Universidad Nacional Autónoma de México

Facultad de Estudios Superiores Acatlán

Series de Tiempo

Proyecto

Modelo de series de tiempo para
el subcampeonato del Cruz Azul en la Liga Mx.

Camargo Salas Mario Alberto
García Sánchez Cecilia Daniela
Ibarra Guerrero Javier Alonso
Rodríguez Becerril Kenia
Rojas Gutiérrez Rodolfo Emmanuel
Tirado Pellón Diana Karina

22 Mayo 2020

Índice

1. Introducción

Hacia el año de 1881, la antigua Hacienda de Jasso ubicada en el Estado de Hidalgo fue alquilada por Henry Gibbon, de origen británico, con el propósito de levantar una fábrica de cal hidráulica la cual años más tarde fue conocida como la Compañía Manufacturera de Cemento Portland "La Cruz Azul" en el año de 1909.

No obstante, hasta el año de 1964 se crea el Club Deportivo Social y Cultural Cruz Azul, el cual estaba integrado por trabajadores de la planta manufacturera, con el propósito de promover el desarrollo social e individual de los obreros de la fábrica. Posteriormente y bajo este contexto, Cruz Azul forma su primer equipo en la división amateur.

Su época de oro fue marcada a finales de los años 60 y la década de los años 70, en donde se apreciaba como uno de los mejores equipos de fútbol mexicano. Sin embargo, a principios de los años 80 y mediados de los 90, el Cruz Azul padeció la ausencia de títulos, fue hasta finales de los 90 cuando lograron obtener otro campeonato.

En el año de 1997 los celestes fueron victoriosos de levantar el título, sin embargo, el equipo no ha conseguido ganar un campeonato nuevamente y desde aquél momento ha perdido finales de manera inconcebible, ya que si bien este equipo es considerado uno de los "grandes"de la Liga Mx, únicamente ha sido capaz de acumular subcampeonatos de la primera división, sosteniendo el triste estigma de ser el equipo de fútbol mexicano que pierde el triunfo en los últimos minutos del partido.

Debido a la adversidad que ha enfrentado el Cruz Azul en los últimos años, en México se ha originado un nuevo verbo conocido como "cruzazulear", que de acuerdo con Brooks(2018) es "un sinónimo de quedarse a la orilla de conseguir un logro". Así como también se ha convertido en un tema de interés social, ya que la pregunta que se encuentra en el aire está relacionada con si el Cruz Azul será o no campeón nuevamente.

Adicionalmente, se sabe que el fútbol mexicano tiene factores internos y externos que intervienen en la vida cotidiana de un aficionado, esta influencia se puede presentar de manera positiva o negativa, debido a que el estado de ánimo de un aficionado de fútbol depende en gran medida de si su equipo gana o pierde en cuanto enciende el televisor, la radio o asiste al estadio para atender el evento deportivo en cuestión. Sin embargo, uno de los problemas deriva en la triste tendencia que el Cruz Azul ha tenido en los últimos años al quedar únicamente como subcampeón de la Liga Mx, ante esta situación, la pregunta que se ha generado es si este equipo de fútbol será al menos subcampeón una vez más.

No obstante, también se han originado inconformidades y poca credibilidad en el equipo por parte de la sociedad, debido a que el equipo de fútbol Cruz Azul más que ser partícipe en el deporte, éste se ha involucrado más en el negocio del fútbol.

Un claro ejemplo es que en 2019 se emitió un video como comunicado de prensa, en donde los presidentes de los Consejos de Administración y Vigilancia de la Cooperativa La Cruz Azul S.C.L. revelan que cada año ésta invierte más de mil 200 millones de pesos al equipo fútbol en cada temporada y aún así, el Cruz Azul no ha podido ganar un campeonato desde hace más de 22 años.

Por lo anterior y otras razónes más, se ha vuelto un tema de interés debido al impacto que se ha visto en la sociedad actual, en donde el tema sobre si será al menos subcampeón el Cruz Azul reincide en un problema "moderno", el cual ha causado sinfín de mofa o tristeza para algunos, no obstante el Cruz Azul es uno de los equipos con mayor popularidad en el fútbol mexicano; en el año 2019 una encuesta mostró que para el 9.8 % de la población encuestada, el Cruz Azul es el tercer favorito en México (Diario Récord, comunicación personal, 11 de junio de 2019).

Debido a la situación expuesta previamente y agregando que, el Cruz Azul es uno de los equipos a los que en los últimos años se le ha negado el campeonato de la Liga Mx en la fase final, es de interés analizar bajo modelos de series de tiempo, si el equipo de fútbol Cruz Azul puede ser al menos subcampeón de la Liga Mx. La hipótesis para este trabajo se centra en si el equipo de fútbol Cruz Azul puede ser al menos subcampeón de la Liga Mx en las siguientes diez temporadas.

Teniendo en cuenta que para la Clausura 2020, se hizo un estudio matemático que ha determinado las posibilidades de que el Cruz Azul gane esta temporada. De acuerdo con Redacción Vamos Cruz Azul(2020) "el portal FiveThirtyEight.com hizo una proyección basada en los números que arrojaron las diez Jornadas disputadas al momento. El estudio incluyó 36 ligas alrededor del mundo y, una vez que reunió las probabilidades para cada partido, ejecutó simulaciones para jugar la temporada de cada liga 20,000 veces usando esos pronósticos." Dicho estudio posiciona al Cruz Azul como el equipo con mayor probabilidad para ganar la Clausura 2020 con un 18 % de probabilidad por encima de Tigres con un 17 % y América con 13 %.

De forma que, el presente trabajo busca la implementación de modelos de series de tiempo asociados al desempeño que ha tenido el equipo de fútbol Cruz Azul a través de los años en los que se tienen registros, con el propósito de describir, modelar la probabilidad de que este equipo de fútbol sea al menos subcampeón de la Liga Mx y eventualmente pronosticar el fenómeno que subyace de la serie de tiempo para conocer de manera analítica si existe la probabilidad de que el Cruz Azul sea al menos subcampeón de la Liga Mx.

2. Análisis descriptivo de la base de datos

Para este trabajo, se ha consultado la base de datos de la página oficial de la Liga Mx `https://ligamx.net/`, con la cual se generaran modelos de series de tiempo para pronosticar si existe la probabilidad de que el Cruz Azul sea subcampeon en la Liga MX, dichos modelos se estudiarán desde un punto de vista analítico.

Dicha base de datos incluye los registros historicos de la trayectoria del Cruz Azul desde el año 1987 al año 2019. Teniendo en cuenta que a partir del segundo semestre 1996 el torneo en el que se disputaba el campeonato cambió la forma de realizarse, se decidió omitir los datos anteriores a esa fecha y trabajar únicamente con los correspondientes al año 1996 en adelante.

Esto da como resultado una 'nueva' base de datos que esta conformada por 12 columnas, entre ellas los partidos jugados, ganados, empatados y perdidos en cada uno de los torneos. Además, se encuentran los goles en contra acumulados en el torneo, así como los goles a favor, y la diferencia que existe entre ellos, por ejemplo, si el Cruz Azul anotó 26 goles y recibió 15 goles, entonces la diferencia entre ellos sería 11. Nótece que ésta operación admite resultados negativos, por lo que una cantidad negativa significaría que el club recibió más goles de los que anotó en ese torneo. Estos resultados permiten realizar algunos primeros cálculos, por ejemplo, se obtuvo que el valor esperado de dichas diferencias de goles resultó ser 5.89, así como una varianza de 50.6188.
Adicionalmente, se programó una función que calcula la distribución empírica de un conjunto de datos, así como la función de distribución empírica bayesiana con sus respectivos intervalos de confianza. Dichas funciones fueron usadas para estimar la distribución de las diferencias de goles, obteniendo la siguiente gráfica:

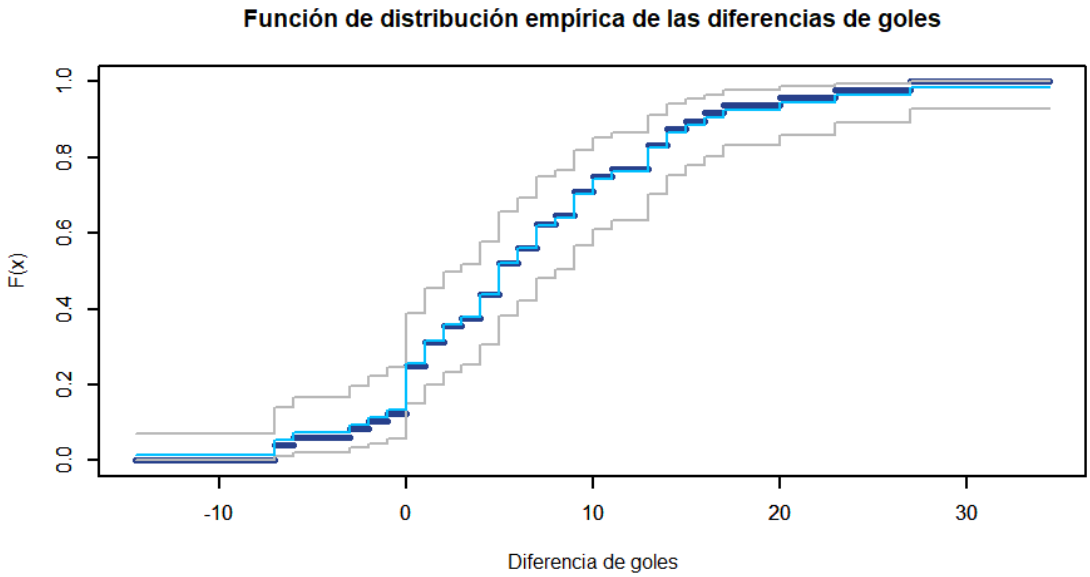

Figura 1: Función de distribución empírica de la diferencia de goles.

