33.3%	66.7%	5.3%	3.5%	Holt-Winters
	ARIMA	16,995.8	5.3%	0.1	14,217.4	288,858.6	50.0%	50.0%	3.2%	7.5%	
TC	Holt-Winters	43,304.8	4.0%	0.0	36,995.8	1,875,304.8	41.7%	58.3%	4.8%	3.5%	Holt-Winters
	ARIMA	60,381.4	5.6%	0.1	50,889.8	3,645,913.6	33.3%	66.7%	4.7%	6.0%	
Cirugía	Holt-Winters	193.2	3.2%	0.0	170.7	37.3	50.0%	50.0%	3.1%	3.4%	Holt-Winters
	ARIMA	309.6	4.9%	0.0	258.6	95.9	8.3%	91.7%	0.2%	5.3%	
Ginecología	Holt-Winters	104.7	4.0%	0.0	87.8	11.0	25.0%	75.0%	5.5%	3.6%	Holt-Winters
	ARIMA	150.2	5.4%	0.1	118.1	22.6	33.3%	66.7%	5.9%	5.2%	
Medicina	Holt-Winters	152.8	3.0%	0.0	132.6	23.4	50.0%	50.0%	3.1%	2.8%	Holt-Winters
	ARIMA	183.1	3.0%	0.0	133.0	33.5	50.0%	50.0%	4.4%	1.6%	
Neonatología	Holt-Winters	34.1	2.7%	0.0	30.9	1.2	83.3%	16.7%	2.8%	2.2%	Holt-Winters
	ARIMA	44.6	2.9%	0.0	34.1	2.0	16.7%	83.3%	3.1%	2.9%	
Obstetricia	Holt-Winters	126.8	1.5%	0.0	110.8	16.1	41.7%	58.3%	1.6%	1.4%	Holt-Winters
	ARIMA	159.5	1.8%	0.0	134.9	25.4	41.7%	58.3%	1.8%	1.8%	
Pediatría	Holt-Winters	107.7	3.5%	0.0	88.0	11.6	41.7%	58.3%	4.6%	2.7%	Holt-Winters
	ARIMA	118.3	3.9%	0.0	97.9	14.0	33.3%	66.7%	5.1%	3.3%	
Psiquiatría	Holt-Winters	36.2	6.4%	0.1	29.6	1.3	41.7%	58.3%	5.0%	7.3%	Holt-Winters
	ARIMA	43.2	7.3%	0.1	33.5	1.9	50.0%	50.0%	5.1%	9.2%	

Nota:

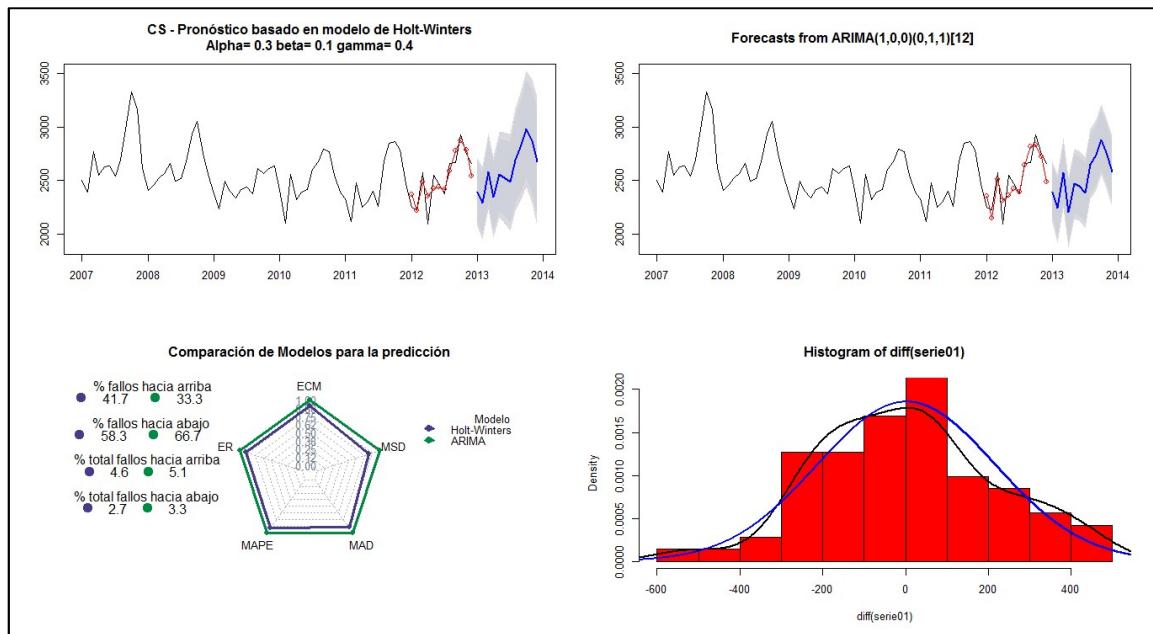
- El ECM, ER, MAPE, MAD y MSD busca que sus valores sean pequeños.
- El PFAr y PFAB busca que los valores ronden el 50%, esto significaría que las estimaciones no están siendo subestimadas ni sobreestimadas.
- El PTFAr y PTFAb busca que sus valores sean pequeños.

Es importante resaltar que, para las variables relacionadas a la consulta externa, las medidas de ajuste ECM, MAD y MSD, tiene valores bastante altos con relación a las variables de hospitalización y horas, pero esto se debe especialmente por el volumen de cada variable.

En lo que respecta a la decisión de seleccionar el modelo de Holt-Winters o ARIMA, la decisión se tomó con base en las medidas de ajuste, para ello se calculan dos índices (I_A y I_{HW}) y se evalúa cual modelo da los mejores resultados.

Ahora bien, como se ha venido mencionando, la herramienta da una recomendación del modelo a utilizar, sin embargo, queda a criterio del usuario si utiliza el modelo recomendado o el alternativo, para ello se crea la siguiente visualización, la cual le permitirá analizar la consistencia de los pronósticos para ambas técnicas (existe un panel como el siguiente para cada variable, ver anexos).

Figura 2. Panel de estimación para la cantidad de consultas subsecuentes. CCSS (2007-2012).



En la gráfica de la parte superior izquierda está la serie correspondiente a la cantidad de hospitalizaciones en el servicio de pediatría entre enero del 2007 y diciembre del 2012, además, se muestra una línea de color rojo que representa el valor estimado con el método de Holt-Winters para la ventana de desempeño (enero 2012 a diciembre 2012), la cual permitió evaluar el ajuste de lo estimado contra lo real, por último, se puede observar una línea en color azul que representa el pronóstico (enero 2013 a diciembre 2013) y bandas de confianza del 90% y 95% para dichos pronósticos (estas bandas son parametrizables, se podría indicar niveles de confianza mayor si se prefiere). En la parte superior derecha se encuentra el pronóstico para el modelo de ARIMA.

La gráfica de la parte inferior izquierda muestra un comparativo para cada modelo, según los distintos índices de ajuste, el modelo que tenga valores más pequeños será el mejor, en este caso el mejor modelo es el de Holt-Winters. Adicionalmente, a la izquierda se muestra los indicadores de PFAr y PFAb, estos índices permiten verificar si el modelo está subestimando o sobreestimando los valores reales, un valor de 50% indicaría que las estimaciones están distribuidas por encima y por debajo del valor real de forma uniforme, en este caso, los modelos ARIMA tiene un 67% de fallos hacia abajo, por encima del 58% que tiene Holt-Winters, por lo que está subestimando en mayor medida los valores reales de la serie.

En cuanto a los indicadores de PTFAr y PTFAb, estos intentan mostrar la magnitud que tienen las diferencias entre lo pronosticado y el valor real, ya sea que el fallo fue hacia abajo o hacia arriba, en este caso la magnitud de los errores fue levemente mejor para el modelo de Holt-Winters.

El gráfico de la parte inferior derecha permite verificar visualmente la normalidad de los residuos. El tamaño de las barras se obtiene a partir de los errores en la estimación, la línea azul es la distribución teórica de una normal y la negra una estimación de la distribución de los errores. En este caso, la distribución de los residuos parece acercarse al valor teórico, por lo que se podría afirmar que no hay problemas con la evaluación de este supuesto.

Para observar los resultados de las demás variables, en la sección de anexos se encuentran un panel para comparar la estimación de cada una de las variables bajo análisis.

V. DISCUSION

En esta sección se menciona los principales logros y conclusiones que se ha obtenido durante el desarrollo de la investigación para tener claro el aporte de este trabajo, además, las limitaciones que se han identificado y que podría ser de interés analizarlas para que en un futuro se puedan mejorar la funcionalidad y confiabilidad de la herramienta.

- Se obtienen pronósticos para las 14 variables bajo análisis, reduciendo al máximo posible el error asociado a las estimaciones, además, se generan intervalos de confianza para dichos pronósticos. Los resultados son bastante consistentes con el comportamiento observado y la herramienta tiene la capacidad de generar estimaciones automáticas para matrices mayores.
- Se obtuvo evidencia clara que el modelo de Holt-Winters ajusta mejor que un ARIMA las series analizadas, queda pendiente comprobar si esto se repite en un ejercicio con otras variables.
- La función auto.arima parece no tener la suficiente capacidad de competir contra la función utilizada para calcular iterativamente modelos de Holt-Winters, por lo que podría ser interesante intentar desarrollar una función para modelos ARIMA que pueda obtener mejores resultados.
- Los modelos programados en la herramienta tienen la limitación de no controlar eventos extraños en la serie (valores extremos), como por ejemplo el cierre de consultorios, lo que provocaría una disminución importante en la cantidad de consultas, o bien lo contrario, la ampliación de un centro de salud. Una solución a esta posible situación podría ser agregar una nueva funcionalidad a la herramienta, que permita identificar esta situación y sea capaz de agregar por ejemplo variables de intervención para modelar estas eventualidades.
- En la CCSS existe una gran cantidad de series de tiempo que podrían ser estimadas, sin embargo, la herramienta actual a pesar de poder realizar estimaciones en un conjunto de variables mucho mayor al actual, no está optimizada para estimar grandes cantidades de forma simultánea (“Big Data”), por lo que sería valioso lograr que la herramienta realice estas estimaciones con procesos más óptimos, como por ejemplo programación en paralelo, que básicamente lo que permitiría es distribuir el trabajo de la estimación en los diferentes procesadores de un computador, lo que lo hace mucho más eficiente y por ende brindaría una capacidad mucho mayor de procesamiento.

VI. ANEXOS

6.1 Sintaxis en R

```

suppressWarnings(suppressMessages(library(itsmr)))
suppressWarnings(suppressMessages(library(forecast)))
suppressMessages(library(FactoMineR))
suppressWarnings(suppressMessages(library(dtw)))
suppressWarnings(suppressMessages(library(rattle)))
suppressWarnings(suppressMessages(library(fmsb)))
suppressWarnings(suppressMessages(library(plotrix)))

ER <- function(Pron, Real) {
  return(sum(abs(Pron - Real))/abs(sum(Real)))
}

# mean squared error (MSE)
ECM <- function(Pron, Real) {
  N <- length(Real)
  ss <- sum((Real - Pron)^2)
  return((1/N) * ss)
}

# Error absoluto porcentual promedio (MAPE)
MAPE <- function(Pron, Real) {
  N <- length(Real)
  sMAPE <- sum(abs(Real - Pron)/Real)
  return((1/N) * sMAPE)
}

# Desviación media absoluta (MAD)
MAD <- function(Pron, Real) {
  N <- length(Real)
  sMAD <- sum(abs(Real - Pron))
  return((1/N) * sMAD)
}

# Desviación media al cuadrado (MSD)
MSD <- function(Pron, Real) {
  N <- length(Real)
  sMSD <- sum(abs(Real - Pron)^2)
  return((1/N) * sMSD)
}

#Porcentaje de fallos arriba
PFArriba <- function(Pron, Real) {
  Total <- 0
  N <- length(Pron)
  for (i in 1:N) {
    if (Pron[i] >= Real[i])
      Total <- Total + 1
  }
}

```

```

}

return(Total/N)
}

#Porcentaje de fallos abajo
PFAabajo <- function(Pron, Real) {
  Total <- 0
  N <- length(Pron)
  for (i in 1:N) {
    if (Pron[i] < Real[i])
      Total <- Total + 1
  }
  return(Total/N)
}

#Porcentaje total de fallos arriba
PTFArriba <- function(Pron, Real) {
  Total <- 0
  SReal <- 0
  N <- length(Pron)
  for (i in 1:N) {
    if (Pron[i] >= Real[i]) {
      Total <- Total + (Pron[i] - Real[i])
      SReal <- SReal + abs(Real[i])
    }
  }
  return(Total/SReal)
}

#Porcentaje total de fallos abajo
PTFAabajo <- function(Pron, Real) {
  Total <- 0
  SReal <- 0
  N <- length(Pron)
  for (i in 1:N) {
    if (Pron[i] < Real[i]) {
      Total <- Total + abs(Pron[i] - Real[i])
      SReal <- SReal + abs(Real[i])
    }
  }
  return(Total/SReal)
}

#Función de calibración para los modelos de Holt-Winters
calibrar <- function(serie.aprendizaje, serie.testing) {
  error.c <- Inf
  alpha.i <- 0.1 # alpha no puede ser cero
  while (alpha.i <= 1) {
    beta.i <- 0
    while (beta.i <= 1) {
      gamma.i <- 0
      while (gamma.i <= 1) {
        mod.i <- HoltWinters(serie.aprendizaje, alpha = alpha.i, beta = beta.i, gamma = gamma.i)
      }
    }
  }
}

```

```

res.i <- predict(mod.i, n.ahead = length(serie.testing))
error.i <- sqrt(ECM(res.i, serie.testing))
if (error.i < error.c) {
  error.c <- error.i
  mod.c <- mod.i
}
gamma.i <- gamma.i + 0.1
}
beta.i <- beta.i + 0.1
}
alpha.i <- alpha.i + 0.1
}
return(mod.c)
}

#directorio
setwd("C:/Users/gcorderom/Documents/Capacitaciones/Practica profesional I")

#Lectura de la base de datos
serie <- read.csv("Series de tiempo.csv", header = T, dec = ".", sep = ",")

#PARAMETROS MODIFICABLES PARA LA SERIE
año inicial<-2007
periodo<-12 # cantidad de periodos
periodos_predecir<-12#periodos a utilizar para la ventade de desempeño
n testing<-12
intervalos<-c(90,95) #intervalos de confianza

for(i in 2:dim(serie)[2]){
  serie01 <- ts(serie[,i], start = c(año inicial, 1), freq = periodo)
  n<-length(serie01)
  nVO<-length(serie01)-n testing
  serie01.aprende <- serie01[1:nVO]
  nVD<-length(serie01.aprende)+1
  serie01.aprende <- ts(serie01.aprende, start = año inicial, freq = periodo)
  serie01.test <- serie01[nVD:n]

  p.HW <- calibrar(serie01.aprende, serie01.test)
  modHW <- HoltWinters(serie01.aprende, alpha = p.HW$alpha, beta = p.HW$beta, gamma = p.HW$gamma)
  predHW <- predict(modHW, n.ahead = n testing)

  p.arima<-auto.arima(serie01.aprende)
  p<-p.arima$arma[1]
  d<-p.arima$arma[6]
  q<-p.arima$arma[2]
  P<-p.arima$arma[3]
  D<-p.arima$arma[7]
  Q<-p.arima$arma[4]
  PP<-p.arima$arma[5]
  modArima<-arima(serie01.aprende,order=c(p,d,q),seasonal=list(order=c(P,D,Q),period=PP))
  predArima <- predict(modArima, n.ahead = n testing)

  par(mfrow=c(2,2))
}

```

```

#modelos
modelofinalHW <- HoltWinters(serie01, alpha = p.HW$alpha, beta = p.HW$beta, gamma = p.HW$gamma)
modelofinalArima<-arima(serie01,order=c(p,d,q),seasonal=list(order=c(P,D,Q),period=PP))

MinPronosticos<-min(
  forecast(modelofinalHW,h=periodos_predecir,level=c(intervalos))$lower[,2],
  forecast(modelofinalArima,h=periodos_predecir,level=c(intervalos))$lower[,2])
MaxPronosticos<-max(
  forecast(modelofinalHW,h=periodos_predecir,level=c(intervalos))$upper[,2],
  forecast(modelofinalArima,h=periodos_predecir,level=c(intervalos))$upper[,2])

#pronosticos
plot(forecast(modelofinalHW,h=periodos_predecir,level=c(intervalos)),ylim=c(MinPronosticos,MaxPronosticos),
main=paste(names(serie)[i], "- Pronóstico basado en modelo de Holt-Winters
Alpha=",p.HW$alpha,"beta=",p.HW$beta,"gamma=",p.HW$gamma))
lines(predHW,type="o",col="red")
plot(forecast(modelofinalArima,h=periodos_predecir,level=c(intervalos)),ylim=c(MinPronosticos,MaxPronosticos))
lines(predArima$pred,type="o",col="red")

ECM1<-sqrt(ECM(predHW, serie01.test))
ECM2<-sqrt(ECM(predArima$pred, serie01.test))
ER1<-ER(predHW, serie01.test)
ER2<-ER(predArima$pred, serie01.test)
PFArriba1<-PFArriba(predHW, serie01.test)
PFArriba2<-PFArriba(predArima$pred, serie01.test)
PFAbajo1<-PFAbajo(predHW, serie01.test)
PFAbajo2<-PFAbajo(predArima$pred, serie01.test)
PTFArriba1<-PTFArriba(predHW, serie01.test)
PTFArriba2<-PTFArriba(predArima$pred, serie01.test)
PTFAbajo1<-PTFAbajo(predHW, serie01.test)
PTFAbajo2<-PTFAbajo(predArima$pred, serie01.test)
MAPE1<-MAPE(predHW, serie01.test)
MAPE2<-MAPE(predArima$pred, serie01.test)
MAD1<-MAD(predHW, serie01.test)
MAD2<-MAD(predArima$pred, serie01.test)
MSD1<-MSD(predHW, serie01.test)
MSD2<-MSD(predArima$pred, serie01.test)

# Comparacion
ComparaECM <- rbind(ECM1, ECM2)/max(ECM1, ECM2)
ComparaER <- rbind(ER1, ER2)/max(ER1, ER2)
ComparaPFArriba <- rbind(PFArriba1, PFArriba2)/max(PFArriba1, PFArriba2)
ComparaPFAbajo <- rbind(PFAbajo1, PFAbajo2)/max(PFAbajo1, PFAbajo2)
ComparaPTFArriba <- rbind(PTFArriba1, PTFArriba2)/max(PTFArriba1, PTFArriba2)
ComparaPTFAbajo <- rbind(PTFAbajo1, PTFAbajo2)/max(PTFAbajo1, PTFAbajo2)
ComparaMAPE <- rbind(MAPE1, MAPE2)/max(MAPE1, MAPE2)
ComparaMAD <- rbind(MAD1, MAD2)/max(MAD1, MAD2)
ComparaMSD <- rbind(MSD1, MSD2)/max(MSD1, MSD2)

Compara <- as.data.frame(cbind(ComparaECM,ComparaER,ComparaMAPE,ComparaMAD,ComparaMSD))
Min <- data.frame(t(rep(0, 5)))
Max <- data.frame(t(rep(1, 5)))