Donde la gráfica azul marino corresponde a la distribución empírica, la azul cielo a la bayesiana y las grises a los intervalos de confianza de la estimación bayesiana.

Continuando con el análisis de la base de datos, también se puede encontrar una columna que representa la cantidad total de puntos que consiguió el Cruz Azul en cada torneo. Ésta variable es de especial interés pues sin necesidad de hacer cálculos se sabe a priori que está estrechamente relacionada con el desempeño del equipo, por lo que también se realizaron los cálculos pertinentes, es decir, se obtuvo que esta variable tiene un valor esperado de 26.85 y una varianza de 30.52.

Al igual que con la diferencia de goles, también se utilizó la función programada en R para estimar la distribución empírica de los puntos obtenidos por el equipo, obteniendo la siguiente gráfica:

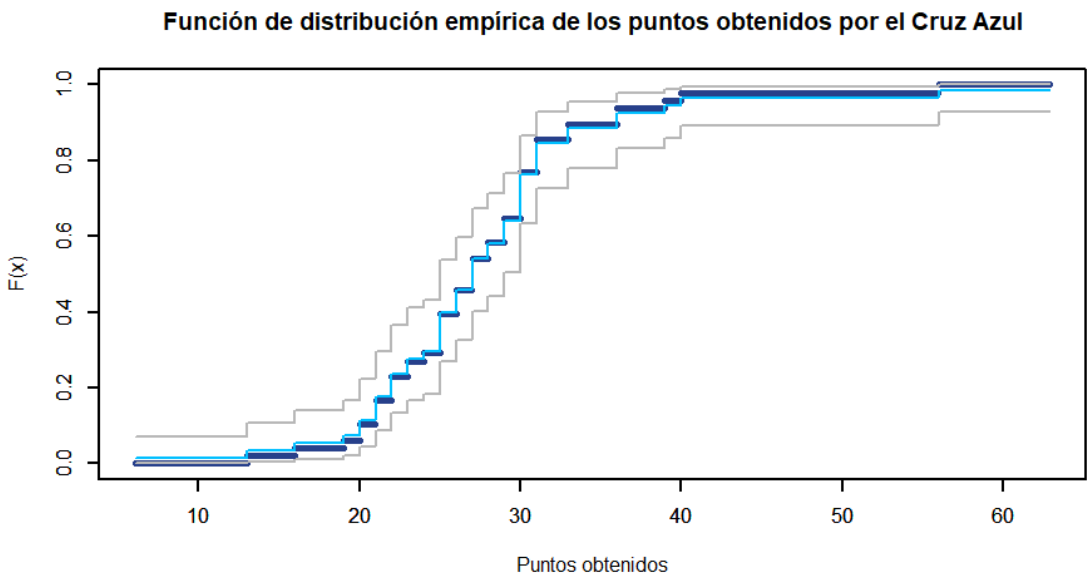

Figura 2: Función de distribución empírica de los puntos obtenidos.

Al igual que la figura 1, la gráfica azul marino corresponde a la distribución empírica, la azul cielo a la bayesiana y las grises a los intervalos de confianza de la estimación bayesiana. Finalmente, se tiene una columna con la posición que alcanzó el equipo en la fase de eliminatorias de cada torneo y una columna que indica el máximo logro que obtuvo el club, es decir, si fue campeón, subcampeón, si únicamente llegó a cuartos de final, o si desafortunadamente no consiguió clasificar. Dado que el objetivo de este trabajo es determinar la probabilidad de que el Cruz Azul sea al menos subcampeón, entonces ésta última columna es precisamente la variable objetivo.

Nótese que la variable objetivo no es numérica, es decir, sus valores no estan representados por números si no por palabras o 'categorías', es decir, se trata de una variable categórica la cuál es imposible de manipular con las técnicas de modelaje tradicionales, por lo que para poder tratarla como una serie de tiempo se debe someter a algún proceso de 'adaptación', y para lograr esto se decidió utilizar la siguiente técnica.

2.1. Modelo Logístico Multinomial

Por lo expuesto con anterioridad, se decidio crear un indice de eficiencia del equipo, basado en la variable categórica de interes, el cual tome valores en algún subconjunto no numerable, compacto

y convexo de $\mathbb{R}$. Para ello, se aplicó regresión logística multinomial, con el propósito de modelar la probabilidad de que el Cruz Azul llegara a ser campeón.

Para el modelo logístico multinomial, considérese la siguiente situación para K posibles resultados, para los cuales ejecutamos $K-1$ modelos binarios de regresión logística, en donde uno de los resultados se elige como pivote, mientras que el resto, los $K-1$, se comparan contra dicho pivote.

Si se considera al resultado K como pivote, se tiene que

$$\begin{aligned}
\ln \frac{\mathbb{P}(Y_i = 1)}{\mathbb{P}(Y_i = K)} =& \beta_1 \cdot X_i \\
\ln \frac{\mathbb{P}(Y_i = 2)}{\mathbb{P}(Y_i = K)} =& \beta_2 \cdot X_i \\
&\vdots \\
\ln \frac{\mathbb{P}(Y_i = K-1)}{\mathbb{P}(Y_i = K)} =& \beta_{K-1} \cdot X_i
\end{aligned}$$

Si se aplica la función exponencial a ambos lados de las igualdades, y después se despejan las probabilidades, se obtiene que

$$\begin{aligned}
\mathbb{P}(Y_i = 1) =& \mathbb{P}(Y_i = K) e^{\beta_1 \cdot X_i} \\
\mathbb{P}(Y_i = 2) =& \mathbb{P}(Y_i = K) e^{\beta_2 \cdot X_i} \\
&\vdots \\
\mathbb{P}(Y_i = K-1) =& \mathbb{P}(Y_i = K) e^{\beta_{K-1} \cdot X_i}
\end{aligned}$$

de este modo,

$$\mathbb{P}(Y_i = K) = 1 - \sum_{k=1}^{K-1} \mathbb{P}(Y_i = k) = 1 - \sum_{k=1}^{K-1} \mathbb{P}(Y_i = K) e^{\beta_k \cdot X_i}$$

entonces,

$$\mathbb{P}(Y_i = K) = \frac{1}{1 + \sum_{k=1}^{K-1} e^{\beta_k \cdot X_i}}$$

Esto se puede utilizar para hallar otras probabilidades

$$\begin{aligned}
\mathbb{P}(Y_i = 1) =& \frac{e^{\beta_1 \cdot X_i}}{1 + \sum_{k=1}^{K-1} e^{\beta_k \cdot X_i}} \\
\mathbb{P}(Y_i = 2) =& \frac{e^{\beta_2 \cdot X_i}}{1 + \sum_{k=1}^{K-1} e^{\beta_k \cdot X_i}} \\
&\vdots \\
\mathbb{P}(Y_i = K-1) =& \frac{e^{\beta_{K-1} \cdot X_i}}{1 + \sum_{k=1}^{K-1} e^{\beta_k \cdot X_i}}
\end{aligned}$$

Los cálculos del modelo anterior no pueden realizarse a mano, por lo que se tiene que recurrir a la programación. La implementación de la regresión logística multinomial en **R** requiere del uso del paquete *nnet*, y se realiza a través del siguiente código

1. Deben calibrarse los parámetros del modelo a mediante el uso de un método numérico.

```
model<-multinom(Logro ~ .-1,data=cf)
```

2. Después, debe almacenarse la matriz de parámetros estimados.

```
A<-as.matrix(summary(model)$coeff)
A
```

3. El siguiente paso es construir la matriz de observaciones de variables explicativas.

```
cf<-select(cf,JG:DIF)
cf<-mutate(cf,uno=rep(1,nrow(cf)))
cf<-as.matrix(select(cf,c(JG,JE,JP,DIF)))
dim(cf)
```

4. La matriz A tiene por filas a los vectores β_i, y la matriz cf tendría, de forma similar, en su fila i, al vector X_i de variables explicativas asociado a la observación i.

```
A%*%cf[1,]
p.cruz<-sapply(1:nrow(cf),function(x)(exp(A[1,]%*%cf[x,])+1)/(1+sum(exp(A%*%cf[x,]))))
```

3. Complicaciones

- La variable a trabajar resultó ser una variable categórica, por lo que fue necesario adaptar la información utilizando una técnica de análisis multivariado, la cual fue aplicar un modelo de regresión logística multinomial.
- La base de datos tenía poca información, 48 observaciones, esto provocó que fuera más difícil proponer modelos no triviales para la Serie de Tiempo. Para esto, se decidió trabajar con cuatro modelos distintos, se abordarán más a fondo en las secciones siguientes, para poder tomar una decisión más acertada al final.
- La prueba de confiabilidad de Diebold y Mariano arrojaba resultados similares para todos los modelos propuestos, esto será expuesto en las secciones siguientes, por lo que la elección de modelo se tornó en una labor complicada. Para afrontar esto se decidió optar por otro criterio de selección basado en los pronósticos de los modelos.
- Una de las propuestas inciales fue ajustar un modelo bayesiano a los datos, pues con esto se podría pronosticar si el Cruzl Azul llegaría a ser campeón, sin embargo, los cálculos resultaron ser más complicados de los esperado y se decidió descartar esta alternativa. Por lo que se trabajó con un modelo clásico, y se tuvo que cambiar el objetivo del pronóstico, a responder la pregunta de si el Cruz Azul llegaría a ser, al menos, subcampeón, pues, al tener más observaciones de este tipo, la probabilidad de que suceda resultaría no nula.

4. Estimación del modelo

Con base en la metodología Box – Jenkins es posible identificar y estimar un modelo estadístico que pueda ser interpretado como generador de la información muestral, es decir, de la probabilidad de que el cruz azul llegue a ser al menos subcampeón. Para ello se sigue una serie de pasos que sugiere Box – Jenkins para llegar al mejor modelo de ajuste a la muestra.

4.1. Verificación de propiedades a la información muestral

Sea $T = \{1996\_2,\ 1997\_1,\ 1997\_2,\ ,1998\_1,\ \cdots,\ 2018\_2,\ 2019\_1,\ 2019\_2\}$ los cuales representan las temporadas jugadas por el Cruz Azul, además los años que tienen el _1 fueron temporadas jugadas en el periodo enero-mayo y los que tienen _2 entre los meses junio-diciembre. Se define la serie $\{P_t\}_{t\in T}$, como la probabilidad al tiempo $t \in T$ de que el Cruz Azul sea al menos subcampeón en un torneo organizado por la Liga Mx. En la sección anterior se estimaron los valores que conforman a $\{P_t\}_{t\in T}$ y al ser dependientes del tiempo, se puede decir de manera formal que $\{P_t\}_{t\in T}$ es una serie de tiempo. Así, se genera la gráfica que corresponde a dicha serie la cual se presenta a continuación:

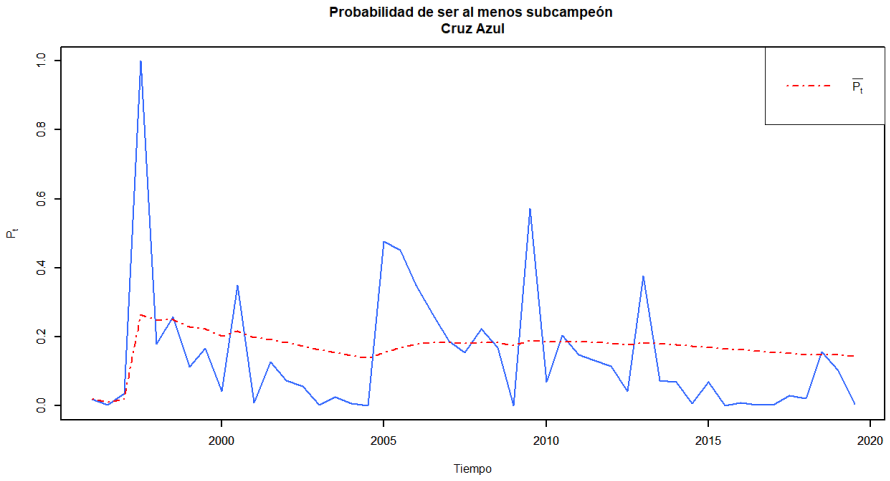

Figura 3: Probabilidad de que el equipo Cruz Azul sea al menos subcampeón

Analizando la serie de tiempo $\{P_t\}_{t\in T}$ en la figura 3, se observan niveles altos de probabilidad cada cierto periodo de tiempo, el cual es aproximadamente cada 8 temporadas. Por otro lado, es posible apreciar las medias móviles con la línea roja punteada en la gráfica, estas medias móviles representan un indicio de que la serie se va estabilizando a lo largo del tiempo.

Para continuar con el análisis, primero se define la función de autocorrelación muestral (ACF), $\hat{\rho}_n$, para un proceso cov-estacionario, tal que $\hat{\rho}_n : \{0, 1, ..., n-1\} \to \mathbb{R}$:

$$\hat{\rho}_x(h) = \frac{\hat{\gamma}_x(h)}{\hat{\gamma}_x(0)} = \frac{\frac{1}{n-h}\sum_{t=1}^{n-h}(x_t - \bar{x}_n)(x_{t-h} - \bar{x}_n)}{\frac{1}{n}\sum_{t=1}^{n-h}(x_t - \bar{x}_n)} \tag{1}$$

la cual define la autocorrelación entre los valores que se encuentran a h intervalos de distancia.

Por otro lado, la función de autocorrelación parcial (PACF) se define como:

$$\rho_{ss} = \frac{\rho_s - \sum_{j=1}^{s-1} \rho_{s-1,j}\ \rho_{s-j}}{1 - \sum_{j=1}^{s-1} \rho_{s-1}\ \rho_j} \tag{2}$$

la cual es una medida de la correlación entre observaciones de una serie de tiempo que se encuentran separadas por s unidades de tiempo (y_t y y_{t-s}), después de ajustarse para la presencia de los demás términos de desfase más corto ($y_{t-1}, y_{t-2}, ..., y_{t-k-1}$).

En el lenguaje de programación R, por medio de las funciones

```
acf(ts,lag.max)
acf(ts,lag.max, type = "partial")
```

se puede apreciar de forma gráfica los valores que toman la ACF (1) y PACF(2) muestrales con rezagos: $\{1, 2, \ \ldots, \ lag.max\}$, para una serie de tiempo ts.

Así, se graficará la función de autocorrelación (ACF) y la función de autocorrelación parcial (PACF) asociadas a la serie $\{P_t\}$, con el propósito de identificar rezagos con los que pueda haber una correlación significativa, así como indicios de patrones.

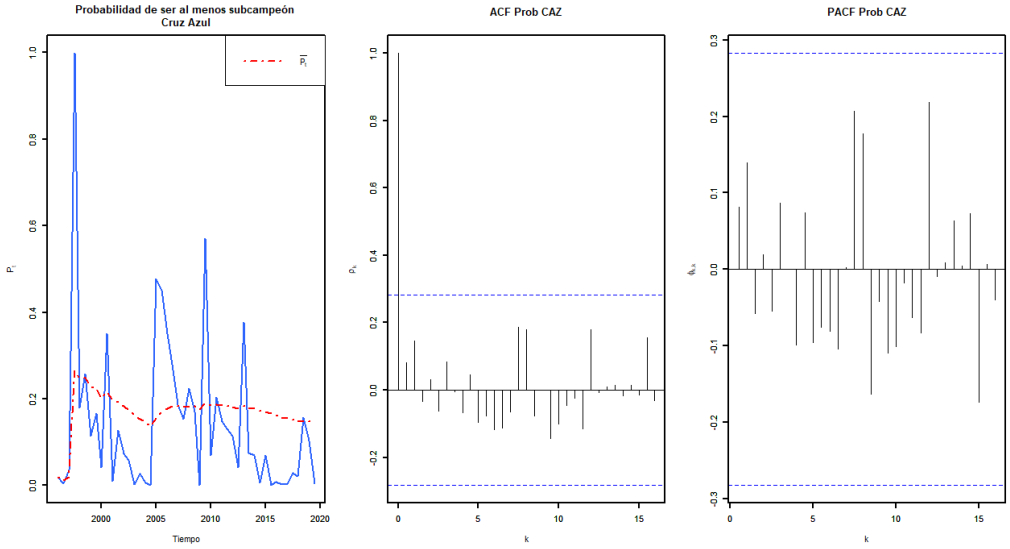

Figura 4: ACF Y PACF de la serie

Se observa que por la forma en la que se presentan los rezagos, se puede inferir que se trata de un ruido blanco, pues no se identifica mucha correlación entre ellos, sin embargo, será de mucha utilidad para implementar modelos adecuados. Las funciones asociadas a la ACF y PACF de la serie se presentan a continuación, junto con la gráfica asociada a las probabilidades de que el Cruz Azul sea al menos subcampeón en la Liga Mx.

De modo que con base en la gráfica anterior, se determina que no es necesario aplicar una transformación a la serie original para reducir la varianza, debido que los valores que toma la serie se

encuentran entre 0 y 1, y de aplicarla, no habría ningún cambio significativo o bien, podría incluso deteriorar los datos causando una distancia mayor entre ellos y consecuentemente se generaría un efecto de heteroscedasticidad.