```

```

colnames(Compara)<-colnames(Min)<-colnames(Max)<-c("ECM","ER","MAPE","MAD","MSD")
Compara2 <- rbind(Max, Min, Compara)

radarchart(Compara2, maxmin = TRUE, axislabel = "slategray4",
           centerzero = FALSE, seg = 8, cglcol = "gray67", pcol = c("slateblue4", "springgreen4"),
           plty = 1, plwd = 3, title = "Comparación de Modelos para la predicción")

legend(1.5, 1, legend = c("Holt-Winters", "ARIMA"), seg.len = 0.5, title = "Modelo",
       pch = 21, bty = "n", lwd = 3, y.intersp = 0.5, horiz = FALSE, col = c("slateblue4","springgreen4"))

legend(-3.5, 1.5, legend = round(rbind(PFArriba1, PFArriba2)*100,0), seg.len = 0.1, title = "% fallos hacia arriba",
       pch = 21, bty = "n", lwd = 6, y.intersp = 0.5, horiz = TRUE, col = c("slateblue4","springgreen4"))
legend(-3.5, 1.0, legend = round(rbind(PFAbajo1, PFAbajo2)*100,0), seg.len = 0.1, title = "% fallos hacia abajo",
       pch = 21, bty = "n", lwd = 6, y.intersp = 0.5, horiz = TRUE, col = c("slateblue4","springgreen4"))
legend(-3.7, 0.5, legend = round(rbind(PTFArriba1, PTFArriba2)*100,0), seg.len = 0, title = "% total fallos hacia arriba",
       pch = 21, bty = "n", lwd = 6, y.intersp = 0.5, horiz = TRUE, col = c("slateblue4","springgreen4"))
legend(-3.7, 0.0, legend = round(rbind(PTFAbajo1, PTFAbajo2)*100,0), seg.len = 0, title = "% total fallos hacia abajo",
       pch = 21, bty = "n", lwd = 6, y.intersp = 0.5, horiz = TRUE, col = c("slateblue4","springgreen4"))

#Histograma
hist(diff(serie01), prob = T, col = "red", ylim=c(0,max(density(diff(serie01))$y)*1.15))
lines(density(diff(serie01)), lwd = 2)
mu <- mean(diff(serie01))
sigma <- sd(diff(serie01))
x <- seq(min(density(diff(serie01))$x), max(density(diff(serie01))$x), length = 100000)
y <- dnorm(x, mu, sigma)
lines(x, y, lwd = 2, col = "blue")

#CREACION DE LA TABLA PARA EXPORTAR LOS PRONOSTICOS
METODO_RECOMENDADO<-if (sum(Compara2[3,])<sum(Compara2[4,])) "HOLT-WINTERS" else "ARIMA"
A1<-data.frame(forecast(modelofinalHW,h=periodos_predecir,level=c(intervalos)))
A<-cbind(rownames(A1),"Holt-Winters",METODO_RECOMENDADO,names(serie)[i],A1)
colnames(A)[1]<-"FECHA"
colnames(A)[2]<-"METODO"
colnames(A)[4]<-"VARIABLE"
B1<-data.frame(forecast(modelofinalArima,h=periodos_predecir,level=c(intervalos)))
B<-cbind(rownames(B1),"ARIMA",METODO_RECOMENDADO,names(serie)[i],B1)
colnames(B)[2]<-"METODO"
colnames(B)[1]<-"FECHA"
colnames(B)[4]<-"VARIABLE"

RESULTADO_I<-rbind(A,B)
RESULTADO<-if (i==2) RESULTADO_I else rbind(RESULTADO,RESULTADO_I)

#CREACION DE LA TABLA PARA EXPORTAR LOS PRONOSTICOS
AJUSTE_I<- cbind(c("Holt-Winters","ARIMA"),names(serie)[i],round(
  cbind(rbind(ECM1,ECM2),rbind(ER1,ER2),rbind(PFArriba1,PFArriba2),rbind(PFAbajo1,PFAbajo2),
         rbind(PTFArriba1,PTFArriba2),rbind(PTFAbajo1,PTFAbajo2),rbind(MAPE1,MAPE2),rbind(MAD1,MAD2),
         rbind(MSD1,MSD2)),3)
)
AJUSTE<-if (i==2) AJUSTE_I else rbind(AJUSTE,AJUSTE_I)

```

```
colnames(AJUSTE)<-c("METODO","VARIABLE","ECM","ER","PFAr","PFAb","PTFAr","PTFAb","MAPE","MAD","MSD")  
}  
  
write.csv(as.data.frame(RESULTADO), file = "PRONOSTICOS.csv", row.names = FALSE)  
write.csv(AJUSTE, file = "AJUSTE.csv", row.names = FALSE)
```

6.2 Paneles de estimaciones

Figura 3. Panel comparativo de la estimación de la cantidad de consultas de primera vez en el año. CCSS (2007-2012).

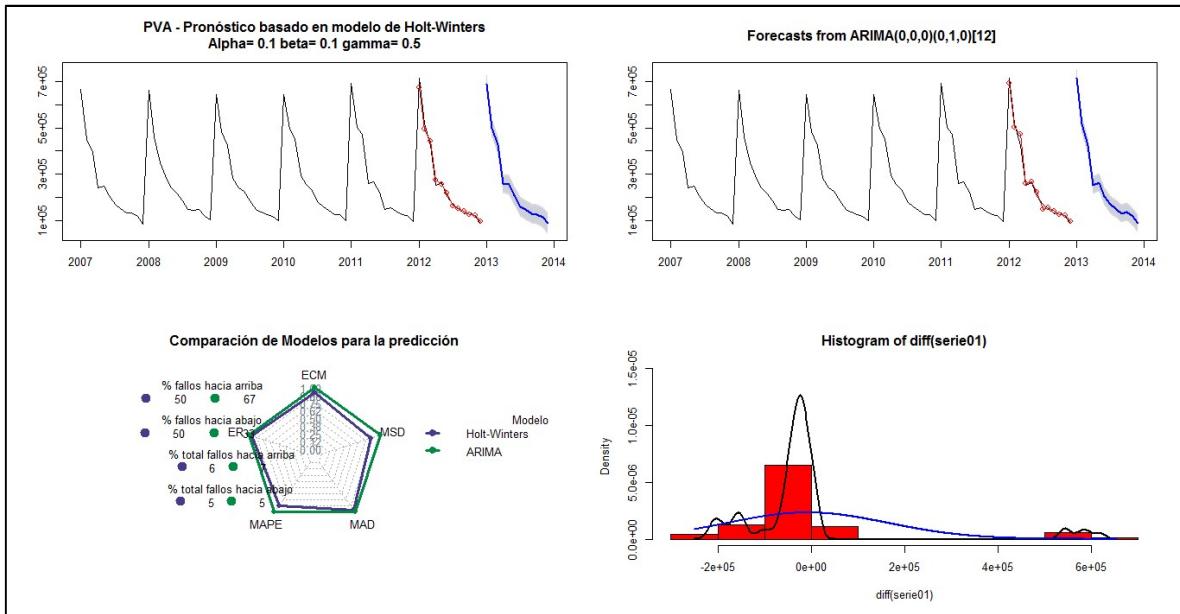


Figura 4. Panel comparativo de la estimación de la cantidad de consultas de primera vez en la especialidad. CCSS (2007-2012).

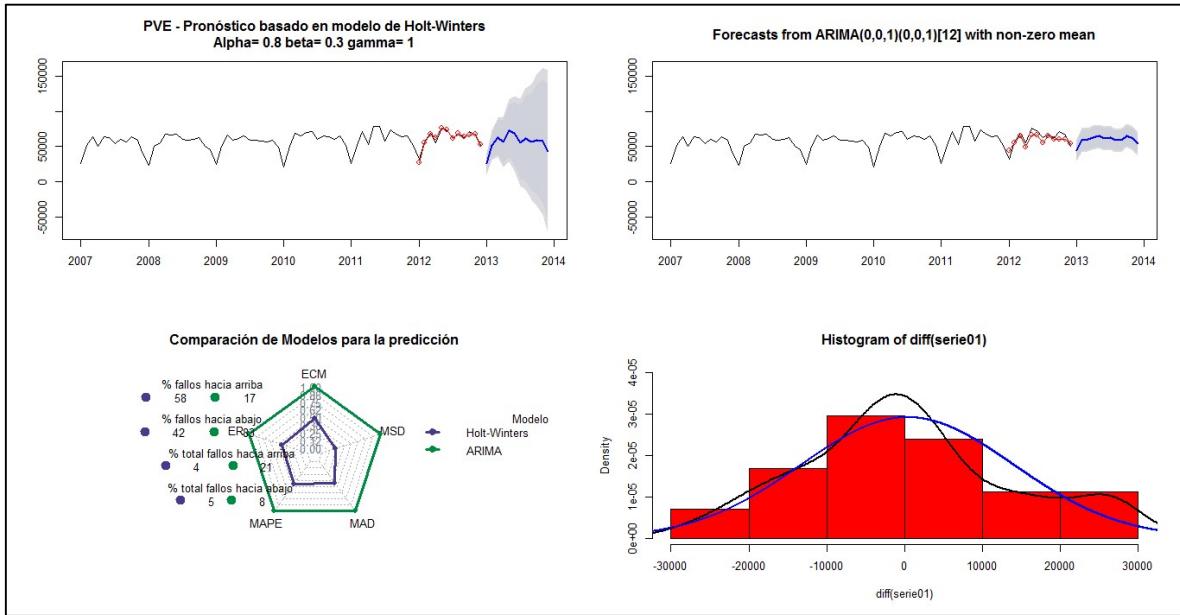


Figura 5. Panel comparativo de la estimación de la cantidad de consultas de primera vez en la vida. CCSS (2007-2012).

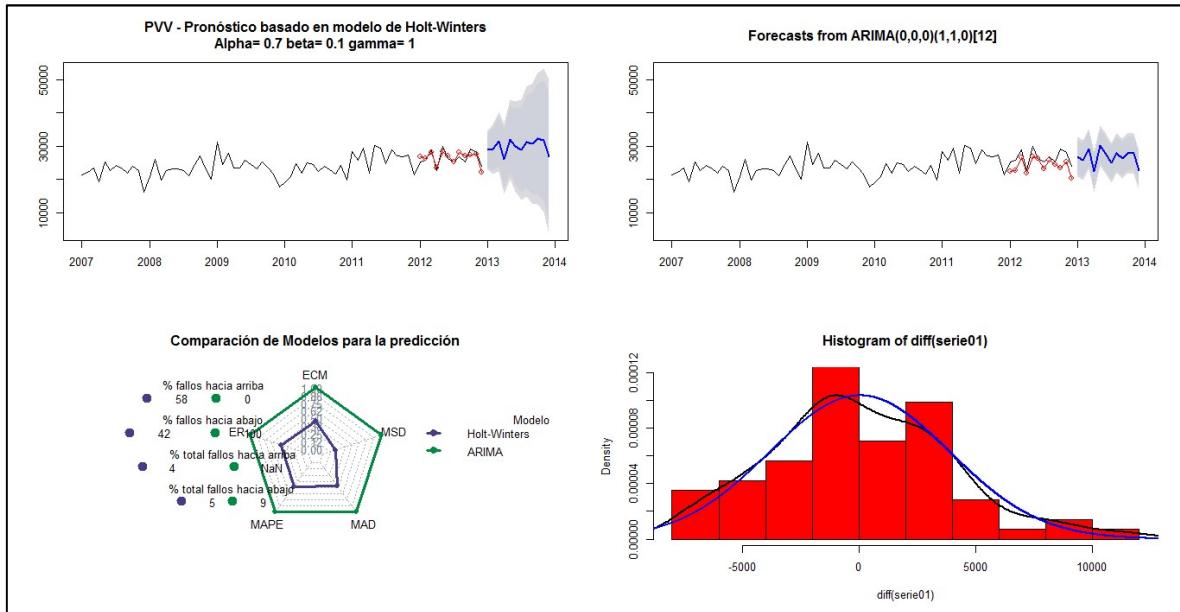


Figura 6. Panel comparativo de la estimación de la cantidad de consultas subsecuentes. CCSS (2007-2012).

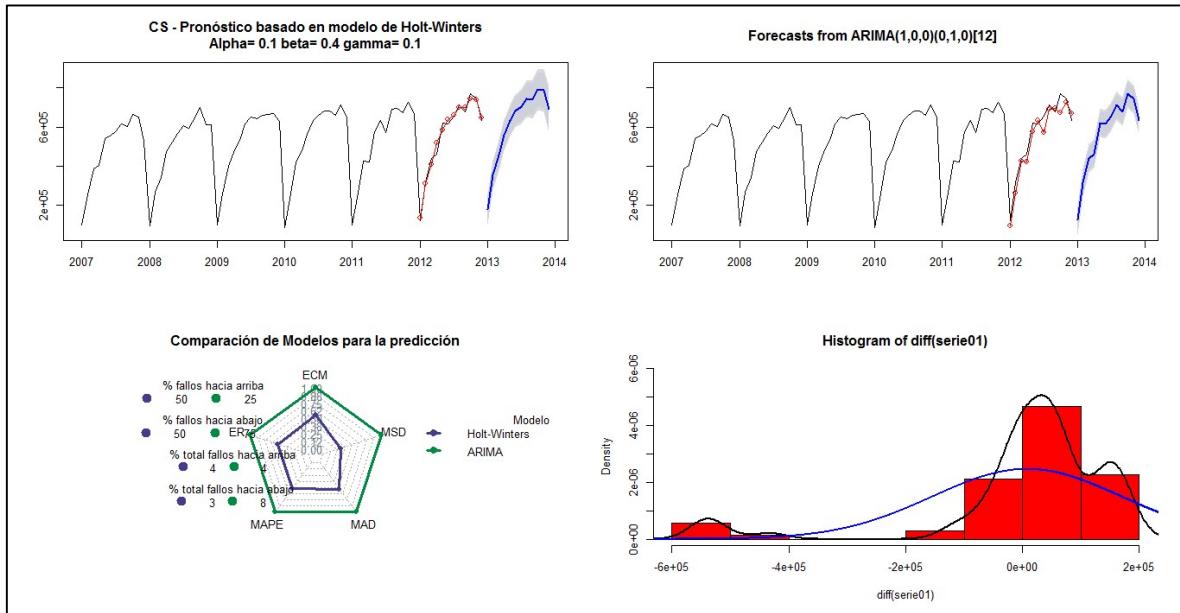


Figura 7. Panel comparativo de la estimación de la cantidad de horas programadas. CCSS (2007-2012).

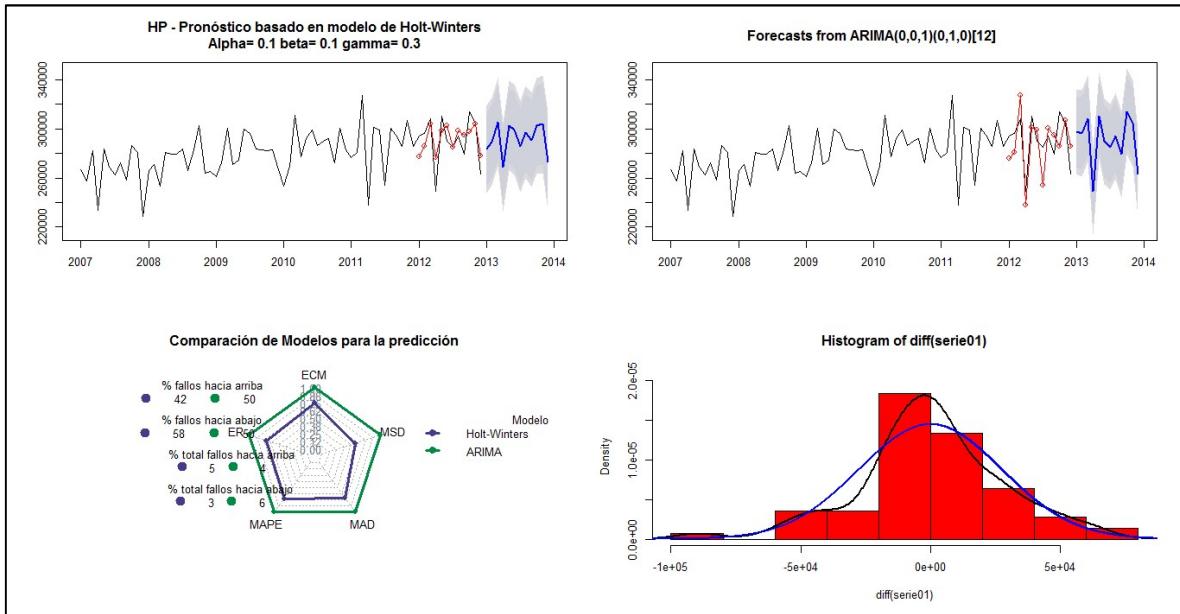


Figura 8. Panel comparativo de la estimación de la cantidad de horas utilizadas. CCSS (2007-2012).

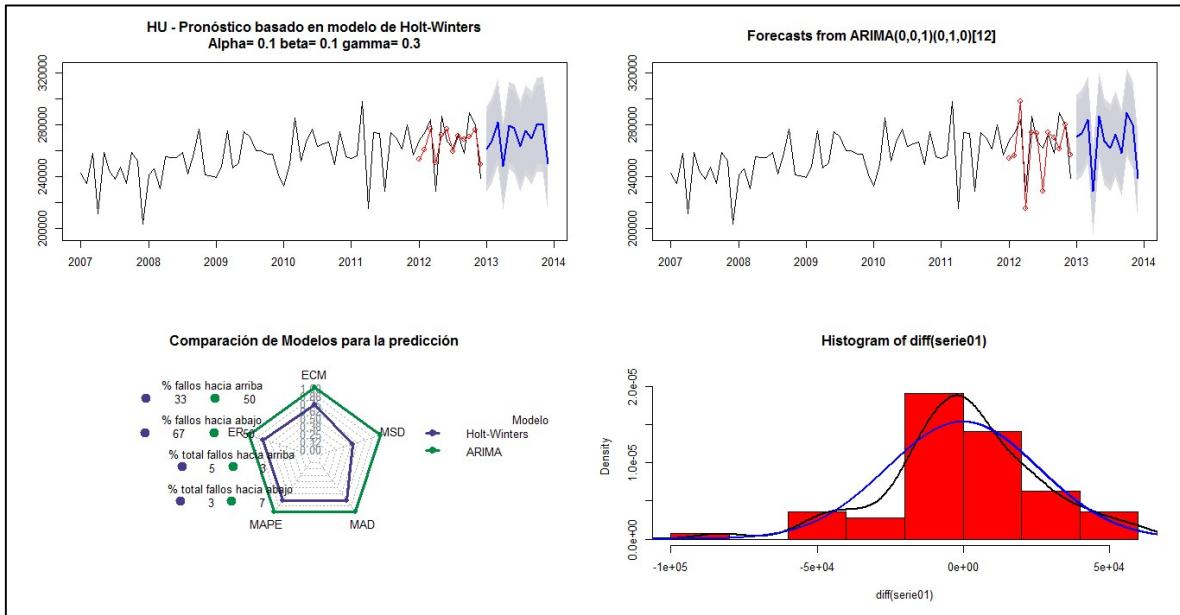


Figura 9. Panel comparativo de la estimación de la cantidad total de consultas. CCSS (2007-2012).

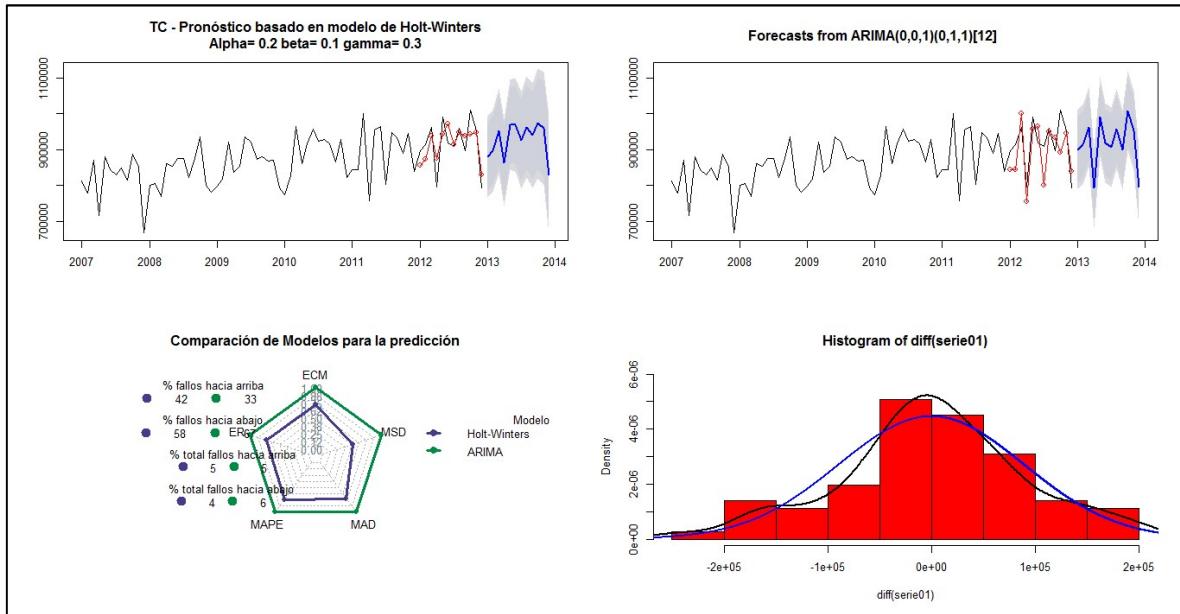


Figura 10. Panel comparativo de la estimación de la cantidad de hospitalizaciones en el servicio de cirugía. CCSS (2007-2012).

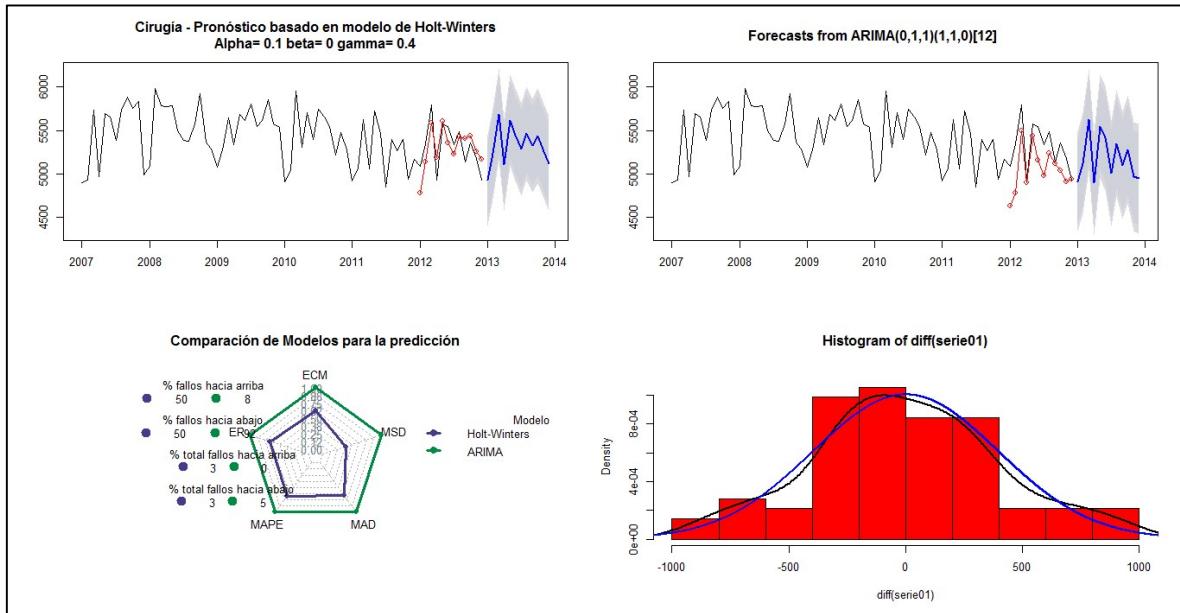


Figura 11. Panel comparativo de la estimación de la cantidad de hospitalizaciones en el servicio de ginecología. CCSS (2007-2012).

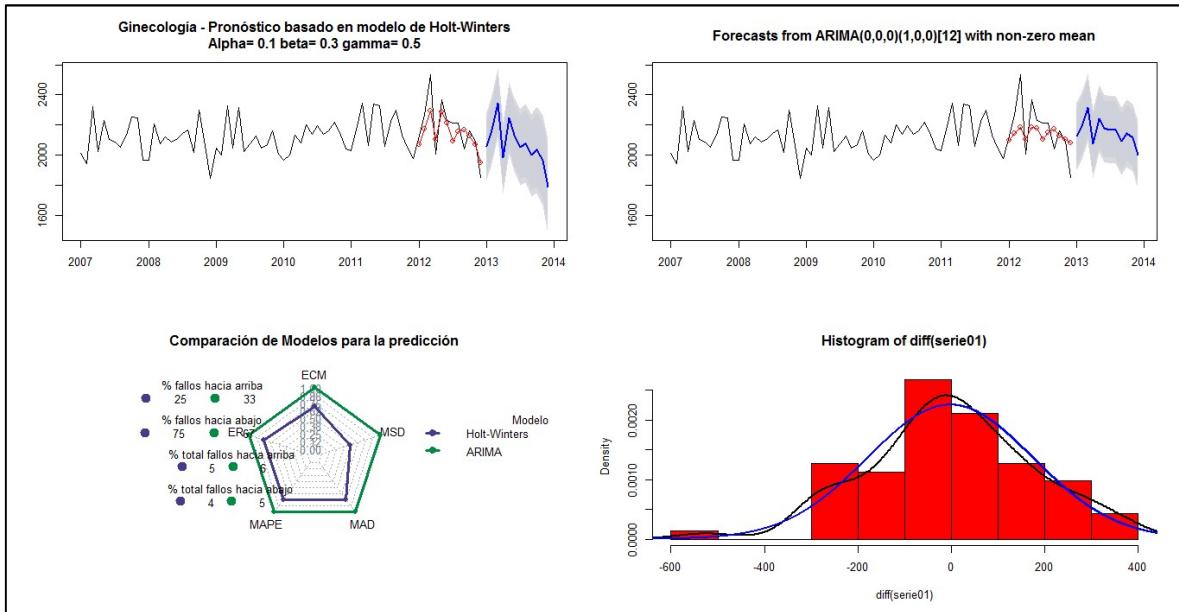


Figura 12. Panel comparativo de la estimación de la cantidad de hospitalizaciones en el servicio de medicina. CCSS (2007-2012).

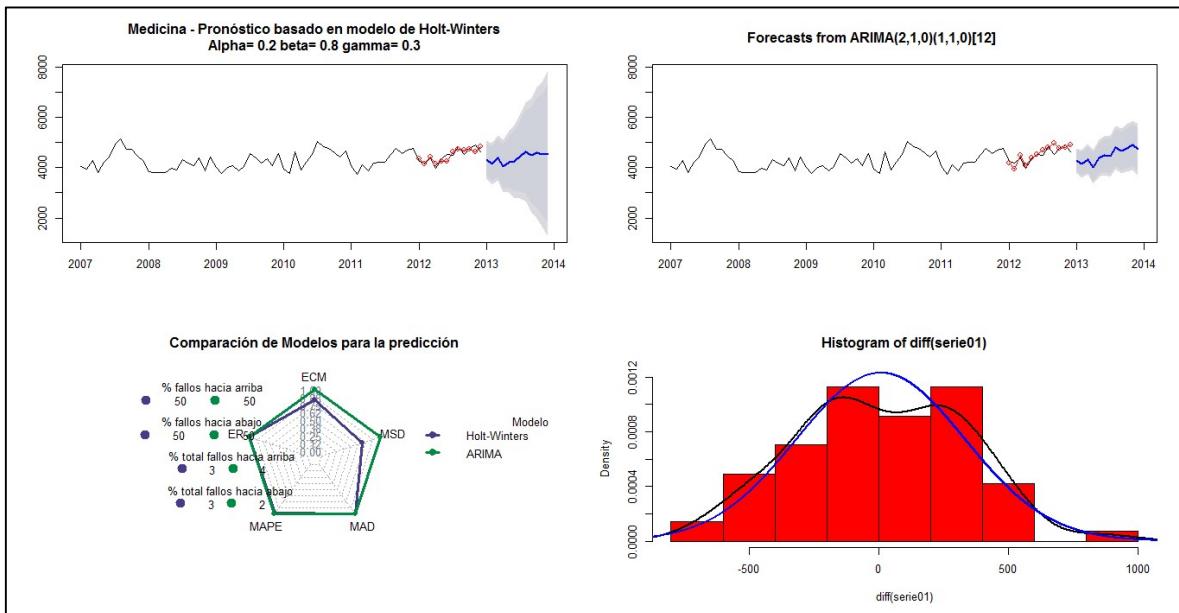


Figura 13. Panel comparativo de la estimación de la cantidad de hospitalizaciones en el servicio de neonatología. CCSS (2007-2012).

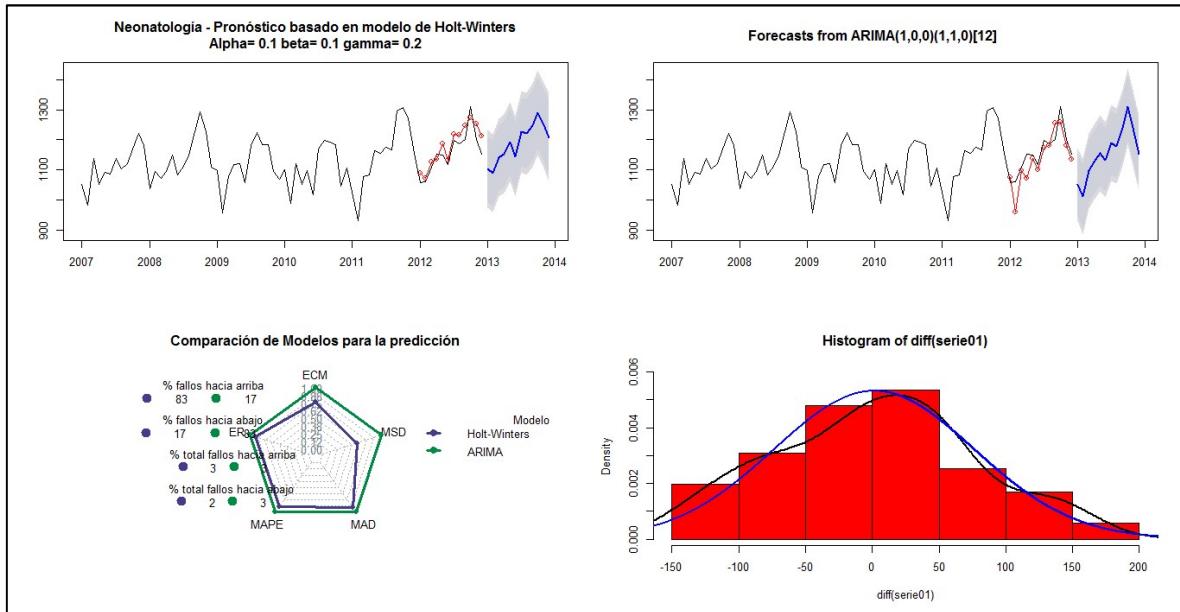


Figura 14. Panel comparativo de la estimación de la cantidad de hospitalizaciones en el servicio de obstetricia. CCSS (2007-2012).

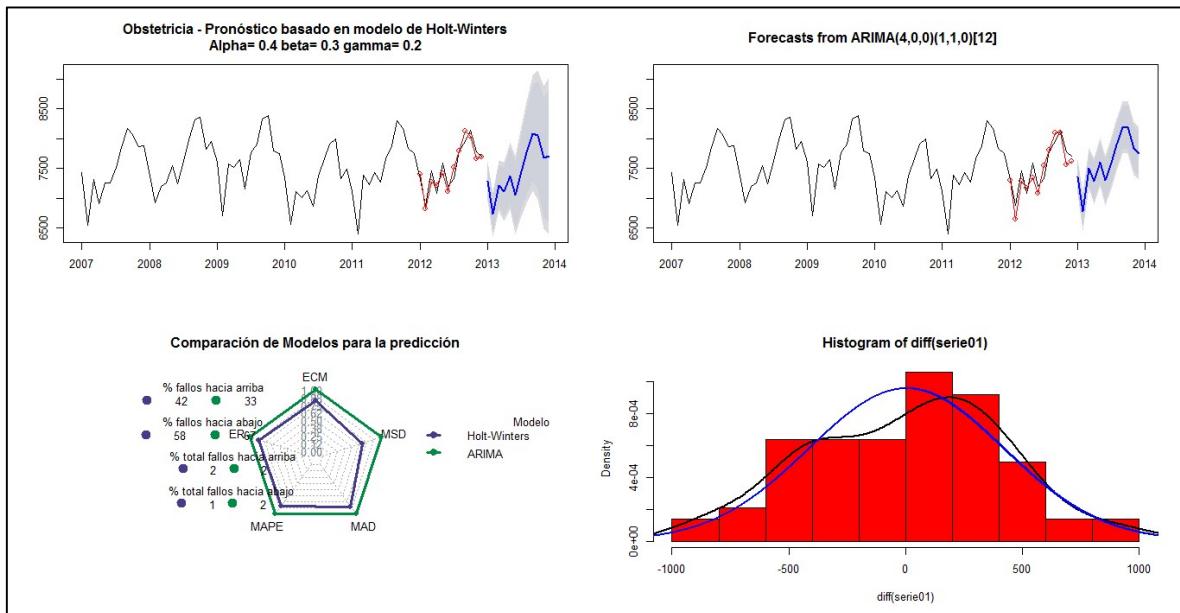


Figura 15. Panel comparativo de la estimación de la cantidad de hospitalizaciones en el servicio de pediatría. CCSS (2007-2012).

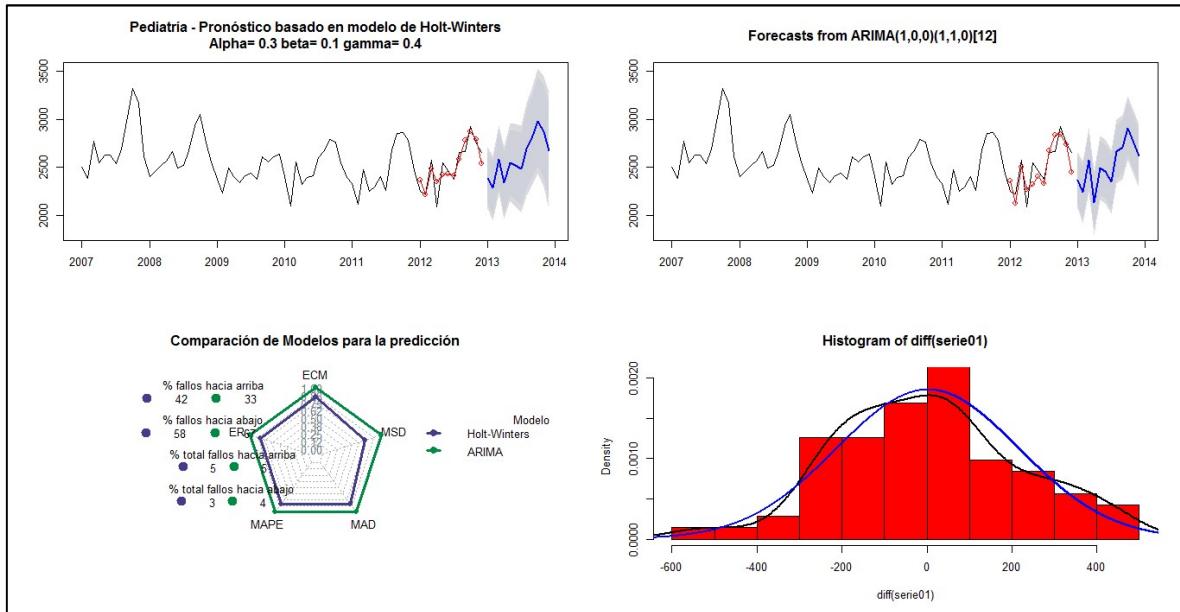
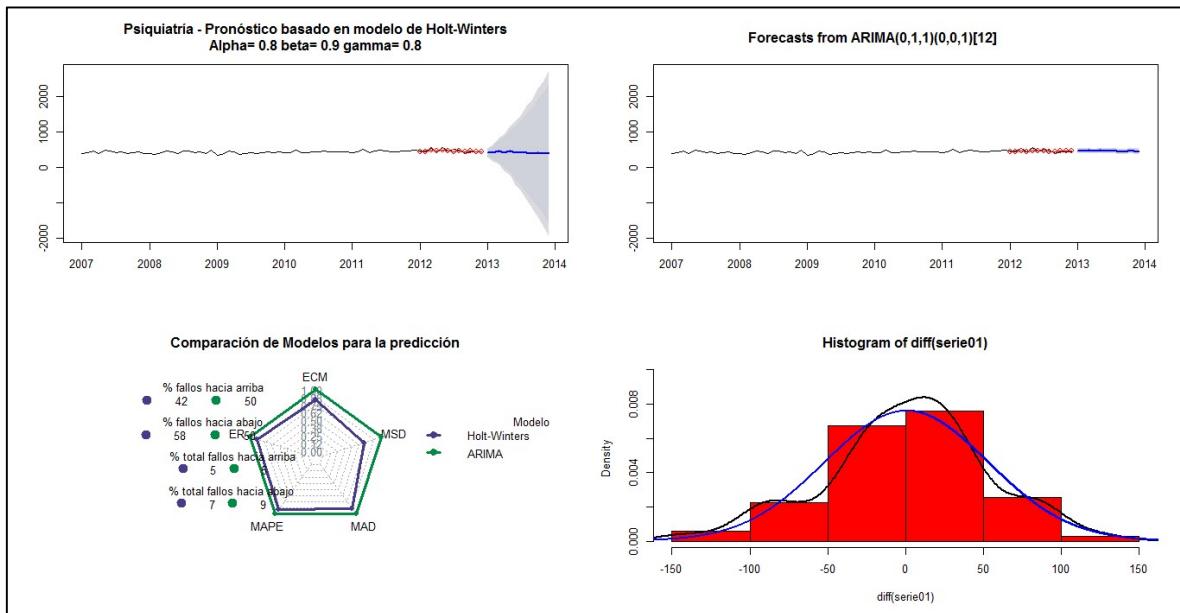


Figura 16. Panel comparativo de la estimación de la cantidad de hospitalizaciones en el servicio de psiquiatría. CCSS (2007-2012).



VII. Bibliografía

Cabrer Borrás, B. (2005). Hacia la automatización de la modelización de las series temporales.

CRAN R. (s.f.). Recuperado el 3 de 8 de 2015, de <https://cran.r-project.org/web/packages/forecast/forecast.pdf>

Hernández Rodríguez, Ó. (2011). *Series Cronológicas*. San José: Editorial UCR.

Hyndman, R. (s.f.). *inside-R*. Recuperado el 1 de 8 de 2015, de <http://www.inside-r.org/packages/cran/forecast/docs/auto.arima>

Hyndman, R., & Khandakar, Y. (2008). Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software*.

Kabacoff, R. (s.f.). *Quick-R*. Recuperado el 7 de 31 de 2015, de <http://www.statmethods.net/advstats/timeseries.html>

Leonard, M. (2002). *Large-Scale Automatic Forecasting: Millions of Forecasts*.

Méllard, G., & Pastels, J.-M. (2000). Automatic ARIMA modeling including interventions, using time series expert software. *International Journal of Forecasting*, 497-508.

Meyer, D. (s.f.). Recuperado el 1 de 8 de 2015, de <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/HoltWinters.html>

Ord, K., & Lowe, S. (1996). Automatic Forecasting. *The American Statistician*, 88-94.