Adicionalmente, se puede visualizar que la curva asociada a las medias móviles es positiva y además tiene una ligera tendencia decreciente, por lo tanto, se ha tomado la decisión de hacer la transformación de las primeras diferencias, generando una nueva serie $Z_t = \Delta P_t$. A continuación se presentan la ACF y PACF de las primeras diferencias de probabilidades P_t.

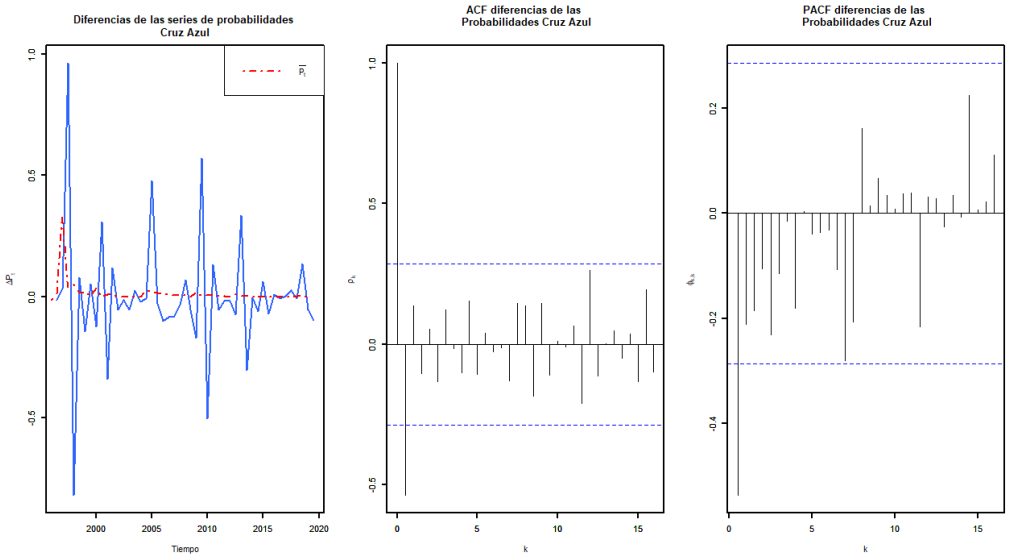

Figura 5: ACF Y PACF de las primeras diferencias de la serie

Haciendo un análisis de la gráfica anterior, se observa que se ha eliminado la tendencia de la serie y además, ésta se encuentra centrada en cero. En adición, después de aplicar primeras diferencias, a priori en la ACF y la PACF se aprecia que existe correlación con el rezago anterior.

Por otro lado, si se tiene una serie de tiempo $\{Z_t\}$, se define el operador de rezagos B, aplicado al elemento Z_t, como $BZ_t = Z_{t-1}$ y si se aplica el operador de forma recursiva, para $r \in \mathbb{Z}$ se deduce que $B^r(Z_t) = Z_{t-r}$.

En general, un modelo de series de tiempo puede representarse por medio de los polinomios característicos sobre el operador rezago definido previamente como B, de la siguiente manera:

- Para un modelo ARIMA$(p,\ d,\ q)$:

$$\phi(B)Z_t = \theta(B)\varepsilon_t \tag{3}$$

  donde $Z_t = \Delta^d(W_t)$, $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ y $\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$

- Para un modelo SARIMA $(P,\ D,\ Q)_s$:

$$\Phi(B^s)Z_t = \delta + \Theta(B^s)\varepsilon_t \tag{4}$$

  donde $Z_t = \Delta_s^D W_t$ con $\Delta_s W_t = W_t - W_{t-s}$, además $\Phi(B) = 1 - \Phi_1 B^s - \cdots - \Phi_P B^{Ps}$ y $\Theta(B) = 1 + \Theta_1 B^s + \cdots + \Theta_Q B^{Qs}$.

Por otra parte para evaluar qué modelo ajusta de mejor manera a los datos, se hace uso de la función *arima*() en R donde, el primer parámetro x es una serie de tiempo univariante y enseguida el parámetro orden (*order*) de esta función especifica la parte no estacional, que incluye los tres elementos (p, d, q) que corresponden al orden del modelo AR, el grado de diferenciación y el orden del modelo MA.

De la misma forma, el parámetro estacional (*seasonal*) determina la parte estacional del modelo ARIMA(p,d,q) más el periodo, este parámetro es una lista que incluye el orden (*order*) y el periodo (*period*) de los elementos, de no especificar este parámetro se considera la frecuencia de la serie de tiempo.

Posteriormente, se proponen dos modelos que ajusten lo mejor posible a la serie de tiempo, los cuales son presentados a continuación y de igual manera, se incluyen las funciones implementados en R que se utilizaron para estimar los parámetros de dichos modelos.

4.2. Identificación de modelos tentativos y estimación de sus parámetros

4.2.1. Modelo 1: ARIMA(0,1,1)

Para el primer modelo, se considera un modelo ARIMA$(0, 1, 1)$ para P_t, que es equivalente a un modelo de medias móviles MA(1) asociado a la serie de las primeras diferencias $Z_t = \Delta P_t$, que de acuerdo a la expresión (3) está dado por $Z_t = \varepsilon_t + \theta_1 \varepsilon_{t-1}$. Para esto es posible dar dos justificaciones, la primera de ellas se deduce de forma analítica y la otra se realiza acorde a fenómenos que se observan en la realidad del equipo.

La primera consiste en analizar las gráficas de la figura 5, en la cual la ACF da la impresión de truncarse en $k = 1$, lo que genera sospecha de que el choque aleatorio que afecta a una observación de P_{t-1}, también afecta al valor de P_t. Además, la PACF toma un valor significativo para el primer rezago y a partir de ahí, presenta un comportamiento decreciente (que son caracterísiticas deseables para un modelo MA(1)). No obstante, en algunos valores para los rezagos k, se aprecian niveles altos en la PACF, lo cual se considerará más adelante.

Para la segunda, primero debemos reescribir el modelo en términos de $P_t = P_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1}$ y de esta manera es fácil suponer que cada temporada existe una mayor correlación con el rezago inmediato anterior para la serie original debido a que la plantilla de jugadores del Cruz Azul suele participar en dos torneos consecutivos.

La siguiente línea de código implementada en R calcula el parámetro asociado a este modelo MA(1)

```
(ARIMA_0_1_1 <- arima(ts, order = c(0,1,1)))
```

con lo anterior se obtiene lo que se muestra a continuación

```
Call:
arima(x = ts, order = c(0, 1, 1))

Coefficients:
        ma1
      -0.9068
s.e.   0.0886
```

De esta manera se tiene la siguiente estimación para el parámetro $\theta_1 = -0.9068$, de modo que el modelo 1 correspondiente a un MA(1) está dado por

$$Z_t = \varepsilon_t + (-0.9068)\varepsilon_{t-1}$$

4.2.2. Modelo 2: ARIMA(1,1,2)

La motivación para el modelo 2 surge de un argumento similar al que se utilizó para el modelo que se presentó previamente, así mismo es necesario incorporar el rezago Z_{t-1} a este modelo con el propósito de describir al rezago Z_t, considerando dos razones sustanciales, la primera de ellas está relacionada con el comportamiento de la PACF, donde se observa que el primer rezago es significativo, de modo que esto sugiere contemplar un elemento autoregresivo de un rezago en el modelo, dado que es el único rezago que supera la banda de confianza impuesta por la PACF.

Con respecto a la segunda razón, si Z_t se vuelve a escribir en términos de P_t y teniendo en cuenta que el modelo considera al elemento Z_{t-1}, en la serie de tiempo, el modelo para la probabilidad estaría considerando los primeros dos rezagos, y es razonable, por consiguiente el argumento es semejante al que se empleó en el modelo 1, puesto que las peculiaridades del equipo de fútbol se conservan un periodo y la interpretación del segundo rezago está relacionada con que la Liga Mx se lleva a cabo en las mismas épocas del año cada dos periodos.

De manera análoga a la justificación del modelo anterior y con base en la PACF, se puede observar que el primer rezago tiene una correlación negativa, la cual es significativa, de esta se ha propuesto un modelo ARIMA$(1, 1, 2)$, que por medio de (3), se puede expresar como:

$$(1 - \phi_1 B)Z_t = (1 + \theta_1 B + \theta_2 B^2)\varepsilon_t$$

o equivalentemente

$$Z_t = \phi_1 Z_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} \tag{5}$$

el cual contempla el elemento autorregresivo Z_{t-1} y dos componentes de medias móviles, no obstante, la estimación del parámetro que acompaña al elemento ε_{t-1} se establece en cero.

Por otra parte, con base en el análisis de la serie Z_t y considerando este modelo, se puede destacar que los primeros dos choques aleatorios de la serie original son absorbidos por el primer rezago de la serie, lo que se ve reflejado en una autocorrelación importante para $h = 1$ en la ACF de la figura 5.