Terrádez, M., & Ángel A., J. (s.f.). UOC. Recuperado el 12 de 8 de 2015, de http://www.uoc.edu/in3/emath/docs/Series_temporales.pdf

PRÁCTICA PROFESIONAL II

DESARROLLO DE UNA HERRAMIENTA WEB PARA CALIBRACIÓN DE SERIES DE TIEMPO BAJO UN ENFOQUE DE MINERÍA DE DATOS

Gabriel Cordero Mora

RESUMEN

La presente investigación tiene como objetivo desarrollar una herramienta de modelación automática de series de tiempo, que sea capaz de generar pronósticos para distintas variables de manera simultánea. Para este trabajo se utiliza como ejercicio práctico una base de datos de 14 variables con información de las atenciones médicas que se brindan a nivel general en los distintos centros de salud de la Caja Costarricense del Seguro Social (CCSS). Estos datos provienen del sitio web de la CCSS por lo que son de carácter público, asegurando la replicabilidad de esta investigación.

Esta herramienta tiene la capacidad de evaluar los modelos de Suavizamiento Exponencial, específicamente Holt-Winters y también los modelo ARIMA, ambos conocidos por su capacidad de producir pronósticos univariantes. El procedimiento que sigue esta herramienta inicia con la calibración de un modelo para cada técnica y los compara para determinar cuál de los dos genera estimaciones con un menor error. Posteriormente se calculan pronósticos para cada una de las variables, tanto con el modelo de Holt-Winters como con el ARIMA, utilizando los parámetros que según la calibración de modelos, minimiza el error.

Para tomar la decisión de cuál modelo es el que produce los mejores pronósticos, se define una ventana de desempeño que consiste en un conjunto de datos dentro del periodo analizado que se utilizan para evaluar la bondad de los pronósticos. Se calculan varios índices de ajuste entre los cuales se puede mencionar el error relativo, el error cuadrático medio, el error absoluto porcentual promedio y la desviación media absoluta; a partir de las cuales se crea un nuevo índice que evalúa estas medidas en forma conjunta.

VIII. INTRODUCCIÓN

La presente investigación tiene como objetivo desarrollar una herramienta de modelación automática de series de tiempo, que brinde pronósticos mediante el uso de una plataforma web que sea accesible y amigable para usuarios que no son especialistas en el tema de series de tiempo o programación. Para lograr esto, se utiliza la plataforma de programación R y el paquete denominado "shiny", el cual permite desarrollar aplicaciones más amigables para un usuario final que la tradicional consola de R o RStudio.

Se seleccionaron 14 variables que se utilizan como ejercicio práctico en esta investigación, las cuales fueron tomadas de la página web de la Caja Costarricense del Seguro Social (CCSS, 2000). La mitad de las variables corresponden a consulta externa: total de consultas, consultas subsecuentes, consultas de primera vez en el año, consultas de primera vez en la especialidad, consultas de primera vez en la vida, horas programadas y horas utilizadas. Las restantes 7 variables correspondientes a la cantidad de hospitalizaciones según servicio: cirugía, ginecología, medicina, neonatología, obstetricia y pediatría. Estas variables corresponden a una agregación de todos los centros de salud de la CCSS, pero la herramienta tiene la capacidad de procesar estas mismas variables a niveles más específicos, como por ejemplo a nivel de un hospital, área de salud, EBAIS o bien, variables de otros campos de análisis.

En la actualidad el análisis de series de tiempo es de gran importancia para las distintas empresas, organizaciones o instituciones; cualquiera de ellas requiere conocer el comportamiento futuro de ciertos fenómenos con el fin de planificar y prevenir ciertas situaciones. Las técnicas que permiten predecir lo que ocurrirá con una variable en el futuro a partir del comportamiento de esa variable en el pasado se conocen como técnicas univariadas. En este caso, disponer de una estimación del volumen de consultas que se demandará en un periodo futuro, permitirá a las autoridades de la CCSS contar con el personal necesario para abastecer la demanda de servicios en salud y a su vez presupuestar los costos asociados a estas consultas.

Una de las problemáticas que enfrentan muchas empresas en la actualidad, entre ellas la CCSS, es que no cuentan con el suficiente personal capacitado ni con el tiempo para generar este tipo de pronósticos, o bien no se cuenta con los recursos para adquirir una herramienta de modelado automático de series temporales que pueda generar rápidamente estas estimaciones con una precisión aceptable.

Por lo anterior, el desarrollo de una herramienta en un código abierto y gratuito como R resulta de gran utilidad, pues el usuario será capaz de realizar estimaciones automáticas mediante una plataforma web que hará que su experiencia sea más agradable; además, no es necesario que cuente con conocimientos especializados en el tema de series de tiempo o en programación, solamente un pequeño entrenamiento en el uso de la herramienta y manejo de la base de datos que se esté analizando, para que sea capaz de validar la consistencia de los resultados.

Hoy en día, los algoritmos de predicción automática más populares se basan en modelos ARIMA o de Suavizamiento Exponencial; en este último caso los del tipo Holt-Winters son los más utilizados cuando las series tienen estacionalidad. Por esta razón en esta investigación se utilizan ambas técnicas.

Estos modelos ya están incorporados en paquetes de modelación automática, sin embargo, el elevado costo que suelen tener estas herramientas o bien el conocimiento técnico que requieren, representan una limitación para algunas instituciones o empresas que no cuentan con personal especializado o bien con los recursos financieros.

IX. VARIABLES Y MÉTODOS

9.1 Variables

Los datos utilizados para esta investigación provienen del sitio web de la CCSS, por lo que son datos de carácter público (CCSS, 2000).

Esta información hace referencia principalmente a las atenciones médicas que se dan a nivel general en los distintos centros de salud, es decir, se trata de datos agregados de todos los centros de salud; sin embargo, como se mencionó anteriormente, esta herramienta tiene la capacidad de estimar modelos a niveles más desagregados como hospitales, áreas de salud, EBAIS o bien otros campos de análisis.

En total se estarán analizando 14 variables, las cuales se describen en las secciones siguientes. Todas estas variables tienen una periodicidad mensual, corresponden a una medición acumulada a fin de mes y los datos disponibles van de enero del 2007 a diciembre del 2012, lo que representa seis años de información o su equivalente de 72 observaciones. En este caso la periodicidad es mensual, sin embargo, la herramienta tiene la opción de analizar variables con una periodicidad distinta, como por ejemplo trimestral, semestral o anual.

9.1.1 Variables de Consulta Externa

La consulta externa se presenta en dos niveles de atención dentro de la CCSS. En el **primer nivel** de atención se ofrece servicios de promoción de la salud, prevención, curación de la enfermedad y rehabilitación de menor complejidad, mientras que el **segundo nivel** de atención apoya al primer nivel de atención a través de una red de establecimientos formada por hospitales regionales, hospitales periféricos y clínicas mayores ubicadas en la gran área metropolitana (Sáenz, Acosta, Muiser, & Bermúdez, 2011).

Las variables que en esta oportunidad se van a analizar son las siguientes:

- **Consultas de primera vez en la vida (CPVV):** Atención brindada en la consulta externa a un paciente que nunca ha sido atendido dentro de la CCSS en esta modalidad.
- **Consultas de primera vez en el año (CPVA):** Atención brindada en la consulta externa a un paciente que no ha sido atendido en esta modalidad durante el año en curso, pero sí en años anteriores.
- **Consultas de primera vez en la especialidad (CPVE):** Atención brindada en la consulta externa a un paciente que ya ha sido atendido en esta modalidad durante el año en curso y que es su primera vez en ese año para la especialidad en específico en la que fue atendido.
- **Consultas subsecuentes (CS):** Atención brindada en la consulta externa y que no cumple ninguna de las condiciones de las consultas anteriores.
- **Total de consultas:** Total de atenciones brindadas en la consulta externa y se calcula como la suma de CPVV, CPVA, CPVE y CS.
- **Horas programadas:** Total de horas programadas de los médicos para atender la consulta externa.
- **Horas utilizadas:** Total de horas utilizadas por los médicos para atender la consulta externa.

9.1.2 Variables de hospitalizaciones

El servicio de hospitalización se presenta en los dos niveles de atención más altos de la CCSS (II y III). El segundo nivel ofrece internamiento de períodos cortos en hospitales regionales, mientras que el tercer nivel cuenta con servicios de internamiento y servicios médico-quirúrgicos de alta complejidad tecnológica que se brindan en los hospitales nacionales y especializados (Sáenz, Acosta, Muiser, & Bermúdez, 2011). En este punto las variables analizadas fueron 7 y cada una hace referencia a la cantidad de hospitalizaciones a pacientes en los servicios de cirugía, ginecología, medicina general, neonatología, obstetricia, pediatría y psiquiatría.

9.2 Métodos de estimación

A continuación, se mencionan las técnicas que se estarán utilizando para realizar las proyecciones. Es importante mencionar que no se profundizará en cada método, únicamente se indican las características generales, pues el objetivo principal de esta investigación no es profundizar en la parte teórica de las mismas. Por consiguiente, se omite en este documento la explicación de algunos conceptos básicos en el análisis de series de tiempo como por ejemplo los componentes de tendencia, ciclicidad, estacionalidad y aleatoriedad, así como detalles específicos de cada técnica.

El paquete de R en el cual se basa la gran parte de este trabajo es "*forecast*". Este paquete cuenta con una serie de métodos y herramientas para mostrar y analizar previsiones de series de tiempo

univariadas, incluyendo el Suavizado Exponencial y modelado ARIMA automático (Hyndman R. J., 2015).

9.2.1 Métodos de suavización exponencial

Estos métodos tienen como ventaja su bajo costo y la sencillez al momento de aplicarlos; además, pueden ser utilizados aun cuando se disponga de una pequeña cantidad de observaciones.

El método de **Suavización exponencial simple** es apropiado para series que no tienen patrones estacionales ni de tendencia y cuya media o nivel cambia lentamente (Hernández, 2011).

Para este caso el pronóstico se puede denotar de la siguiente forma:

$$P_{t+1} = P_t + \alpha(Z_t - P_t)$$

Donde Z, es la variable que se desea predecir, P es el pronóstico y α es una constante que corresponde al valor óptimo que minimiza la suma de los cuadrados de los errores de pronóstico.

Si la serie presenta un patrón de tendencia lineal, el método anterior pierde validez y se hace necesario optar por métodos más complejos. El **Método Lineal de Holt** viene a solventar la limitante del Método de Suavización Exponencial Simple.

Este método utiliza 3 ecuaciones y 2 parámetros, α y β que toman valores entre 0 y 1. Las ecuaciones son las siguientes:

$$\begin{aligned} a_t &= \alpha Z_t + (1 - \alpha)(a_{t-1} + b_{t-1}) \\ b_t &= \beta(a_t - a_{t-1}) + (1 - \beta)b_{t-1} \\ P_{t+m} &= a_t + b_t m \end{aligned}$$

El valor de a_t es una estimación del nivel promedio de la serie en el tiempo t, mientras que b_t es una estimación de la pendiente de la serie Z en el tiempo t.

En el caso de P_{t+m} es la ecuación que permite pronosticar el valor de Z, m periodos por delante del tiempo t, sumando m veces el valor de la pendiente estimada en el tiempo t al nivel de la serie en el tiempo t.

Por su parte, los valores de α y β se obtienen al minimizar la suma de cuadrados de los errores de pronóstico para todos los valores posibles de α y β en el intervalo (0,1).

Un tercer método que toma en cuenta la presencia de estacionalidad en la serie es el denominado Holt-Winters que toma en cuenta el componente tendencia y la estacionalidad.

De este método se pueden encontrar dos variantes, el multiplicativo y el aditivo. La diferencia básica entre uno y otro es que el modelo aditivo considera que la variable observada se puede descomponer en la suma de los factores (tendencia, estacionalidad, ciclicidad y aleatoriedad), mientras el multiplicativo, el comportamiento de la variable observada se expresa como el producto de los componentes de la serie.

La herramienta desarrollada compara los resultados para ambas opciones y recomienda el modelo que mejor se ajuste a los datos.

El método multiplicativo se denota de la siguiente manera:

$$\begin{aligned} a_t &= \alpha \frac{Z_t}{S_{t-s}} + (1 - \alpha)(a_{t-1} + b_{t-1}) & S_t &= \gamma \frac{Z_t}{a_t} + (1 - \gamma)S_{t-s} \\ b_t &= \beta(a_t - a_{t-1}) + (1 - \beta)b_{t-1} & P_{t+m} &= (a_t + b_t m)S_{t-s+m} \end{aligned}$$

El método aditivo se denota de la siguiente manera:

$$\begin{aligned} a_t &= \alpha(Z_t - S_{t-s}) + (1 - \alpha)(a_{t-1} + b_{t-1}) \\ b_t &= \beta(a_t - a_{t-1}) + (1 - \beta)b_{t-1} \\ S_t &= \gamma(Z_t - a_t) + (1 - \gamma)S_{t-s} \\ P_{t+m} &= a_t + b_t m + S_{t-s+m} \end{aligned}$$

En las ecuaciones "s" representa la longitud de la estacionalidad (número de meses o trimestres en un año), a_t representa el nivel de la serie Z en el tiempo t, b_t es la tendencia, S_t es el componente estacional, P_{t+m} es el pronóstico m períodos adelante mientras que los parámetros pueden variar en los siguientes rangos: $0 < \alpha \leq 1$, $0 \leq \beta \leq 1$, $0 \leq \gamma \leq 1$.

En caso de que el parámetro $\gamma = 0$, el componente $S_t = S_{t-s}$, mientras que si $\beta = 0$, sucede algo similar, se obtiene como resultado que $b_t = b_{t-1}$, por último, α tiene como restricción que debe ser mayor a 0.

9.2.2 Modelos ARIMA (enfoque Box-Jenkins)

Un modelo ARIMA tiene asociado una función de autocorrelación teórica y una función de autocorrelación parcial teórica que lo caracteriza. El enfoque de Box-Jenkins compara estas funciones teóricas con las respectivas funciones muestrales de autocorrelación y de autocorrelación parcial con el fin de identificar inicialmente un proceso probabilístico ARIMA que represente razonablemente las observaciones de la serie temporal que se analiza. Una vez

identificado el proceso, se procede a estimar los parámetros que lo definen y, luego, a realizar un diagnóstico para evaluar el ajuste del proceso a los datos. Si no se diera un buen ajuste, el enfoque de Box-Jenkins repite de nuevo las etapas de identificación, estimación y diagnóstico hasta encontrar un modelo ARIMA que describa adecuadamente los datos (Hernández, 2011).

Estos modelos utilizan la notación $ARIMA(p, d, q)$ para representar un modelo autoregresivo(AR) integrado (I) de promedios móviles (MA) para Z_t , donde:

- p indica el orden de la parte autorregresiva del modelo, es decir, el número de rezagos de Z_t que se incorporan a la ecuación.
- d indica el número de veces que se ha diferenciado Z_t para satisfacer la condición de estacionaridad que se define más adelante.
- q especifica el orden de la parte de promedios móviles del modelo, es decir, el número de rezagos de los residuos a_t que se incorporan a la ecuación.

En términos generales un modelo ARIMA que únicamente tiene el componente autorregresivo $AR(p)$, se denomina como un modelo no estacional autorregresivo de orden p y está dado por la siguiente ecuación:

$$Z_t = C + \varphi_1 Z_{t-1} + \varphi_2 Z_{t-2} + \cdots + \varphi_p Z_{t-p} + a_t$$

donde C y las φ_i son constantes desconocidas, y a_t es ruido blanco.

Un modelo no estacional de promedios móviles de orden q para Z_t $MA(q)$ está dado por la siguiente ecuación:

$$Z_t = C + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q}$$

donde C y las θ_i son constantes desconocidas, y a_t es ruido blanco.

Un modelo más complejo puede ser el modelo mixto no estacional autorregresivo de promedios móviles de orden (p,q) , $ARMA(p, q)$, que está dado por la siguiente ecuación:

$$Z_t = C + \varphi_1 Z_{t-1} + \varphi_2 Z_{t-2} + \cdots + \varphi_p Z_{t-p} + a_t + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q}$$

Estos modelos son adecuados para series temporales que son estacionarias. Se dice que una serie es estacionaria si cumple las siguientes condiciones:

- La serie Z_t debe tener una media que toma el mismo valor sea cual sea el valor de t .
- Debe tener una varianza constante para cada valor de t .

- La correlación entre valores de la serie en t y t-k, Z_t y Z_{t-1} , vale lo mismo para k dada toda t.

En la práctica la mayoría de las series temporales no son estacionarias, sin embargo, se pueden convertir fácilmente en series estacionarias, calculando la diferenciación (d) entre valores consecutivos. En ocasiones, una segunda diferenciación es necesaria en caso de que la primera diferenciación no satisfaga las condiciones mencionadas anteriormente. Usualmente d es un valor pequeño como 1 ó 2.

Un modelo $ARIMA(0, d, 0)$ es una serie temporal que se convierte en ruido blanco después de ser diferenciado d veces. Este modelo se puede expresar utilizando la siguiente ecuación:

$$(1 - B)^d Z_t = a_t$$

Para que se puede considerar que la serie temporal es ruido blanco, se deben cumplir 2 condiciones:

- $\text{corr}(Z_t, Z_{t+k}) = 0 \forall k \neq 0$
- $E[Z_t]$ es constante para todo t (usualmente $Z_t = 0$)

Por lo que, en términos generales, un modelo $ARIMA(p, d, q)$ se puede representar de la siguiente manera:

$$(1 - \varphi_1 B + \varphi_2 B^2 + \cdots + \varphi_p B^p)(1 - B)^d = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)a_t$$

En el caso que la serie tenga estacionalidad se define como $ARIMA(p, d, q)(P, D, Q)$, el primer paréntesis (p, d, q) se refiere a la parte regular de la serie y el segundo paréntesis (P, D, Q) se refiere a las estacionales o parte cíclica de la serie temporal.

9.3 Desarrollo de aplicación con Shiny

Con el fin de brindar al usuario final una herramienta accesible y amigable, se decide utilizar la librería Shiny de R para desarrollar la aplicación web.

La implementación y administración de aplicaciones Shiny en la nube es relativamente sencillo, no requiere conocimientos de HTML, poseer un servidor ni conocimiento para configurar un *firewall* para implementar y administrar las aplicaciones; lo único necesario es tener conocimientos en el lenguaje de programación R. Además, no se requiere *hardware*, instalación o contrato de compra anual a no ser que el consumo de datos se vuelva masivo y sobrepase las cuotas de una licencia gratuita.

Las aplicaciones creadas con Shiny pueden correr tanto a nivel local en el computador del usuario como en un servidor; también pueden ser almacenadas en un servidor y luego ejecutadas directamente en el ordenador del usuario.

Una aplicación shiny se compone de al menos dos ficheros: **ui.R** que es la interfaz del usuario y **server.R** que es un *script* almacenado en el servidor. Estos dos ficheros deben estar contenidos en un mismo directorio con el nombre de la aplicación. El fichero ui.R es el encargado de crear la interfaz que estará viendo el usuario, dicha interfaz es dinámica y tiene un papel muy importante dentro de la aplicación, dado que es la encargada de enviar al fichero server.R, los valores especificados por el usuario; por su parte, server.R se encarga de utilizar los valores introducidos por el usuario para ejecutar el código que haya sido definido. El intercambio de datos de entrada y salida entre las dos partes de la aplicación se hace de manera automática.

Para ilustrar el uso de Shiny, a continuación se muestra el código contenido en los ficheros que permiten generar un histograma. La estructura básica de un fichero ui.R es la siguiente:

```
ui.R
library(shiny) # carga del paquete

shinyUI(fluidPage(
  titlePanel("Ejemplo del uso del Shiny!"),
  # Barra lateral con una entrada para definir el número de barras
  sidebarLayout(
    sidebarPanel(
      sliderInput("barras",
                 "Número de barras:",
                 min = 1,
                 max = 50,
                 value = 30)
    ),
    # Crea un histograma y este se visualiza en la aplicación
    mainPanel(
      plotOutput("distPlot")
    )
  )))

```

La instrucción anterior permite enviar un dato al fichero server.R, este dato corresponde "barras" el cual indica la cantidad de barras que desea que tenga el histograma. Adicionalmente, tiene una función llamada plotOutput la cual se encarga traer el histograma creado en el fichero server.R, para que sea visible al usuario en la interfaz de la aplicación.

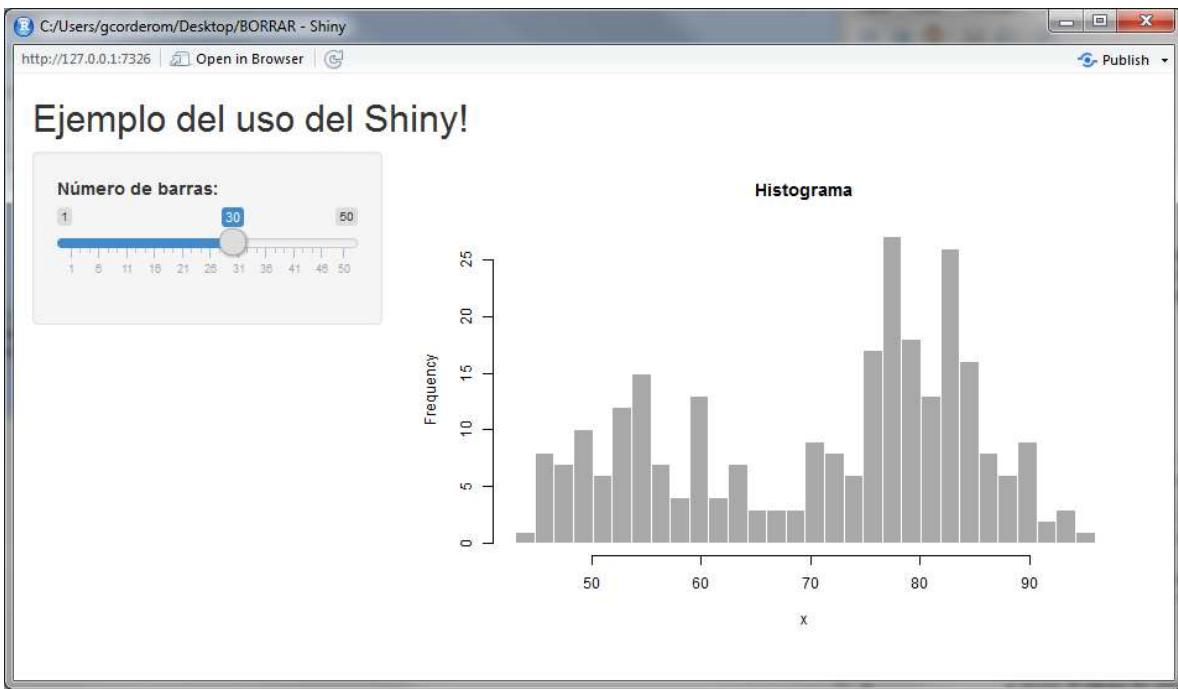
Una estructura básica de un fichero server.R puede ser la siguiente:

```
Server.R
library(shiny)

# Define la lógica a nivel de servidor, para graficar el histograma
shinyServer(function(input, output) {
  # Esta función gráfica el histograma e indica mediante renderPlot que la función es reactiva, por lo que debe
  #ejecutarse nuevamente cada vez que el usuario modifique los valores que este puede visualizar en la interfaz
  output$distPlot <- renderPlot({ # distPlot es el objeto llamado por ui.R
    x <- faithful[, 2] # datos utilizados
    barras <- seq(min(x), max(x), length.out = input$barras + 1)
    # Dibuja el histograma con las columnas seleccionadas por el usuario
    hist(x, breaks = barras, col = 'darkgray', border = 'white', main = paste("Histograma"))
  }))}
}
```

Obteniendo como resultado la siguiente salida.

Figura 17. Ejemplo de una interfaz creada con la librería de Shiny.



Para el caso del presente estudio, fue necesario desarrollar una serie de pasos más complejos que los especificados en el ejemplo anterior, los cuales se detallan seguidamente.

9.3.1 Calibración de los modelos

Para iniciar el proceso de calibración es necesario contar con un archivo de datos plano donde cada una de las columnas corresponda a las distintas variables que se desean pronosticar. La estructura de este archivo debe ser similar a la que se presenta en la siguiente figura y es de suma importancia que los datos estén ordenados de los más antiguos a los más recientes.

Tabla 3. Muestra del archivo de datos de las series temporales bajo análisis.

Fecha	PVA	PVE	PVV	CS	HP	HU	TC	Cirugía	Ginecología	Medicina	Neonatología	Obstetricia	Pediatría	Psiquiatría
31/01/2007	664794	25458	21433	100559	266752	242874	812244	4897	2010	4041	1053	7426	2503	392
28/02/2007	449829	51957	22137	254286	257733	234775	778209	4929	1942	3929	982	6542	2388	405
31/03/2007	395182	63628	23444	387668	282457	258041	869922	5733	2321	4284	1137	7318	2771	465
30/04/2007	241198	50096	19302	404983	233790	211703	715579	4971	2026	3814	1052	6915	2545	377
31/05/2007	249955	63763	25301	542136	283601	258782	881155	5699	2229	4204	1093	7255	2627	468
30/06/2007	206362	62294	22769	552398	268676	244524	843823	5656	2102	4422	1087	7261	2632	445
31/07/2007	171654	54571	24173	579470	262320	238086	829868	5388	2088	4925	1139	7517	2541	415
31/08/2007	150536	59812	23502	616021	272125	247408	849871	5732	2051	5140	1105	7848	2688	438
30/09/2007	134509	56503	21842	601282	258556	234693	814136	5879	2135	4720	1121	8178	3027	371

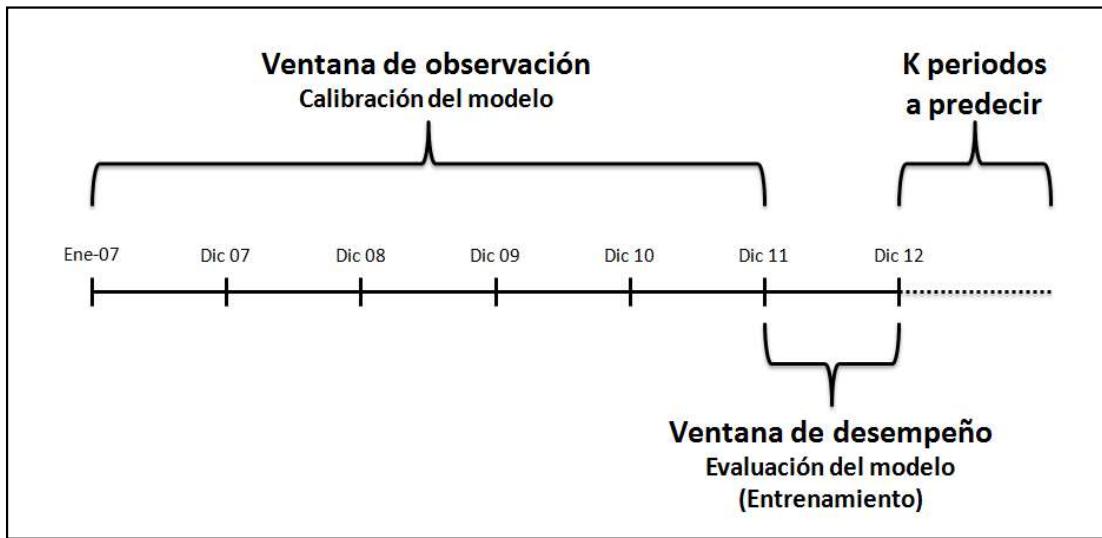
Una vez que se cuenta con el archivo de datos, la herramienta revisa si el archivo tiene valores faltantes que se deben reemplazar por valores válidos o bien valores extremos que se deban ajustar. En caso de que el usuario así lo especifique, se utiliza la función "tsclean" de R, que puede identificar y reemplazar valores extremos o valores perdidos dentro de una serie de tiempo.

Dentro de la función "tsclean" se pueden ejecutar distintos procesos según se necesite. Por ejemplo, para la identificación de valores extremos, se utiliza la función "supsmu" (Friedman's SuperSmoother) para series sin estacionalidad y una descomposición periódica STL (Seasonal Decomposition of Time Series by Loess) para series con estacionalidad. Para estimar los valores faltantes y los reemplazos de valores atípicos, se utiliza la interpolación lineal en la serie (ajustada estacionalmente de ser necesario).

Una vez depurado el archivo de datos, se inicia el proceso para calibrar el mejor modelo de Holt-Winters y ARIMA, mediante la creación de un proceso en R que es capaz de detectar el modelo con los parámetros que generan un menor error de estimación para cada una de estas técnicas.

Para la selección del mejor modelo se divide la información en dos secciones: la primera sección comprende de enero 2007 a diciembre 2011 y se denomina ventana de observación; la segunda sección va de enero 2012 a diciembre 2012 y se denomina ventana de desempeño. En la ventana de observación se calibran los modelos y en la ventana de desempeño se evalúa el ajuste y calidad de las estimaciones.

Figura 18. Ventana de observación y de desempeño para el presente estudio.



En el caso del modelo ARIMA, se utiliza la función `auto.arima` que pertenece a la librería `forecast` de R. Esta función tiene la capacidad de devolver el mejor modelo ARIMA según alguno de los criterios de información siguientes: AIC (*default*), BIC, AICC (Hyndman & Khandakar, 2008).

Por otra parte, para el modelo Holt-Winters se utilizó una función que genera un proceso iterativo en el cual se prueban distintos valores para los parámetros alpha, beta y gamma y se determina cual combinación minimiza el error cuadrático medio del modelo. En este proceso se calculan un total de 900 modelos diferentes para las posibles combinaciones de valores de los parámetros: alpha=0.1, 0.2,...,1 (alpha no puede ser cero); beta=0.0, 0.1, 0.2,..., 1; gamma=0.0, 0.1, 0.2,..., 1.

9.3.2 Selección del mejor modelo: Holt-Winters vs. ARIMA

Una vez que en la etapa de calibración se obtiene el mejor modelo ARIMA y el mejor modelo de Holt-Winters, corresponde seleccionar el que tenga asociado el error más pequeño. Para realizar esta selección se utilizan una serie de medidas para evaluar el ajuste, entre los cuales se puede mencionar el error relativo (RE), error cuadrático medio (MSE), el error absoluto porcentual promedio (MAPE) y la desviación media absoluta (MAD).

A continuación se describen las fórmulas de cálculo de cada una de las **medidas para evaluar el ajuste**:

$$RE = \frac{\sum_{i=1}^N |R_i - P_i|}{\sum_{i=1}^N |R_i|} * 100$$

$$MAPE = \frac{\sum_{i=1}^N \left| \frac{(R_i - P_i)}{P_i} \right|}{N} * 100$$

$$MSE = \frac{1}{N} \sum_{i=1}^N (R_i - P_i)^2$$

$$MAD = \frac{\sum_{i=1}^N |R_i - P_i|}{N}$$

Donde

N: Número de datos pronosticados

R_i: Valor real en el tiempo *i*

P_i: Valor pronosticado en el tiempo *i*

Posteriormente se elaboró un índice para cada modelo que combina los resultados de las cuatro medidas de ajuste. El mejor modelo será aquel que tenga el valor más pequeño de este índice.

$$I_{HW} = \frac{RE_{HW}}{\max(RE_A, RE_{HW})} + \frac{MSE_{HW}}{\max(MSE_A, MSE_{HW})} + \frac{MAPE_{HW}}{\max(MAPE_A, MAPE_{HW})} + \frac{MAD_{HW}}{\max(MAD_A, MAD_{HW})}$$

$$I_{ARIMA} = \frac{RE_A}{\max(RE_A, RE_{HW})} + \frac{MSE_A}{\max(MSE_A, MSE_{HW})} + \frac{MAPE_A}{\max(MAPE_A, MAPE_{HW})} + \frac{MAD_A}{\max(MAD_A, MAD_{HW})}$$

Es importante mencionar que este índice sirve como criterio para seleccionar el mejor modelo, sin embargo, ambos modelos podrían ser satisfactorios; por tanto, queda a criterio del usuario cuál tiene mayor funcionalidad práctica y mayor consistencia según lo observado en los pronósticos generados en la ventana de desempeño.

9.3.3 Construcción de una interfaz para el usuario final

Una vez definida la lógica de trabajo que debe seguir la herramienta para realizar los pronósticos de una manera automática, se procede a realizar la programación en R. Como se mencionó antes, es necesario contar con dos ficheros, server.R y ui.R.

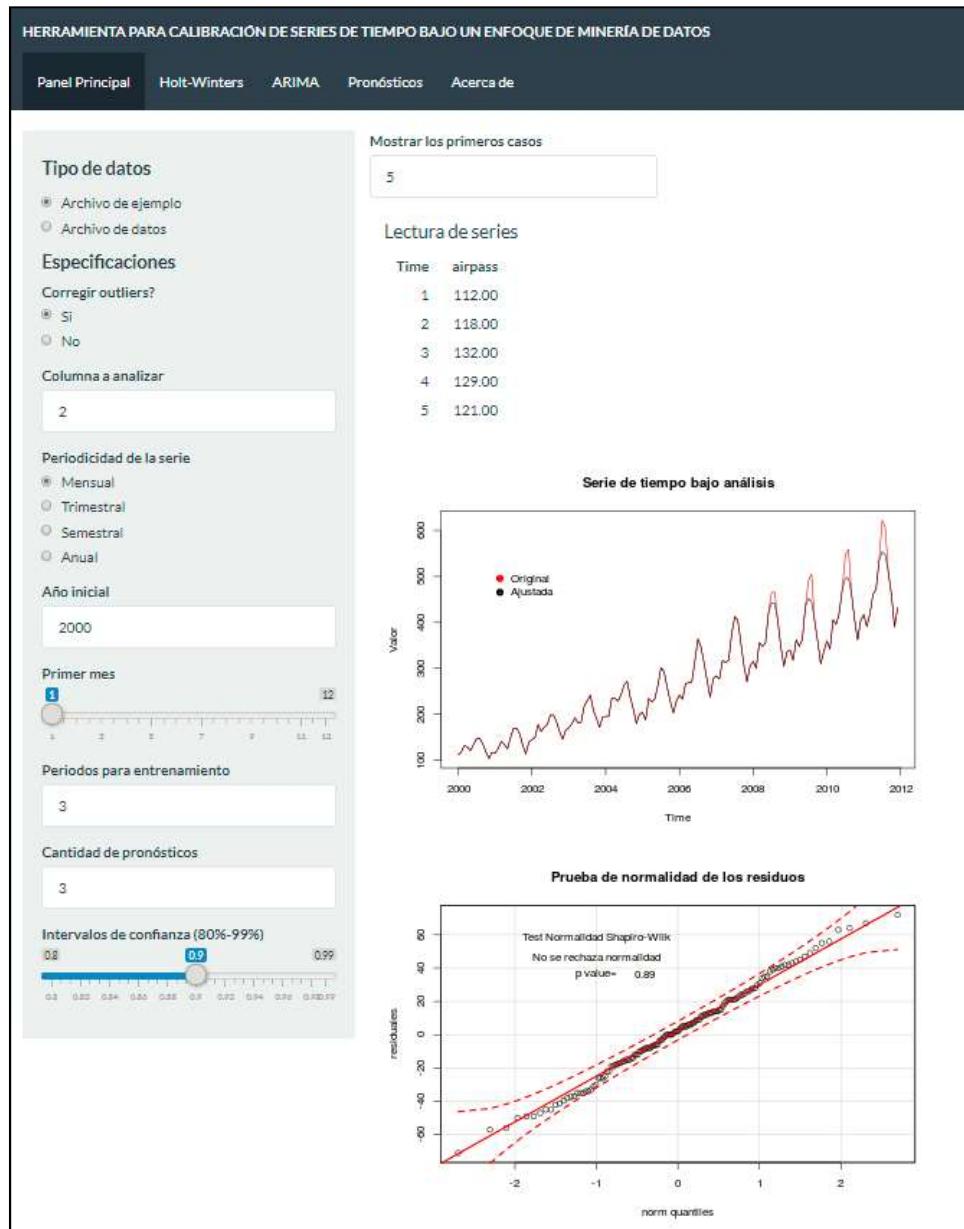
Estos ficheros permiten generar una interfaz para un usuario siempre y cuando el equipo de cómputo cuente con RStudio (RStudio Inc., 2009), plataforma de desarrollo libre y de código abierto para R. La interfaz generada se ejecuta localmente en un equipo, por lo que únicamente puede ser utilizada por el usuario que cuente con los ficheros antes mencionados, ui.R y server.R, lo cual es un inconveniente bastante importante en términos de accesibilidad a más usuarios.

Para resolver este inconveniente se utiliza la plataforma shinyapps de RStudio la cual permite contar con una aplicación accesible a más usuarios a través de la web, por lo que no requiere tener instalado RStudio ni debe ejecutarse código de R. Shinyapps permite publicar los

aplicativos desarrollados con Shiny, para que estén hospedados en un sitio web y sean accesibles a cualquier persona con acceso a internet.

La página web donde está hospedada la herramienta desarrollada es "<https://gcoorderom.shinyapps.io/Pronostica/>" y la interfaz es la siguiente.

Figura 19. Panel principal de la herramienta.



X. RESULTADOS

Con el fin de ejemplificar los resultados que puede generar la herramienta, se utilizarán los datos de la CCSS y se modelará las 14 variables descritas en la sección 2.1.

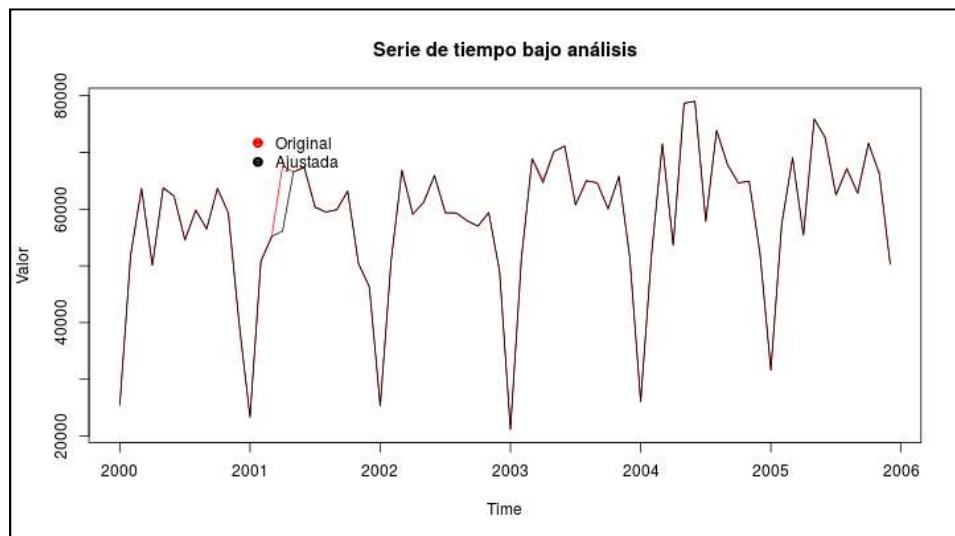
10.1 Análisis descriptivo de las series de tiempo

Esta herramienta genera una salida para cada variable que consta de dos gráficos: un gráfico lineal que permite observar el comportamiento de la serie de tiempo (gráfico 4) y un gráfico donde se evalúa el cumplimiento del supuesto de normalidad de los residuos (gráfico 5).