La siguiente línea de código implementada en R calcula el parámetro asociado a este modelo ARIMA(1,1,2)

```
> (ARIMA_1_1_2<-arima(ts,order = c(1,1,2),fixed = c(NA,0,NA)))
```

con lo anterior se obtiene lo que se muestra a continuación

```
Call:
arima(x = ts, order = c(1, 1, 2), fixed = c(NA, 0, NA))

Coefficients:
         ar1  ma1      ma2
      -0.9376    0  -0.8073
s.e.   0.0905    0   0.1672
```

De esta manera se tienen las siguientes estimaciones para los parámetros $\phi_1 = -0.9376$, $\theta_1 = 0$ y $\theta_2 = -0.8073$ de modo que el modelo 1 correspondiente a un ARIMA(1,1,2) está dado por

$$Z_t = -0.9376 Z_{t-1} + \varepsilon_t - 0.8073 \varepsilon_{t-2} \tag{6}$$

4.2.3. Modelo 3: SARIMA$(2,1,0)_5$

Considérese ahora la serie $Y_t = P_t - P_{t-5}$, pues se han notado indicios de estacionalidad cada 5 rezagos, es importante mencionar que se hizo una búsqueda con el propósito de darle respuesta a este comportamiento, sin embargo, no se encontró un motivo en particular que explicara el fenómeno.

No obstante, fue posible determinar algunos *patrones* que se presentan en la base de datos original, los cuales podrían representar una casualidad, sin embargo, se han considerado para poder generar un buen modelo que se ajuste a los datos. Por ejemplo, para el primer posible patrón se cree que cada vez que el Cruz Azul terminó como subcampeón, al contar 15 temporadas atrás, se observa que para ese rezago, el Cruz Azul no logró clasificar, también se podría considerar como causa que cada vez que se obtuvo el campeonato o subcampeonato, 5 y 10 temporadas adelante apenas llegó a cuartos de final en la mayoría de los casos. Por lo que se considera la serie Y_t previamente definida.

Además existen razones para creer que el valor de esta variable al tiempo t está correlacionado con los rezagos 5, 10 y 15, por lo que se propone un modelo SARIMA $(2,1,0)_5$, que de acuerdo a la expresión de (4) está dado por

$$(1 - \Phi_1 B^5 - \Phi_2 B^{10}) Y_t = \varepsilon_t$$

o equivalentemente:

$$Y_t = \Phi_1 Y_{t-5} + \Phi_2 Y_{t-10} + \varepsilon_t \tag{7}$$

Ahora para analizar más a fondo la serie Y_t, se graficará su ACF y PACF.

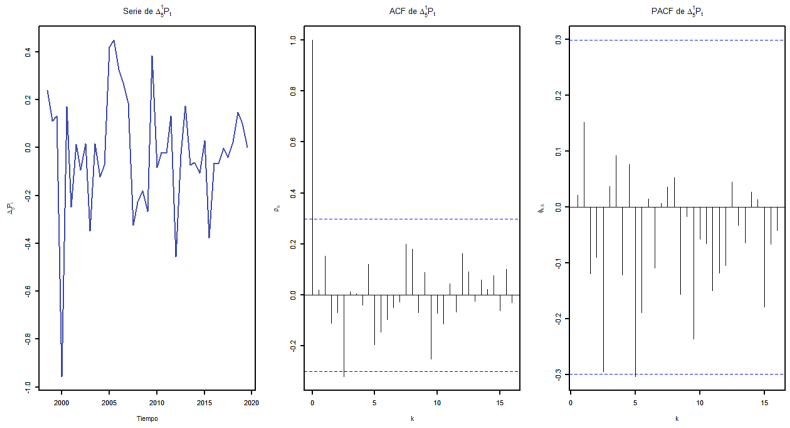

Figura 6: ACF Y PACF del modelo SARIMA$(2,1,0)_5$

Se puede apreciar en la figura 6 la gráfica de la serie Y_t, en la cual se nota que se elima la tendencia y la serie se centra en cero, lo cual es deseable para manejar una serie de tiempo. Por otro lado, en la PACF se observa que Y_t tiene correlación significativa con los rezagos y_{t-5} y y_{t-10}. Siguiendo con el análisis y justificación del modelo, se despejará la expresión $Y_t = P_t - P_{t-5}$ en el modelo

propuesto de (7), de tal forma que se llegue a una expresión para P_t, como sigue:

$$P_t - P_{t-5} = \Phi_1(P_{t-5} - P_{t-10}) + \Phi_2(P_{t-10} - P_{t-15}) + \varepsilon_t$$
$$P_t = (\Phi_1 + 1)P_{t-5} + (-\Phi_1 + \Phi_2)P_{t-10} + (-\Phi_2)P_{t-15}$$

Así, se corroboran las observaciones e identificación de patrones que se notaron en la base de datos original, ya que para modelar la probabilidad de que el cruz azul sea al menos subcampeón se usan las probabilidades de 5, 10 y 15 periodos atrás.

Posteriormente se utiliza la función *arima()* en R para estimar los parametros Φ_1 y Φ_2

```
> (SARIMA_2_1_0_5 <- arima(ts, order = c(0,0,0), seasonal =
                      list(order = c(2,1,0),  period = 5)))
```

con lo anterior se obtiene lo siguiente

```
Call:
arima(x = ts, order = c(0, 0, 0), seasonal = list(order = c(2, 1, 0), period = 5))

Coefficients:
        sar1     sar2
      -0.6969  -0.5526
s.e.   0.1591   0.1400
```

De esta manera se tiene la siguiente estimación para los parámetros $\Phi_1 = -0.6969$, $\Phi_2 = -0.5526$ de modo que el modelo 3 correspondiente a un SARIMA $(2,1,0)_5$ está dado por

$$Y_t = -0.6969Y_{t-5} - 0.5526Y_{t-10} + \varepsilon_t$$

4.2.4. Modelo 4: SARIMA(1,1,0)$_8$

Para esta última propuesta se considera la serie dada por $W_t = P_t - P_{t-8}$, dado que se observan indicios de estacionalidad cada 8 rezagos. Es importante mencionar que se hizo una búsqueda con el propósito de darle respuesta a este comportamiento, sin embargo, no se encontró un motivo en particular que explicara el fenómeno.

No obstante ha sido posible identificar una serie de *patrones* en la base de datos original, los cuales podrían representar únicamente una casualidad, sin embargo, se han considerado para poder generar un buen modelo que se ajuste a los datos. El primero de éstos consiste en que cada 8 rezagos, se cree que el primero de estos patrones consiste en que 8 periodos después de que el Cruz Azul alcanzó el logro de Subcampeón, en la mayoría de los casos no clasificó o llegó a cuartos de final.

Adicionalmente se podría considerar otro posible patrón, que acorde a la figura (3) asociada a la probabilidad de que el Cruz Azul sea al menos subcampeón, se puede observar un posible patrón que consiste en que cada 6 temporadas en la serie de tiempo hubo un aumento en la probabilidad para esos periodos, por otro lado considerando que estos periodos son equidistantes y la base de datos original está compuesta por 48 registros, entonces es posible determinar que cada 8 rezagos se presentó un evento significativo.

Dado lo anterior, se propone un modelo SARIMA$(1,1,0)_8$, que de acuerdo a la expresión de (4) está dado por

$$(1 - \Phi_1 B^8)W_t = \varepsilon_t$$

o equivalentemente:

$$W_t = \Phi_1 W_{t-8} + \varepsilon_t \tag{8}$$

Ahora se analizarán las gráficas de la ACF y la PACF correspondientes a la serie de tiempo $W_t = \Delta_8(P_t)$

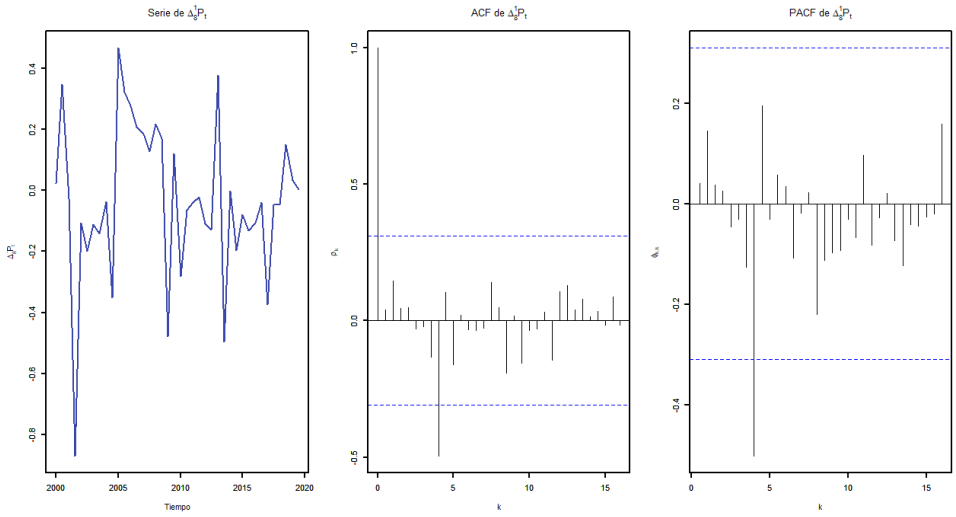

Figura 7: ACF Y PACF del modelo SARIMA$(1,1,0)_8$

En la PACF de este modelo se observa que el octavo rezago supera la banda de confianza impuesta por esta función, es decir, hay rastros de autocorrelación negativa con el elemento W_{t-8}, esto implica para la serie original, que indica la probabilidad de que el cruz azul sea al menos subcampeón, dependerá de las observaciones en los rezagos 8 y 16, como se muestra en la siguiente expresión, que se obtiene despenjando $W_t = \Delta_8(P_t)$ en (8)

$$P_t - P_{t-8} = \Phi_1(P_{t-8} - P_{t-16}) + \varepsilon_t$$
$$P_t = (1 + \Phi_1)P_{t-8} - \Phi_1 P_{t-16} + \varepsilon_t$$