Es importante mencionar que existen otras pruebas que pueden aportar información valiosa sobre la factibilidad de aplicar modelos de series de tiempo, sin embargo, se decidió utilizar únicamente el test de Shapiro-Wilk para tener una idea general del comportamiento de la variable y evaluar el cumplimiento del supuesto de normalidad de los residuos. En caso de aceptarse la hipótesis, los resultados pueden tomarse como válidos, en caso de rechazarse, se debe valorar la posibilidad de realizar una transformación a la variable en estudio (la herramienta no realiza este proceso de manera automática).

En el siguiente ejemplo se muestra el gráfico de la serie en el cual se aprecia que se realizó un ajuste en los valores extremos que presentó la serie (línea roja es la serie original y la línea negra es la serie ajustada).

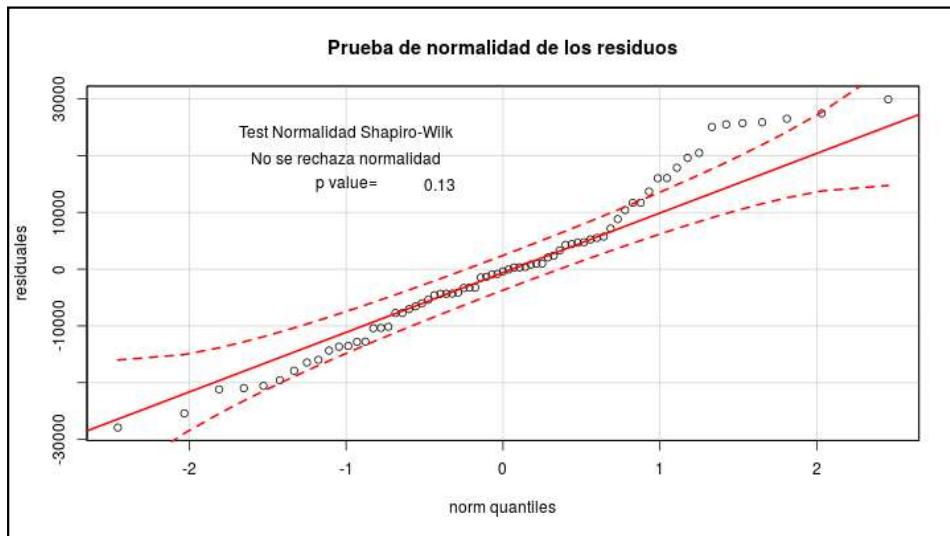
Gráfico 4. Análisis de la serie de tiempo y corrección de valores extremos y valores faltantes.



En el segundo gráfico se puede corroborar si el modelo seleccionado para la serie cumple con el supuesto de normalidad de los residuos. Para ello se agrega el test de Shapiro-Wilk que permite

realizar una prueba de hipótesis para H_0 : los residuos se distribuyen normalmente. Un valor de $p < 0.05$ indicaría que el supuesto de normalidad se debe rechazar, mientras que un valor de $p \geq 0.05$ indica que los residuos se distribuyen normalmente. En el caso que se presenta a continuación, $p = 0.13$ por lo que no se rechaza el supuesto de normalidad de los residuos y se puede seguir analizando la serie.

Gráfico 5. Prueba de la normalidad de los residuos.

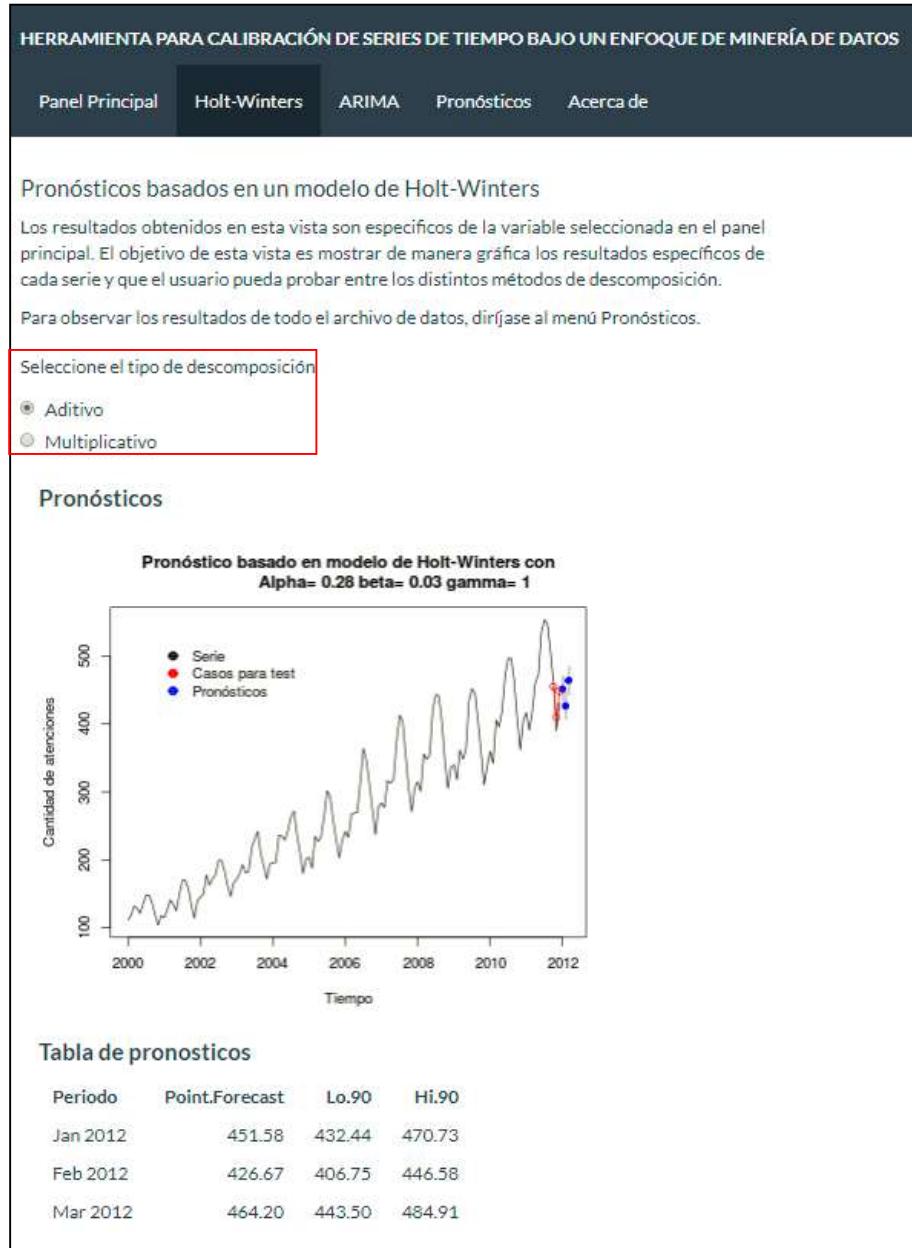


10.2 Exploración de resultados: Holt-Winters -ARIMA

Los dos siguientes paneles permiten revisar los resultados específicos de la variable que se está analizando. En el caso del modelo de Suavizamiento Exponencial se puede cambiar el método de descomposición entre Aditivo y Multiplicativo; sin embargo, esto es únicamente para que el usuario pueda explorar ambas opciones, pues la herramienta por sí misma determina cual método genera mejores resultados en términos de reducción de los errores.

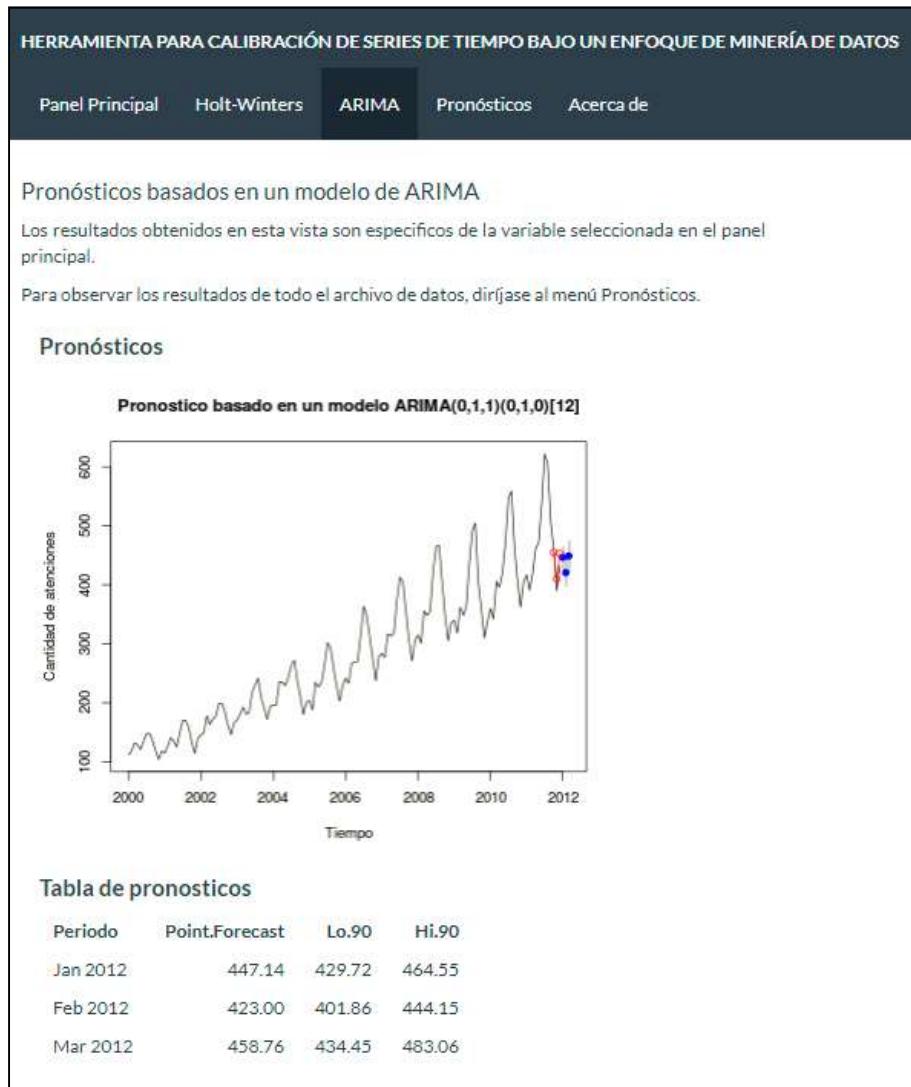
La interfaz para el método de Holt-Winters es la siguiente:

Figura 20. Panel de estimación utilizando Suavizamiento Exponencial.



En el caso de ARIMA, la vista es muy similar, consta de un gráfico y una tabla donde se puede ver los valores que está pronosticando este modelo.

Figura 21. Panel de estimación utilizando ARIMA.

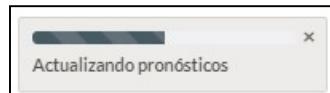


10.3 Generación de pronósticos

Esta vista genera los pronósticos para todas las series de tiempo incluidas en el archivo de datos. En caso de no querer obtener pronósticos para alguna serie en particular, ya sea porque no cumple con el supuesto de normalidad de los residuos o bien el usuario considera por alguna razón que es mejor excluir alguna serie, lo ideal es eliminarlas del archivo y refrescar la lectura de datos; esto provocará que la herramienta necesite menos tiempo de procesamiento y hará el proceso más rápido.

Al ingresar a la interfaz de "Pronósticos" automáticamente se inicia el proceso de cálculo de pronósticos para todas las variables incluidas en el archivo de datos. El siguiente mensaje indica que se está ejecutando el proceso.

Figura 22. Barra indicativa de avance.



Una vez terminado el proceso se obtiene la siguiente tabla, la cual se puede descargar a un archivo de formato .csv con la opción "**Descargar pronósticos**".

Figura 23. Panel de generación de los pronósticos finales.

HERRAMIENTA PARA CALIBRACIÓN DE SERIES DE TIEMPO BAJO UN ENFOQUE DE MINERÍA DE DATOS

Panel Principal Holt-Winters ARIMA **Pronósticos** Acerca de

Pronósticos de todas las series

En la siguiente tabla se encuentran los pronósticos para cada una de las series de tiempo presente en el archivo de datos. Para cada serie se generan dos pronósticos diferentes, uno basado en un modelo de Holt-Winters y otro obtenido mediante un ARIMA, la columna "Método Recomendado" indica cual de los dos pronósticos tiene un menor margen de error según los datos de entrenamiento.

Descargar pronósticos

Periodo	Metodo Pronóstico	Metodo Recomendado	Variable	Pronóstico	LI 90 %	LS 90 %
Jan 2006	Holt-Winters	ARIMA	PVA	711231.10	685841.00	736621.30
Feb 2006	Holt-Winters	ARIMA	PVA	511246.10	485627.30	536864.80
Mar 2006	Holt-Winters	ARIMA	PVA	421916.10	396039.70	447792.50
Jan 2006	ARIMA	ARIMA	PVA	714572.00	686996.75	742147.25
Feb 2006	ARIMA	ARIMA	PVA	516883.00	489307.75	544458.25
Mar 2006	ARIMA	ARIMA	PVA	422220.00	394644.75	449795.25
Jan 2006	Holt-Winters	Holt-Winters	PVE	31193.80	23870.10	38517.50
Feb 2006	Holt-Winters	Holt-Winters	PVE	57551.90	50228.00	64875.80
Mar 2006	Holt-Winters	Holt-Winters	PVE	71816.80	64492.70	79140.90
Jan 2006	ARIMA	Holt-Winters	PVE	30570.60	22198.32	38942.88

XI. DISCUSION

El objetivo de esta sección es mostrar el aporte de este nuevo trabajo con respecto al que fue presentado en la Práctica Profesional I, así como mencionar las limitaciones que se han identificado para que en el futuro se pueda mejorar la funcionalidad y confiabilidad de esta herramienta.

Avances:

- Se realizaron mejoras a la herramienta desde el punto de vista de calidad de la información que se utiliza como insumo para calcular los pronósticos. En la primera versión de la herramienta los modelos se ajustaban a las series originales sin controlar la influencia de eventos extremos en la serie o el problema de los valores faltantes. En la presente versión este problema fue corregido.
- Se desarrolló una interfaz ágil y amigable para un usuario final. En la primera versión era necesario manipular un archivo de código el cual debía ser ejecutado en la consola de R. Actualmente la herramienta está disponible en un sitio web al cual el usuario puede ingresar y analizar las series de tiempo que este desee.
- Se garantiza un rendimiento óptimo de la herramienta dado que está hospedada en un sitio web en el cual se almacenan otras herramientas similares. Este sitio cuenta con servidores dedicados al procesamiento de información, dando la ventaja de que el rendimiento de la herramienta no dependa del equipo que esté utilizando el usuario.
- No hay necesidad de que el usuario tenga instalado R en su equipo. Al contar con una herramienta web, lo único necesario es un equipo de cómputo y conexión a internet, lo que hace que la herramienta sea mucho más accesible a más usuarios.

Pendientes:

- La herramienta actualmente acepta únicamente archivos **".csv"**. Además no tiene la capacidad de reestructurar el archivo de datos, en caso de que las series de tiempo no estén en columnas. Por ejemplo, el caso de un supermercado que tiene información de las ventas realizadas y la base de datos contiene 3 columnas: fecha, producto y cantidad vendida. Por lo que puede ser de utilidad en un futuro desarrollar una funcionalidad que permita trabajar este tipo de archivos, de manera que no sea un requisito que las series de tiempo estén estructuradas en columnas.
- A pesar de que se realizaron pruebas con distintos usuarios para ingresar a la herramienta desde distintos lugares y en forma simultánea, no se ha podido simular un escenario en el

que se realice un volumen considerable de consultas de manera simultánea, que pudieran llegar a afectar el rendimiento de la página web. Por tal razón, no se tiene seguridad del desempeño de la herramienta en ese escenario. No obstante, en caso de ser necesario obtener una herramienta más robusta, se deberá recurrir a una licencia pagada de Shinyapps (ver listado de precios en los anexos).

- Se debe evaluar en un futuro la calidad de los pronósticos recomendados por la herramienta, por lo que es importante repetir este ejercicio en otros campos de estudio y generar una medición que permita conocer la capacidad predictiva de la herramienta.
- Se debe valorar la posibilidad de incluir nuevas técnicas de pronóstico que puedan aumentar la precisión de los valores estimados. También puede ser importante explorar nuevas técnicas para los procesos de detección de valores extremos e imputación de valores perdidos, de manera que se pueda generar un impacto positivo a la calidad de los pronósticos generados por la herramienta.

XII. ANEXOS

12.1 Manual de uso

MANUAL DE USO HERRAMIENTA PARA CALIBRACIÓN DE SERIES DE TIEMPO BAJO UN ENFOQUE DE MINERÍA DE DATOS

1. Contar con un archivo de datos estructurado de la siguiente manera:

PERIODO	CIRUGIA	CIR_AMB	GINE_OBS	MEDICINA	PEDIATRIA	PSIQUIATRIA
1997-01	4568	1272	10277	3588	4878	377
1997-02	4711	1306	9422	3470	4579	406
1997-03	5015	1249	10228	3574	4755	423
1997-04	5173	1455	10091	3664	4792	422
1997-05	5188	1674	10720	3845	4908	421
1997-06	4916	1584	10181	3703	4681	394
1997-07	5040	1580	10383	3958	4609	442
1997-08	5227	1579	10470	3777	4532	475
1997-09	5175	1775	11033	3882	4914	434
1997-10	5661	1891	11528	3978	5406	415
1997-11	5218	1611	10904	3701	5204	412
1997-12	4785	1157	10361	4076	4743	430
1998-01	4840	1537	10243	3587	4801	389
1998-02	4781	1630	9248	3246	4564	392

Nota: La primera columna corresponde al periodo la cual debe estar ordenada en forma ascendente (de más antiguo a más reciente), mientras que de la columna dos en adelante deben contener la información de las variables que se desean pronosticar. En caso de tener otra periodicidad, la fecha no debe tener un formato específico, basta con que esté ordenada de manera ascendente y que no haya faltante de registros, en caso de no contar con el dato de algún periodo, la línea deberá incluirse vacía y la herramienta se encargará de estimar el valor faltante.

2. Ingresar a la página web donde se ubica la herramienta:

<https://gorderom.shinyapps.io/Pronostica/>

3. Al abrir la página, en el "**Panel Principal**" se mostrará un ejemplo correspondiente al total mensual de pasajeros de líneas aéreas internacionales (1949-1960). En la izquierda, se encuentra un panel con las especificaciones necesarias para que la serie sea modelada como el usuario lo defina.

Se puede seleccionar el tipo de datos que se va a utilizar (ejemplo o datos propios), si se desea corregir valores extremos (por defecto Sí), columna a analizar, periodicidad de la serie, año y periodo inicial de la serie, cantidad de periodos para entrenamiento (ventana de desempeño) y pronóstico y por último, los intervalos de confianza que se van a utilizar para generar los pronósticos.

The screenshot shows a configuration interface for time series analysis. On the left, there's a sidebar with several sections:

- Tipo de datos:** A radio button group where "Archivo de ejemplo" is selected.
- Especificaciones:** A section containing "Corregir outliers?" with "Sí" selected, and "Columna a analizar" set to "2".
- Periodicidad de la serie:** A radio button group where "Mensual" is selected.

The main area contains the following settings:

- Año inicial:** A text input field showing "2000".
- Primer mes:** A slider scale from 1 to 12 with a handle at position 1.
- Periodos para entrenamiento:** A text input field showing "3".
- Cantidad de pronósticos:** A text input field showing "3".
- Intervalos de confianza (80%-99%):** A slider scale from 0.8 to 0.99 with a handle at position 0.9.

4. Para analizar un archivo de datos propio, en "**Tipo de datos**" seleccione "**Archivo de datos**" e indique las especificaciones necesarias.

Tipo de datos

Archivo de ejemplo
 Archivo de datos

Elija el archivo

Browse... No file selected

Leer nombres de campo del archivo

Separador

Coma
 Punto y coma
 Tabulador

Comillas

Ninguna
 Doble comilla
 Comilla simple

Especificaciones

Corregir valores extremos y faltantes?

Sí
 No

Una vez que la herramienta lee el archivo de datos, se genera una vista previa de los primeros registros.

HERRAMIENTA PARA CALIBRACIÓN DE SERIES DE TIEMPO BAJO UN ENFOQUE DE MINERÍA DE DATOS

Panel Principal Holt-Winters ARIMA Pronósticos Acerca de

Tipo de datos

Archivo de ejemplo
 Archivo de datos

Elija el archivo

Browse... Info de hospitalizaciones.
Upload complete

Header

Mostrar los primeros casos

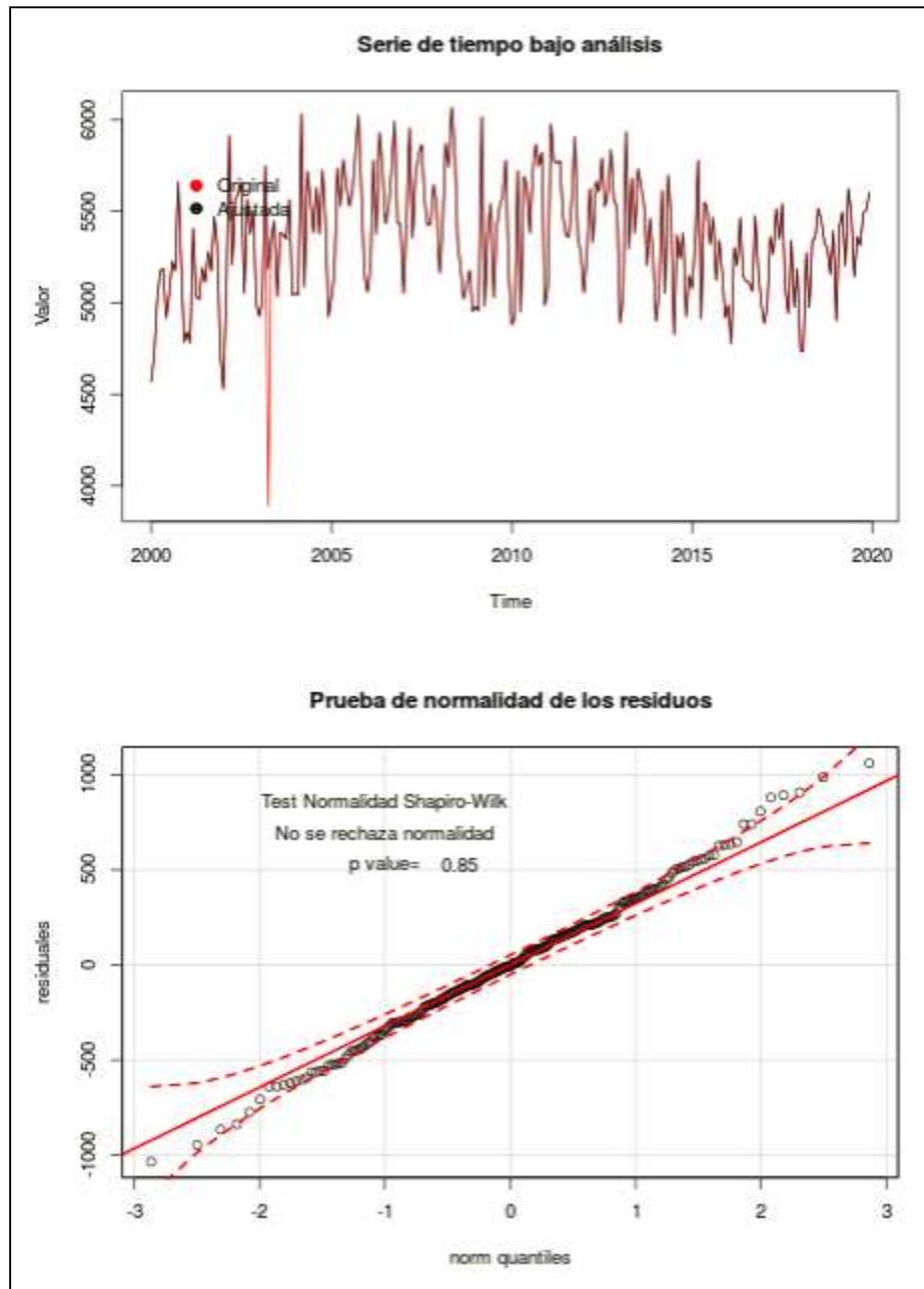
5

Lectura de series

PERIODO	CIRUGIA	CIR_AMB	GINE_OBS	MEDICINA	PEDIATRIA	PSIQUIATRIA
199701	4568	1272	10277	3588	4878	377
199702	4711	1306	9422	3470	4579	406
199703	5015	1249	10228	3574	4755	423
199704	5173	1455	10091	3664	4792	422
199705	5188	1674	10720	3845	4908	421

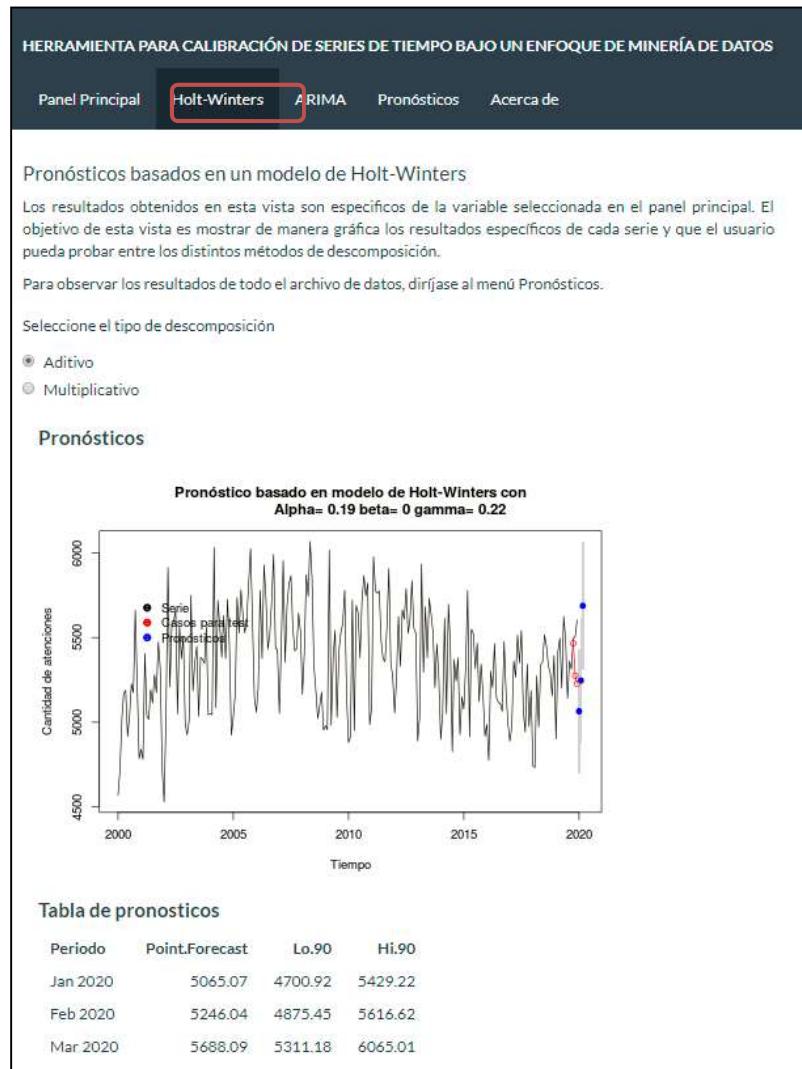
5. Análisis inicial del comportamiento de las variables. La siguiente vista permite realizar dos tareas:

- Analizar la serie de tiempo: en este gráfico se puede ver la serie original vs la serie ajustada sin valores extremos y determinar si se justifica realizar estos ajustes.
- Evaluar el supuesto de la normalidad de los residuos: dentro del gráfico se incluye una leyenda en la cual se indica si se cumple o no el supuesto de normalidad, esto basado en el Test de Normalidad Shapiro-Wilk.

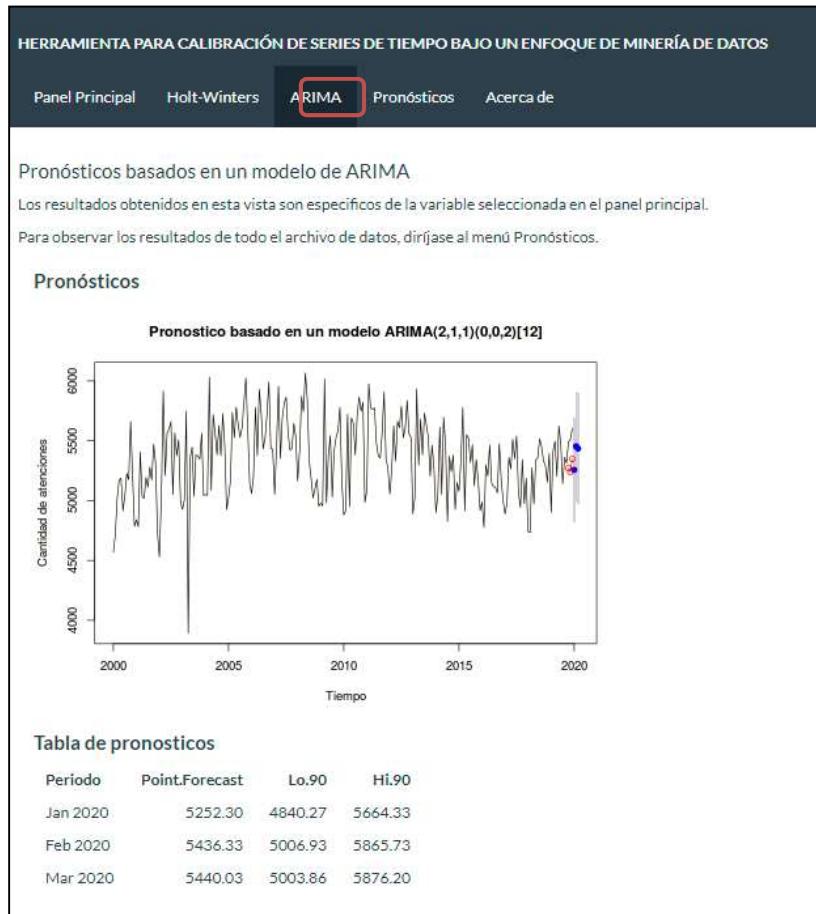


- Análisis de los resultados mediante Holt-Winters. Se encuentra en la sección "**Holt-Winters**" y están disponibles los dos métodos de descomposición: aditivo y

multiplicativo. En los pronósticos finales, la herramienta determina cuál genera los resultados con un menor término de error.



7. Análisis de los resultados mediante ARIMA, se encuentra en la sección "**ARIMA**".



8. Obtención de los pronósticos para todas las variables, panel de "Pronósticos".

HERRAMIENTA PARA CALIBRACIÓN DE SERIES DE TIEMPO BAJO UN ENFOQUE DE MINERÍA DE DATOS

Panel Principal Holt-Winters ARIMA Pronósticos Acerca de

Pronósticos de todas las series

En la siguiente tabla se encuentran los pronósticos para cada una de las series de tiempo presente en el archivo de datos.

Para cada serie se generan dos pronósticos diferentes, uno basado en un modelo de Holt-Winters y otro obtenido mediante un ARIMA, la columna "Método Recomendado" indica cual de los dos pronósticos tiene un menor margen de error según los datos de entrenamiento.

[Descargar pronósticos](#)

Tabla de pronósticos

Periodo	Método Pronóstico	Método Recomendado	Variable	Pronóstico	LI 90 %	LS 90 %
Jan 2020	Holt-Winters	Holt-Winters	CIRUGIA	5065.10	4700.90	5429.20
Feb 2020	Holt-Winters	Holt-Winters	CIRUGIA	5246.00	4875.50	5616.60
Mar 2020	Holt-Winters	Holt-Winters	CIRUGIA	5688.10	5311.20	6065.00
Jan 2020	ARIMA	Holt-Winters	CIRUGIA	5252.30	4840.27	5664.33
Feb 2020	ARIMA	Holt-Winters	CIRUGIA	5436.33	5006.93	5865.73
Mar 2020	ARIMA	Holt-Winters	CIRUGIA	5440.03	5003.86	5876.20
Jan 2020	Holt-Winters	Holt-Winters	CIR_AMB	5311.60	4786.60	5836.60
Feb 2020	Holt-Winters	Holt-Winters	CIR_AMB	6002.30	5468.30	6536.20
Mar 2020	Holt-Winters	Holt-Winters	CIR_AMB	6217.90	5675.20	6760.70
Jan 2020	ARIMA	Holt-Winters	CIR_AMB	5249.52	4612.73	5886.30
Feb 2020	ARIMA	Holt-Winters	CIR_AMB	5978.14	5311.61	6644.68
Mar 2020	ARIMA	Holt-Winters	CIR_AMB	5797.60	5121.32	6473.89
Jan 2020	Holt-Winters	ARIMA	GINE_OBS	8632.40	8244.00	9020.90
Feb 2020	Holt-Winters	ARIMA	GINE_OBS	8015.60	7609.70	8421.50

Como se puede ver en la tabla anterior, la herramienta genera un pronóstico para cada técnica, pero recomienda uno por encima del otro, en este caso Holt-Winters.

En caso de requerir más pronósticos, basta con ir al "**Panel Principal**" y aumentar el número de pronósticos que se desean. Se recomienda utilizar la misma cantidad de periodos para entrenamiento, que los que se desea pronosticar.

Periodos para entrenamiento

Cantidad de pronósticos

En caso de modificar alguna especificación en el modelo que pueda provocar que el modelo se deba generar nuevamente, es importante tener presente que esto llevará tiempo de

procesamiento, por lo que se mostrará un mensaje como el que se muestra en el recuadro inferior.

HERRAMIENTA PARA CALIBRACIÓN DE SERIES DE TIEMPO BAJO UN ENFOQUE DE MINERÍA DE DATOS

Panel Principal Holt-Winters ARIMA Pronósticos Acerca de

Pronósticos de todas las series

En la siguiente tabla se encuentran los pronósticos para cada una de las series de tiempo presente en el archivo de datos.

Para cada serie se generan dos pronósticos diferentes, uno basado en un modelo de Holt-Winters y otro obtenido mediante un ARIMA, la columna "Método Recomendado" indica cual de los dos pronósticos tiene un menor margen de error según los datos de entrenamiento.

Tabla de pronósticos

Periodo	Método Pronóstico	Método Recomendado	Variable	Pronóstico	LI 90 %	LS 90 %
Jan 2020	Holt-Winters	Holt-Winters	CIRUGIA	5065.10	4700.90	5429.20
Feb 2020	Holt-Winters	Holt-Winters	CIRUGIA	5246.00	4875.50	5616.60
Mar 2020	Holt-Winters	Holt-Winters	CIRUGIA	5688.10	5311.20	6065.00
Jan 2020	ARIMA	Holt-Winters	CIRUGIA	5252.30	4840.27	5664.33
Feb 2020	ARIMA	Holt-Winters	CIRUGIA	5436.33	5006.93	5865.73
Mar 2020	ARIMA	Holt-Winters	CIRUGIA	5440.03	5003.86	5876.20
Jan 2020	Holt-Winters	Holt-Winters	CIR_AMB	5311.60	4786.60	5836.60
Feb 2020	Holt-Winters	Holt-Winters	CIR_AMB	6002.30	5468.30	6536.20
Mar 2020	Holt-Winters	Holt-Winters	CIR_AMB	6217.90	5675.20	6760.70
Jan 2020	ARIMA	Holt-Winters	CIR_AMB	5249.52	4612.73	5886.30
Feb 2020	ARIMA	Holt-Winters	CIR_AMB	5978.14	5311.61	6644.68
Mar 2020	ARIMA	Holt-Winters	CIR_AMB	5797.60	5121.32	6473.89

Actualizando pronósticos

9. Descarga de una tabla con los pronósticos.

HERRAMIENTA PARA CALIBRACIÓN DE SERIES DE TIEMPO BAJO UN ENFOQUE DE MINERÍA DE DATOS

Panel Principal Holt-Winters ARIMA Pronósticos Acerca de

Pronósticos de todas las series

En la siguiente tabla se encuentran los pronósticos para cada una de las series de tiempo presente en el archivo de datos.

Para cada serie se generan dos pronósticos diferentes, uno basado en un modelo de Holt-Winters y otro obtenido mediante un ARIMA, la columna "Método Recomendado" indica cual de los dos pronósticos tiene un menor margen de error según los datos de entrenamiento.

12.2 Sintaxis en R

12.2.1 Archivo ui.R

```

library(shiny)
library(shinythemes)
shinyUI(navbarPage(
  theme = shinytheme("cyborg"),
  h5(strong("HERRAMIENTA PARA CALIBRACIÓN DE SERIES DE TIEMPO BAJO UN ENFOQUE DE MINERÍA DE DATOS")),

  tabPanel("Panel Principal",
    sidebarLayout(
      sidebarPanel(
        radioButtons("RD1",label=h4(strong("Tipo de datos")),choices = list("Archivo de ejemplo"=1,"Archivo de datos"=2),selected = 1),
        conditionalPanel(condition="input.RD1==2",fileInput('file1', h4(strong('Elija el archivo'))),
                      accept=c('text/csv',
                               'text/comma-separated-values,text/plain',
                               '.csv')),
        checkboxInput('header', 'Header', TRUE),
        radioButtons('sep', 'Separator',
                    c(Comma=',',
                      Semicolon=';',
                      Tab='\t'),
                    ','),
        radioButtons('quote', 'Quote',
                    c(None='',
                      'Double Quote'="",
                      'Single Quote'=""),
                    ""))
      ,
      h4(strong("Especificaciones")),
      radioButtons("outliers",label=(strong("Corregir outliers?")),choices = list("Si"=1,"No"=2),selected = 2),
      numericInput("Col",strong("Columna a analizar"),value = 2),
      radioButtons("freq", label = strong("Periodicidad de la serie"),
                  choices = list("Mensual" = 12,"Trimestral" = 4,"Semestral"=2,"Anual"=1),
                  selected = 12),
      numericInput("Start",strong("Año inicial"),value = 2000),
      conditionalPanel(condition="input.freq==12",
                     sliderInput("Start2",label="Primer mes",min=1,max=12,value=1)),
      conditionalPanel(condition="input.freq==4",
                     sliderInput("Start2",label="Primer trimestre",min=1,max=4,value=1)),
      conditionalPanel(condition="input.freq==2",
                     sliderInput("Start2",label="Primer semestre",min=1,max=2,value=1)),
      conditionalPanel(condition="input.freq==1",
                     sliderInput("Start2",label="No aplica",min=1,max=1,value=1)),
      numericInput("ntesting",strong("Periodos para entrenamiento"),value = 3),
      numericInput("pronosticos",strong("Cantidad de pronósticos"),value = 3),
      sliderInput("CI2",label="Intervalos de confianza (80%-99%)",min=0.8,max=0.99,value=0.9,step = 0.01
    ),
    mainPanel(fluidRow(
      numericInput("head",strong("Mostrar los primeros casos"),value = 5),