Posteriormente se utiliza la función *arima()* en R para estimar el parámetro Φ_1, como se muestra:

```
(SARIMA_1_1_0_8 <- arima(ts, order = c(0,0,0), seasonal = list(order = c(2,1,0),
                  period = 8)))
```

con lo anterior, se obtiene lo siguiente

```
Call:
arima(x = ts, order = c(0, 0, 0), seasonal = list(order = c(1, 1, 0), period = 8))

Coefficients:
        sar1
      -0.6275
s.e.   0.1253
```

De esta manera, la estimación para el parámetro $\Phi_1 = -0.6275$ de tal forma que permite reexpresar al modelo 4, correspondiente a un SARIMA $(1,1,0)_8$, como:

$$W_t = -0.6275 W_{t-8} + \varepsilon_t$$

4.3. Diagnóstico de los modelos

4.3.1. Prueba de significancia Z

Se introducirá la función *coeftest*() en **R**, la cual resulta de mucha utilidad para saber qué tan significativos son los parámetros y si se ajustan de buena manera a los modelos, los cuales, se han estimado previamente con la función *arima*(). Es importante mencionar que se trata de una función genérica que realiza una prueba Z y una cuasi prueba de Wald, las cuales consisten en lo siguiente; la prueba Z es una prueba estadística de hipótesis basada en el estadístico Z, que sigue la distribución normal estándar bajo la hipótesis nula, esta hipótesis nula consiste en que los parámetros no son significativos. Esta prueba además es de utilidad para saber si el modelo está sobreespecificado o no.

Por lo que respecta a la prueba de Wald, es una prueba estadística parámetrica, la cual es utilizada para poner a prueba el verdadero valor del parámetro basándose en la estimación de la muestra. Del mismo modo, la función *coeftest*() devuelve una matriz de coeficientes que en sus columnas contiene las estimaciones de los parámetros, los errores estándar asociados a éstos, las estadísticas de prueba y los p-values.

Adicionalmente para la parte de los p-values, se incluyen asteriscos, los que informan la frontera del valor de significancia para la cual es mayor que el p-value, es decir, entre más asteriscos tenga, el p-value es más pequeño y hay mayor confianza para no aceptar la hipótesis nula. Se explicará más adelante de manera detallada. A continuación se mostrarán los resultados de esta prueba para cada uno de los modelos.

Nota: A partir de este punto, cada vez que se trabaje con p-values, se dirá que estos son pequeños, o cualquier expresión similar a está, para hacer referencia a la comparación entre estos y un nivel de significancia que estableceremos en el 5 %, por otra parte se dira que son suficientemente grandes para aceptar, si las medianas a lo largo de una serie de rezagos, rebasan al 70 %, está última aclaración solo será tomada en cuenta cuando sea realmente importante el tomar una decisión para continuar con el proyecto. En caso contrario solo serán evidencia necesaría para no rechazar.

Modelo 1

```
> coeftest(ARIMA_0_1_1)
z test of coefficients:
     Estimate Std. Error z value  Pr(>|z|)
ma1 -0.906779   0.088638  -10.23 < 2.2e-16 ***
```

Modelo 2

```
> coeftest(ARIMA_1_1_2)
z test of coefficients:
     Estimate Std. Error  z value  Pr(>|z|)
ar1 -0.937623   0.090537 -10.3562 < 2.2e-16 ***
ma2 -0.807331   0.167172  -4.8293  1.37e-06 ***
```

Modelo 3

```
> coeftest(SARIMA_2_1_0_5)
z test of coefficients:2
     Estimate Std. Error z value  Pr(>|z|)
```

```
sar1 -0.69691    0.15911 -4.3800 1.187e-05 ***
sar2 -0.55263    0.13997 -3.9484 7.869e-05 ***
```

Modelo 4

```
> coeftest(SARIMA_1_1_0_8)
z test of coefficients:
     Estimate Std. Error z value  Pr(>|z|)
sar1 -0.62748    0.12534 -5.0063 5.548e-07 ***
```

El p-value se puede interpretar como la probabilidad de equivocarse, dado que se rechaza la hipótesis nula, la cual es $\mathscr{H}_0$: *Los parámetros no son significativos*. Por lo tanto, como se observa que los p-value asociados a cada parámetro, en cada uno de lo modelos, son muy a cercanos a cero, se puede decir casi con total confianza que se rechaza la hipótesis nula y por lo tanto, los parámetros sí son significativos.

4.3.2. Criterios de Información

Para una serie de tiempo con n observaciones, el criterio de información de Akaike (AIC) y el criterio de información bayesiano (BIC) son medidas de calidad para comparar entre dos o más modelos. Estos están definido como:

$$AIC = n\ln\left(\hat{\sigma}_n^2\right) + 2(1+p+q)$$

$$BIC = n\ln\left(\hat{\sigma}_n^2\right) + (1+p+q)\ln(n)$$

Dado un conjunto de modelos candidatos para los datos, el modelo preferido es el que tiene el valor mínimo en el AIC o BIC. Por lo tanto, estos criterios no solamente recompensan la bondad de ajuste, sino también incluyen una penalización en función del número de parámetros estimados.

Nota: Estudios empíricos muestran que AIC es más eficiente con muestras pequeñas que BIC y, como es el caso, para las observaciones que se tienen de $\{P_t\}$, se dará más importancia al primer criterio.

Por medio de las funciones $AIC()$ y $BIC()$ en R se pueden obtener dichos valores para los cuatro modelos que se propusieron, a continuación se muestran las tablas de resultados.
Para Akaike:

```
> (akaike<-data.frame("ARIMA_0_1_1"=AIC(ARIMA_0_1_1),
                      "ARIMA_1_1_2"=AIC(ARIMA_1_1_2),
                      "SARIMA_2_1_0_5" = AIC(SARIMA_2_1_0_5),
                      "SARIMA_1_1_0_8"=AIC(SARIMA_1_1_0_8)))

      ARIMA_0_1_1   ARIMA_1_1_2   SARIMA_2_1_0_5   SARIMA_1_1_0_8
1     -16.69292     -13.06792     -7.815497        -6.506905
```

Para BIC:

```
> (bayes<-data.frame("ARIMA_0_1_1"=BIC(ARIMA_0_1_1),
                     "ARIMA_1_1_2"=BIC(ARIMA_1_1_2),
                     "SARIMA_2_1_0_5" = BIC(SARIMA_2_1_0_5),
                     "SARIMA_1_1_0_8"=BIC(SARIMA_1_1_0_8)))
```

	ARIMA_0_1_1	ARIMA_1_1_2	SARIMA_2_1_0_5	SARIMA_1_1_0_8
1	-12.99262	-5.667333	-2.531897	-3.129146

A partir de estos criterios, se puede concluir que el modelo que mejor ajusta es el primero, ARIMA$(0, 1, 1)$, aunque, a decir verdad, los cuatro modelos cuentan con buenos niveles con respecto a otros que se propusieron en el proceso y que en su mayoría tenían un valor de AIC positivo.

4.3.3. Validación de supuestos

En esta sección se realizarán diversas pruebas con el fin de validar los supuestos de estacionariedad para los cuatro modelos propuestos. Para esto, se introducirán algunas definiciones.

Si se tiene un modelo de series de tiempo para $\{P_t\}$, por cada observación P_t se puede obtener una estimación $\hat{P}_t$ que depende de las observaciones anteriores y su valor $\hat{P}_t = \mathbb{E}_{I_{t-1}}(P_t)$ que resulta ser la esperanza de P_t, condicionada al evento I_{t-1}, que contiene toda la información que ya ha sucedido hasta el tiempo $t-1$. Así, se definen a los *residuales* como

$$e_t = P_t - \hat{P}_t$$

Esto será util para evaluar si el modelo captura adecuadamente la información. Ya que, es deseable que los residuales e_t no tengan autocorrelación, pues de tenerla, significaría que en éstos, hay información que debería ser usada en el cálculo de la estimación $\hat{P}_t$; por otra parte, deben de cumplir que tengan media cero, así, el pronóstico que arroje el modelo será insesgado. Notemos que éstas son propiedades de un ruido blanco. Para evaluar si se cumplen estas condiciones en todos los modelos se analiza la figura 8 que muestra a los residuales de cada modelo como una serie de tiempo, se aprecia que están centradas en 0.