```

```

column(10,h3("Lectura de series"),tableOutput("tableResumen")),
column(10,h3("Plot"),plotOutput("PlotVariable")),
column(10,h3("Plot"),plotOutput("plot_Shapiro"))
)

),
tabPanel("Holt-Winters",
mainPanel(
fluidRow(
h4('Pronósticos basados en un modelo de Holt-Winters'),
p('Los resultados obtenidos en esta vista son específicos de la variable seleccionada en el panel principal. El objetivo de esta vista es mostrar de manera gráfica los resultados específicos de cada serie y que el usuario pueda probar entre los distintos métodos de descomposición.',align='Justify'),
p('Para observar los resultados de todo el archivo de datos, diríjase al menú Pronósticos.',align='Justify'),
radioButtons("AM",h5("Seleccione el tipo de descomposición"),choices=list("Aditivo"=1,"Multiplicativo"=2),selected=1),
column(10,h4(strong('Pronósticos')),h3(textOutput("text_HW"))),
column(10,plotOutput("Plot_HW")),
column(10,h4(strong('Tabla de pronósticos')),tableOutput("table_HW"))
)
)
),
tabPanel("ARIMA",
mainPanel(
fluidRow(
h4('Pronósticos basados en un modelo de ARIMA'),
p('Los resultados obtenidos en esta vista son específicos de la variable seleccionada en el panel principal.',align='Justify'),
p('Para observar los resultados de todo el archivo de datos, diríjase al menú Pronósticos.',align='Justify'),
column(10,h4(strong('Forecasting Plot')),plotOutput("Plot_ARIMA")),
column(10,h4(strong('Tabla de pronósticos')),tableOutput("table_ARIMA"))
)
)
),
tabPanel("Pronósticos",
fluidRow(
h4('Pronósticos de todas las series'),
p('En la siguiente tabla se encuentran los pronósticos para cada una de las series de tiempo presente en el archivo de datos.',align='Justify'),
p('Para cada serie se generan dos pronósticos diferentes, uno basado en un modelo de Holt-Winteres y otro obtenido mediante un ARIMA, la columna "Método Recomendado indica cual de los dos pronósticos tiene un menor margen de error según los datos de entrenamiento.',align='Justify'),
h4('**Los resultados pueden tardar algunos minutos actualizándose dependiendo el volumen de datos que se estén estimando.'),
p(downloadButton('x3', 'Descargar pronósticos'))
),
fluidRow(
column(10,h4("Tabla de pronósticos"),tableOutput("Table_Pronosticos"))
)
),
tabPanel("Acerca de",

```

```
h3(strong("Descripcion")),
br(),
p("T1",align="Justify"),
p("T2",align="Justify"),
p("T3",align="Justify"),

p("T4",align="Justify"),
br(),
h3(strong("Author")),
p("Author: Gabriel Cordero Mora"),
p("Estudiante de Maestria Profesional en Estadística"),
p("Universidad de Costa Rica"),
p("Email: gabro_cor85@hotmail.com")

)
)
)
```

12.2.2 Archivo server.R

```

if(input$outliers==1)
  TotalTS<-tsclean(ts(serie[,i],start=c(input$Start,input$Start2),frequency = as.numeric(input$freq)))
else
  TotalTS<-ts(serie[,i],start=c(input$Start,input$Start2),frequency = as.numeric(input$freq))

TotalTest1<-(serie[dim(serie)[1]-input$ntesting+1:dim(serie)[1],i])[1:input$ntesting]

serieaprende<-window(TotalTS,start = input$Start,end = c(input$Start+(length(TotalTS)-input$ntesting-1)/as.numeric(input$freq)),frequency = as.numeric(input$freq))

modHW1 <- HoltWinters(serieaprende);predHW1 <- predict(modHW1, n.ahead = input$ntesting)
modHW2 <- HoltWinters(serieaprende,seasonal="multiplicative");predHW2 <- predict(modHW2, n.ahead = input$ntesting)

ECM1<-sqrt(ECM(predHW1, TotalTest1));ECM2<-sqrt(ECM(predHW2, TotalTest1))
ER1<-ER(predHW1, TotalTest1);ER2<-ER(predHW2, TotalTest1)
MAPE1<-MAPE(predHW1, TotalTest1);MAPE2<-MAPE(predHW2, TotalTest1)
MAD1<-MAD(predHW1, TotalTest1);MAD2<-MAD(predHW2, TotalTest1)

# Comparacion
ComparaECM <- rbind(ECM1, ECM2)/max(ECM1, ECM2)
ComparaER <- rbind(ER1, ER2)/max(ER1, ER2)
ComparaMAPE <- rbind(MAPE1, MAPE2)/max(MAPE1, MAPE2)
ComparaMAD <- rbind(MAD1, MAD2)/max(MAD1, MAD2)

Compara <- as.data.frame(cbind(ComparaECM,ComparaER,ComparaMAPE,ComparaMAD))

seasonV <- if (sum(Compara[1,])<=sum(Compara[2,])) "additive" else "multiplicative"
#####
modHW <- HoltWinters(serieaprende,seasonal=seasonV)
predHW <- predict(modHW, n.ahead = input$ntesting)

p.arima<-auto.arima(serieaprende)
p<-p.arima$arma[1]
d<-p.arima$arma[6]
q<-p.arima$arma[2]
P<-p.arima$arma[3]
D<-p.arima$arma[7]
Q<-p.arima$arma[4]
PP<-p.arima$arma[5]
modArima<-arima(serieaprende,order=c(p,d,q),seasonal=list(order=c(P,D,Q),period=PP))
predArima <- predict(modArima, n.ahead = input$ntesting)

#MODELOS
modelofinalHW <- HoltWinters(TotalTS,
                                alpha = if(modHW$alpha==0) NULL else modHW$alpha,
                                beta = if(modHW$beta==0) NULL else modHW$beta,
                                gamma = if(modHW$gamma==0) NULL else modHW$gamma)

```

```

        )
modeloFinalArima<-arima(TotalTS,order=c(p,d,q),seasonal=list(order=c(P,D,Q),period=PP))

ECM1<-sqrt(ECM(predHW, TotalTest1));ECM2<-sqrt(ECM(predArima$pred, TotalTest1))
ER1<-ER(predHW, TotalTest1);ER2<-ER(predArima$pred, TotalTest1)
MAPE1<-MAPE(predHW, TotalTest1);MAPE2<-MAPE(predArima$pred, TotalTest1)
MAD1<-MAD(predHW, TotalTest1);MAD2<-MAD(predArima$pred, TotalTest1)

# Comparacion
ComparaECM <- rbind(ECM1, ECM2)/max(ECM1, ECM2)
ComparaER <- rbind(ER1, ER2)/max(ER1, ER2)
ComparaMAPE <- rbind(MAPE1, MAPE2)/max(MAPE1, MAPE2)
ComparaMAD <- rbind(MAD1, MAD2)/max(MAD1, MAD2)

Compara <- as.data.frame(cbind(ComparaECM,ComparaER,ComparaMAPE,ComparaMAD))

#CREACION DE LA TABLA PARA EXPORTAR LOS PRONOSTICOS
Metodo_Recomendado <- if (sum(Compara[1,])<=sum(Compara[2,])) "Holt-Winters" else "ARIMA"
#Metodo_Recomendado <- "ARIMA"
A1<-data.frame(forecast(modeloFinalHW,h=input$pronosticos,level=input$CI2))
A<-cbind(rownames(A1),"Holt-Winters",Metodo_Recomendado,names(serie)[i],round(A1,1))
colnames(A)[1]<-"Periodo"
colnames(A)[2]<-"Metodo Pronóstico"
colnames(A)[3]<-"Metodo Recomendado"
colnames(A)[4]<-"Variable"
colnames(A)[5]<-"Pronóstico"
colnames(A)[6]<-paste("LI",input$CI2*100,"%")
colnames(A)[7]<-paste("LS",input$CI2*100,"%")
B1<-data.frame(forecast(modeloFinalArima,h=input$pronosticos,level=input$CI2))
B<-cbind(rownames(B1),"ARIMA",Metodo_Recomendado,names(serie)[i],B1)
colnames(B)[2]<-"Metodo Pronóstico"
colnames(B)[3]<-"Metodo Recomendado"
colnames(B)[1]<-"Periodo"
colnames(B)[4]<-"Variable"
colnames(B)[5]<-"Pronóstico"
colnames(B)[6]<-paste("LI",input$CI2*100,"%")
colnames(B)[7]<-paste("LS",input$CI2*100,"%")

RESULTADO_I<-rbind(A,B)
RESULTADO<-if (i==2) RESULTADO_I else rbind(RESULTADO,RESULTADO_I)

}

RESULTADO

})

output$tableResumen<-renderTable({
  MyData()[1:input$head,]
})
output$PlotVariable<-renderPlot({

```

```

serie_O<-ts(MyData()[,input$Col],start=input$Start,frequency = as.numeric(input$freq))

if(input$outliers==1)
  serie<-tsclean(ts(MyData()[,input$Col],start=input$Start,frequency = as.numeric(input$freq)))
else
  serie<-ts(MyData()[,input$Col],start=input$Start,frequency = as.numeric(input$freq))

if(is.null(MyData())!=T){
  plot(serie_O,ylab="Valor",col="red",main=c("Serie de tiempo bajo análisis"))+lines(serie)
  legend(input$Start+1, max(serie)*0.95, legend = c("Original","Ajustada"), seg.len = 0.5,
    pch = 21, bty = "n", lwd = 3, y.intersp = 1, horiz = FALSE, col = c("red", "black"))
}
})

output$plot_Shapiro<-renderPlot({

  serie_O<-ts(MyData()[,input$Col],start=input$Start,frequency = as.numeric(input$freq))

  if(input$outliers==1)
    serie<-tsclean(ts(MyData()[,input$Col],start=input$Start,frequency = as.numeric(input$freq)))
  else
    serie<-ts(MyData()[,input$Col],start=input$Start,frequency = as.numeric(input$freq))

  qqPlot(diff(serie),ylab="residuales",main="Prueba de normalidad de los residuos")
  text(-1,max(diff(serie))*0.8,"Test Normalidad Shapiro-Wilk")
  shap<-shapiro.test(diff(serie))
  text(-1,max(diff(serie))*0.65,
    if (shap$p.value > 0.05 ) "No se rechaza normalidad" else "Se rechaza la hipótesis de normalidad"
  )
  text(-1,max(diff(serie))*0.5,"p value=")
  text(-0.4,max(diff(serie))*0.5,round(shap$p.value,2))

}
)

output$text_HW<-renderText({
  names(MyData()[,input$Col])
})

output$Plot_HW<-renderPlot{

  # Create a Progress object
  style <- isolate('notification')
  progress <- shiny::Progress$new(style = style)
  progress$set(message = "Creando gráfico", value = 0.5)
  on.exit(progress$close())

  if(input$outliers==1)
    TotalTS<-tsclean(ts(MyData()[,input$Col],start=input$Start,frequency = as.numeric(input$freq)))
  else
    TotalTS<-ts(MyData()[,input$Col],start=input$Start,frequency = as.numeric(input$freq))
}

```

```

serieaprende<-window(TotalTS,start = input$Start,end = c(input$Start+(length(TotalTS)-input$ntesting-1)/as.numeric(input$freq)),frequency = as.numeric(input$freq))
if(input$AM==1)
  modHW <- HoltWinters(serieaprende)
else
  modHW <- HoltWinters(serieaprende,seasonal="multiplicative")

predHW <- predict(modHW, n.ahead = input$ntesting)
modelofinalHW <- HoltWinters(TotalTS,
                               alpha = if(modHW$alpha==0) NULL else modHW$alpha,
                               beta = if(modHW$beta==0) NULL else modHW$beta,
                               gamma = if(modHW$gamma==0) NULL else modHW$gamma)

plot(forecast(modelofinalHW,h=input$pronosticos,level=input$CI2),
      main=paste("Pronóstico basado en modelo de Holt-Winters con

Alpha=",round(modHW$alpha,2),"beta=",round(modHW$beta,2),"gamma=",round(modHW$gamma,2)),
      ylab="Cantidad de atenciones",xlab="Tiempo")
lines(predHW,type="o",col="red")
legend(input$Start+1, max(TotalTS)*0.95, legend = c("Serie","Casos para test","Pronósticos"), seg.len = 0.5,
       pch = 21, bty = "n", lwd = 3, y.intersp = 1, horiz = FALSE, col = c("black","red","blue"))
)}

output$table_HW<-renderTable({

# Create a Progress object
style <- isolate('notification')
progress <- shiny::Progress$new(style = style)
progress$set(message = "Creando tabla", value = 0.5)
on.exit(progress$close())

if(input$outliers==1)
  TotalTS<-tsclean(ts(MyData()[,input$Col],start=input$Start,frequency = as.numeric(input$freq)))
else
  TotalTS<-ts(MyData()[,input$Col],start=input$Start,frequency = as.numeric(input$freq))

serieaprende<-window(TotalTS,start = input$Start,end = c(input$Start+(length(TotalTS)-input$ntesting-1)/as.numeric(input$freq)),frequency = as.numeric(input$freq))

if(input$AM==1)
  modHW <- HoltWinters(serieaprende)
else
  modHW <- HoltWinters(serieaprende,seasonal="multiplicative")

predHW <- predict(modHW, n.ahead = input$ntesting)
modelofinalHW <- HoltWinters(TotalTS,
                               alpha = if(modHW$alpha==0) NULL else modHW$alpha,
                               beta = if(modHW$beta==0) NULL else modHW$beta,
                               gamma = if(modHW$gamma==0) NULL else modHW$gamma)

pron_HW<-data.frame(forecast(modelofinalHW,h=input$pronosticos,level=input$CI2))
pron_HW<-cbind(rownames(pron_HW),pron_HW)
colnames(pron_HW)[1]<-"Periodo"
}

```

```

colnames(pron_HW)[2]<-"Pronóstico"
pron_HW
})

output$Plot_ARIMA<-renderPlot({

# Create a Progress object
style <- isolate('notification')
progress <- shiny::Progress$new(style = style)
progress$set(message = "Creando gráfico", value = 0.5)
on.exit(progress$close())

if(input$outliers==1)
  TotalTS<-tsclean(ts(MyData()[,input$Col],start=input$Start,frequency = as.numeric(input$freq)))
else
  TotalTS<-ts(MyData()[,input$Col],start=input$Start,frequency = as.numeric(input$freq))
serieaprende<-window(TotalTS,start = input$Start,end = c(input$Start+(length(TotalTS)-input$ntesting-1)/as.numeric(input$freq)),frequency = as.numeric(input$freq))

p.arima<-auto.arima(serieaprende)
p<-p.arima$arma[1]
d<-p.arima$arma[6]
q<-p.arima$arma[2]
P<-p.arima$arma[3]
D<-p.arima$arma[7]
Q<-p.arima$arma[4]
PP<-p.arima$arma[5]
modArima<-arima(serieaprende,order=c(p,d,q),seasonal=list(order=c(P,D,Q),period=PP))
predArima <- predict(modArima, n.ahead = input$ntesting)
modelofinalArima<-arima(TotalTS,order=c(p,d,q),seasonal=list(order=c(P,D,Q),period=PP))

plot(forecast(modelofinalArima,h=input$pronosticos,level=input$CI2),
      main=paste("Pronostico basado en un modelo ARIMA(",p,",",d,",",q,")(",P,",",D,",",Q,")[,PP,]",sep=""),ylab="Cantidad de atenciones",xlab="Tiempo")
lines(predArima$pred,type="o",col="red")

})

output$table_ARIMA<-renderTable({

# Create a Progress object
style <- isolate('notification')
progress <- shiny::Progress$new(style = style)
progress$set(message = "Creando tabla", value = 0.5)
on.exit(progress$close())

if(input$outliers==1)
  TotalTS<-tsclean(ts(MyData()[,input$Col],start=input$Start,frequency = as.numeric(input$freq)))
else
  TotalTS<-ts(MyData()[,input$Col],start=input$Start,frequency = as.numeric(input$freq))

serieaprende<-window(TotalTS,start = input$Start,end = c(input$Start+(length(TotalTS)-input$ntesting-1)/as.numeric(input$freq)),frequency = as.numeric(input$freq))

```

```

p.arima<-auto.arima(serieaprende)
p<-p.arima$arma[1]
d<-p.arima$arma[6]
q<-p.arima$arma[2]
P<-p.arima$arma[3]
D<-p.arima$arma[7]
Q<-p.arima$arma[4]
PP<-p.arima$arma[5]
modArima<-arima(serieaprende,order=c(p,d,q),seasonal=list(order=c(P,D,Q),period=PP))
predArima <- predict(modArima, n.ahead = input$ntesting)
modelofinalArima<-arima(TotalTS,order=c(p,d,q),seasonal=list(order=c(P,D,Q),period=PP))

pron_ARIMA<-data.frame(forecast(modelofinalArima,h=input$pronosticos,level=input$CI2))
pron_ARIMA<-cbind(rownames(pron_ARIMA),pron_ARIMA)
colnames(pron_ARIMA)[1]<-"Periodo"
colnames(pron_ARIMA)[1]<-"Pronóstico"
pron_ARIMA

})

output$Table_Pronosticos<-renderTable({

# Create a Progress object
style <- isolate('notification')
progress <- shiny::Progress$new(style = style)
progress$set(message = "Actualizando pronósticos", value = 0.5)
on.exit(progress$close())

MyPronosticos()
#datatable(MyPronosticos(), options = list(searchHighlight = TRUE), filter = 'top',)

})

# download the filtered data
output$x3 = downloadHandler('Pronosticos.csv', content = function(file) {
  s = input$x1_rows_all
  write.csv(MyPronosticos(), file)
})

})

```

12.3 Costo de licencias Shiny

Tabla 4. Costos de implementación de las distintas modalidades de licencias de Shiny.

LIBRE	VERSIÓN DE INICIO	BÁSICA	ESTÁNDAR	PROFESIONAL
\$ 0 / mes	\$ 9 / mes (o \$100 / año)	\$ 39 / mes (o \$440 / año)	\$ 99 / mes (o \$1.100 / año)	\$ 299 / mes (\$3.300 / año)
5 Aplicaciones	25 Aplicaciones	Aplicaciones Ilimitadas	Aplicaciones Ilimitadas	Aplicaciones Ilimitadas
25 horas activas	100 horas activas	500 horas activas	2.000 horas activas	10.000 horas activas
Soporte comunitario	Soporte Premium	Soporte Premium	Soporte Premium	Soporte Premium
		Aumento de rendimiento	Aumento de rendimiento	Aumento de rendimiento
			Autenticación	Autenticación
				Dominios personalizados

XIII. BIBLIOGRAFÍA

CCSS. (2000). *Estadísticas Generales de Consulta Externa*. Recuperado el 15 de 05 de 2016, de <http://ccssvdcapp03.ccss.sa.cr/bincre/RpWebEngine.exe/Portal?&BASE=CEXTER>

CRAN R. (s.f.). Recuperado el 3 de 8 de 2015, de <https://cran.r-project.org/web/packages/forecast/forecast.pdf>

Hernández Rodríguez, Ó. (2011). *Series Cronológicas*. San José: Editorial UCR.

Hyndman, R. J. (2015). *forecast: Forecasting functions for time series and linear models_*. R package version 6.2. Obtenido de <http://github.com/robjhyndman/forecast>

Hyndman, R. J., & Khandakar, Y. (2008). Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software* , 1-22.

Leonard, M. (2002). *Large-Scale Automatic Forecasting: Millions of Forecasts*.

Méllard, G., & Pastels, J.-M. (2000). Automatic ARIMA modeling including interventions, using time series expert software. *International Journal of Forecasting* , 497-508.

Meyer, D. (s.f.). Recuperado el 1 de 8 de 2015, de <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/HoltWinters.html>

Ord, K., & Lowe, S. (1996). Automatic Forecasting. *The American Statistician* , 88-94.

R_Core_Team. (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Obtenido de <https://www.R-project.org/>.

RSStudio Inc. (2009). Recuperado el 4 de 10 de 2017, de <https://www.rstudio.com/>

Sáenz, M. d., Acosta, M., Muiser, J., & Bermúdez, J. L. (2011). *Sistema de salud de Costa Rica*. México.