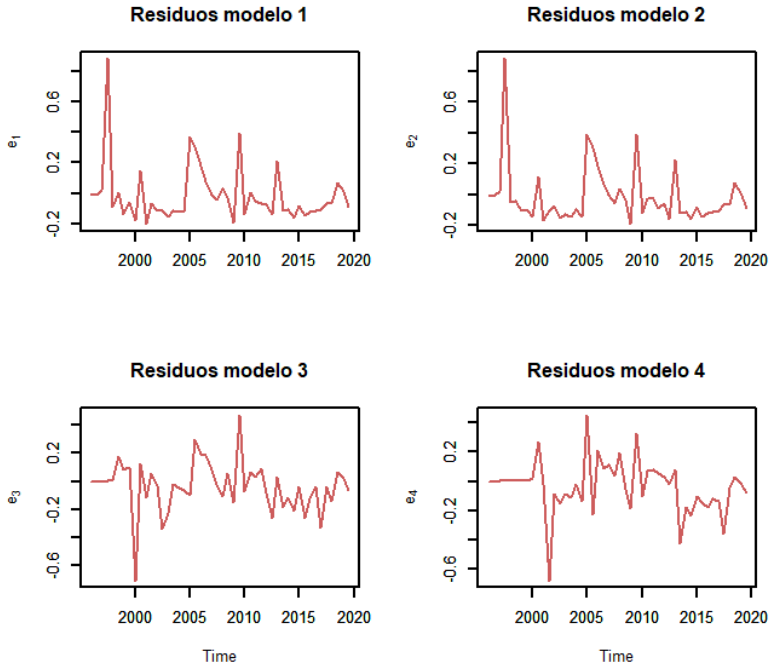

Figura 8: Series de los residuos asociados a los modelos 1, 2, 3 y 4

En la siguiente figura se muestran los histogramas de los residuales para cada modelo, se observa

que en todos los casos existe una gran acumulación en el intervalo más cercano al cero, lo que se puede interpretar como que la media de los residuales toma valores cercanos a cero.

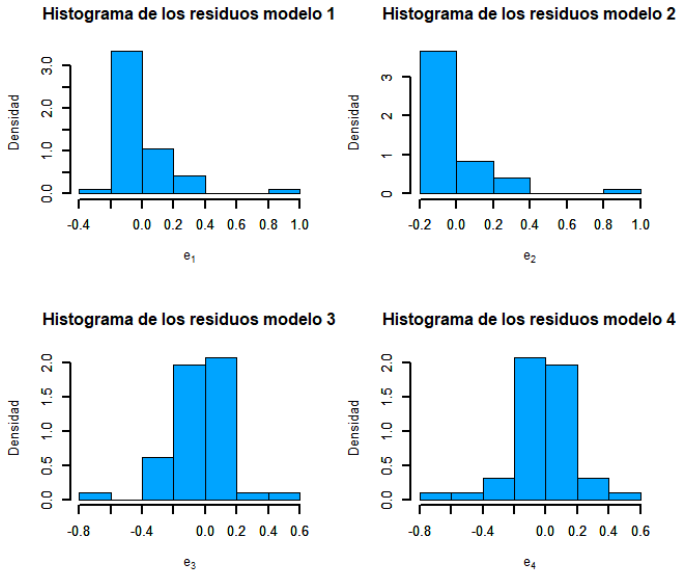

Figura 9: Histogramas de los residuos asociados a los modelos 1, 2, 3 y 4

Con la intención de visualizar que no hay correlación entre los residuales se analizan las ACF y PACF correspondientes a los residuos que se generan por cada modelo. En todos los casos vemos un comportamiento típico de un ruido blanco, es decir, la función de autocorrelación evaluada en 0, vale 1 y para cualquier otro rezago, se mantiene dentro de las bandas punteadas, por lo que se puede decir que toman el valor de 0 con cierto nivel de confianza. De la PACF se deduce que no existe correlación significativa entre los elementos e_t y e_{t-k} con $k = 1, 2, \ldots, 15$

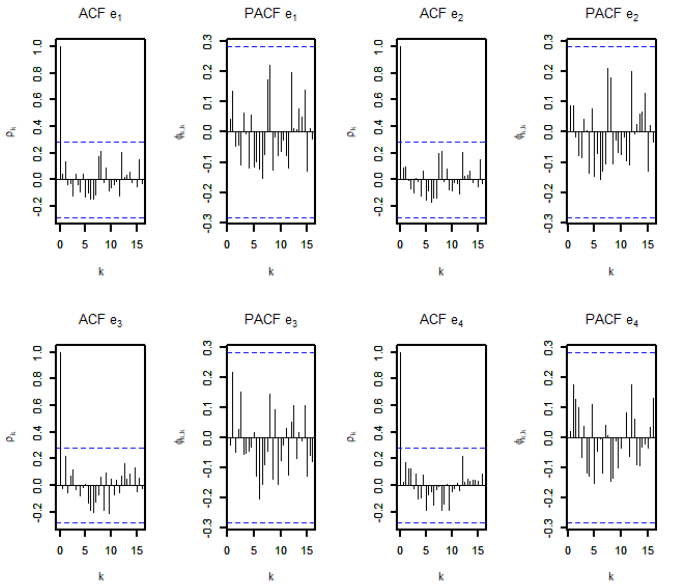

Figura 10: ACF y PACF de los residuos asociados a los modelos 1, 2, 3 y 4

4.3.4. Prueba de Dickey Fuller

Sea un modelo autorregresivo (1) expresado como $y_t = \alpha y_{t-1} + \varepsilon_t$. Al sustraer y_{t-1} de ambos lados el resultado es:

$$\Delta y_t = (\alpha - 1) y_{t-1} + \varepsilon_t \tag{9}$$

La ecuación (9) es la base de la prueba $Dickey - Fuller$. El estadístico de la prueba es el estadístico t sobre la variable dependiente rezagada. Así, si $\alpha > 1$, el coeficiente de la variable dependiente rezagada será positivo; y si $\alpha = 1$, $(\alpha - 1)$ será igual a 0, aún así, para ambos casos, y_t es no estacionaria.

Por otro lado, la hipótesis nula en esta prueba es, $\mathscr{H}_0 = \alpha$ es igual a 1 , o bien que existe una raíz unitaria y por lo tanto la serie no es estacionaria.

Por medio de la función en R $adf.test(ts, k)$, se realiza la prueba de Dickey- Fuller para los valores $k = \{1, 2, \ldots, 18\}$ y se obtienen los siguientes resultados:

- Modelo 1:ARIMA(0, 1, 1)

```
> adf.test(res1)
> adf_pv_ts<-sapply(1:18,function(i)adf.test(res1,k=i)$p.value)
> summary(adf_pv_ts)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.01093 0.21360 0.45343 0.35719 0.48860 0.74434
```

- Modelo 2:ARIMA(1,1,2)

```
> adf.test(res2)
> adf_pv_ts<-sapply(1:18,function(i)adf.test(res2,k=i)$p.value)
> summary(adf_pv_ts)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0100  0.1755  0.4248  0.3532  0.4947  0.7301
```

- Modelo 3: SARIMA$(2, 1, 0)_5$

```
> adf.test(res3)
> adf_pv_ts<-sapply(1:18,function(i)adf.test(res3,k=i)$p.value)
> summary(adf_pv_ts)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0263  0.3715  0.5091  0.4436  0.5787  0.7047
```

- Modelo 4: SARIMA$(1, 1, 0)_8$

```
> adf.test(res4)
> adf_pv_ts<-sapply(1:18,function(i)adf.test(res4,k=i)$p.value)
> summary(adf_pv_ts)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.01876 0.42615 0.49987 0.48004 0.60632 0.69795
```

En los párafos anteriores se definió la prueba *Dickey - Fuller* para un modelo AR(1), esta prueba se puede generalizar, para modelos con más rezagos, aunque la hipótesis nula se mantiene, siendo $\mathscr{H}_0$*: La serie no es estacionaria.* Así, dados los resultados que se obtuvieron para cada uno de los cuatro modelos, se observa que la mediana de los p-values que considera desde el orden del primer hasta el décimo octavo rezago, toma valores cercanos a 0.5, esto no aporta información útil para rechazar o aceptar la hipótesis nula, por lo que a continuación se realizará una prueba que resulta ser más confiable cuando se trata de aceptar o rechazar que una serie sea ruido blanco.

4.3.5. Prueba de Ljung Box

Ahora, se considera la prueba de Ljung Box. Sea $X_t = \mu + \varepsilon_t$ donde $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ es ruido blanco estricto con varianza unitaria, entonces

$$\sqrt{n}\hat{\rho}_n(h) \xrightarrow{d} N(0,1) \quad \forall 1 \leq h \leq n-1$$

Más aún

$$Q(H) = \sum_{h=1}^{H} (\sqrt{n}\hat{\rho}_n(h))^2 \xrightarrow{d} X^2_{(H)}$$

Las hipótesis que asume esta prueba están dadas por $\mathscr{H}_0 : \varepsilon_t$ se distribuye como ruido blanco.

Nota: Q(H) es conocida como el estadístico de Box-Pierce. No obstante, Ljung y Box modificaron el estadístico Q(H) con el propósito de mejorar la aproximación de Q(H) a la $X^2_{(H)}$. Por lo tanto, se define el estadístico de Ljung-Box como

$$Q^*(H) = n(n+2)\sum_{h=1}^{H} \frac{1}{n-h}\hat{\rho^2_n}(h) \xrightarrow{d} X^2_{(H)}$$

De esta manera se tiene que $\mathscr{H}_0 : \mu + \varepsilon_t$ con $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ ruido blanco estricto. Y además si $Q^* > X^2_{(H)}(\alpha)$ entonces se rechaza $\mathscr{H}_0$.

El siguiente punto consiste en la implementación en R, la cual se da por medio de la función: $Box.test(ts, type = "Ljung - Box", lag)$ para $lag = 1, 2, \ldots, 18$ y para cada uno de los rezagos asociados a los modelos propuestos, se tienen de resultados que se presentan a continuación:

- Modelo 1: ARIMA(0, 1, 1)

```
> Box.test(res1,type = "Ljung-Box")
> lb_pv_ts<-sapply(1:18,function(i)as.numeric(Box.test(res1,
                    type = "Ljung-Box",lag = i)$p.value))
> summary(lb_pv_ts)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.5808  0.7656  0.8790  0.8331  0.9434  0.9708
```

- Modelo 2: ARIMA(1, 1, 2)

```
> Box.test(res2,type = "Ljung-Box")
>lb_pv_ts<-sapply(1:18,function(i)as.numeric(Box.test(res2,
                  type = "Ljung-Box", lag = i)$p.value))
```

```
>summary(lb_pv_ts)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.4978  0.6576  0.8700  0.8029  0.9434  0.9761
```

- Modelo 3: SARIMA$(2,1,0)_5$

```
> Box.test(res3,type = "Ljung-Box")
> lb_pv_ts<-sapply(1:18,function(i)as.numeric(Box.test(res3,
                    type = "Ljung-Box",lag = i)$p.value))
> summary(lb_pv_ts)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.2954  0.6075  0.7003  0.7048  0.8398  0.9416
```

- Modelo 4: SARIMA$(1,1,0)_8$

```
> Box.test(res4,type = "Ljung-Box")
> lb_pv_ts<-sapply(1:18,function(i)as.numeric(Box.test(res4,
                    type = "Ljung-Box",lag = i)$p.value))
> summary(lb_pv_ts)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.4469  0.6849  0.7378  0.7091  0.7848  0.8877
```

En la prueba de Ljung Box, la hipótesis nula es $\mathscr{H}_0$: *La serie es ruido blanco*. Por lo tanto, el obtener valores altos en las medianas de los p-value, indica que se podría aceptar la hipótesis. Dado que para todos los modelos, resulta ser el caso de que la mediana de los p-value es mayor que 0.7 se puede aceptar $\mathscr{H}_0$ y decir que todos los procesos residuales asociados a un modelo, son ruido blanco y por lo tanto se cumplen los supuestos de estacionariedad y se confirma que ningún modelo está subespecificado.

Posteriormente, los residuales de cada modelo fueron elevados al cuadrado, ya que también serán analisados por medio de la prueba de Ljung Box y la prueba de efectos ARCH de Engle. Dichos residuales cuadráticos tienen las siguientes gráficas

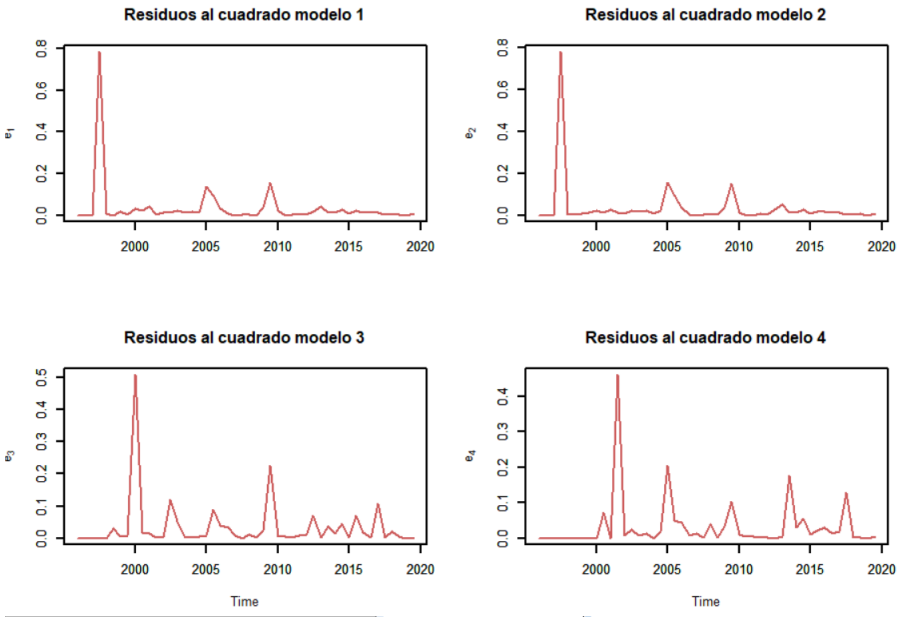

Figura 11: Gráfica de los residuales cuadrados.

4.3.6. Prueba Ljung Box para el cuadrado de los residuales.

Bajo la misma lógica y recursos de programación utilizados en la sección anterior, se aplicó la prueba de Ljung Box pero esta ocasión al cuadrado de los residuales.

En cada uno de los modelos se consideró inicialmente el caso de un sólo resago, y posteriormente los casos de uno a 18 resagos, obteniendo únicamente el 'resumen' de los p-value's arrojados por la función.

- Modelo 1: ARIMA(0, 1, 1)

```
> Box.test(res12,type = "Ljung-Box")
> lb_pv_ts<-sapply(1:18,function(i)as.numeric(Box.test(res12,type =
            "Ljung-Box",lag = i)$p.value))
> summary(lb_pv_ts)
    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.6307  0.8439  0.9405  0.9054  0.9928  0.9995
```

- Modelo 2: ARIMA(1, 1, 2)

```
> Box.test(res22,type = "Ljung-Box")
    > lb_pv_ts<-sapply(1:18,function(i)as.numeric(Box.test(res22,type =
            "Ljung-Box",lag = i)$p.value))
    > summary(lb_pv_ts)
        Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     0.6735  0.8653  0.9462  0.9082  0.9901  0.9981
```

- Modelo 3: SARIMA$(2, 1, 0)_5$

```
> Box.test(res32,type = "Ljung-Box")
> lb_pv_ts<-sapply(1:18,function(i)as.numeric(Box.test(res32,type =
            "Ljung-Box",lag = i)$p.value))
> summary(lb_pv_ts)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.4633  0.5836  0.6505  0.6598  0.7390  0.8340
```

- Modelo 4: SARIMA$(1,1,0)_8$

```
> Box.test(res42,type = "Ljung-Box")
> lb_pv_ts<-sapply(1:18,function(i)as.numeric(Box.test(res42,type =
            "Ljung-Box",lag = i)$p.value))
> summary(lb_pv_ts)
    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1986  0.3836  0.5270  0.5539  0.7437  0.8891
```

Nótece que al igual que los residuales, se obtuvieron valores altos en las medianas de los p-value, por lo que se puede no rechazar la hipótesis de que ésta serie sea ruido blanco.

4.3.7. Prueba ARCH de Engle

Para determinar si existen o no efectos del modelo ARCH en los residuales al cuadrado, de los distintos modelos propuestos, se hace uso de la prueba ARCH de Engle. La cual se explicará en las siguientes líneas.

Considere una serie de tiempo

$$y_t = \mu_t + \varepsilon_t \quad t \in Z$$

Donde μ_t es la esperanza condicional del proceso, y ε_t el choque aleatorio asociado. Además, suponga que dichos choques pueden generarse como:

$$\varepsilon_t = \sigma_t s_t$$

Donde, $\{s_t\}_{t\in Z}$ es un proceso de variables aleatorias independientes e idénticamente distribuidas, con media 0 y varianza 1. De este modo:

$$\mathbb{E}\left(\varepsilon_t \varepsilon_{t+h}\right) = 0 \quad \forall\ h \neq 0$$

Sea I_t, el conjunto de información disponible al tiempo t, y considerando los supuestos anteriores, observe que la varianza condicional de y_t estará dada por:

$$\mathbb{V}(y_t|I_t) = \mathbb{V}(\varepsilon_t|I_{t-1}) = \mathbb{E}(\varepsilon_t^2|I_{t-1}) = \sigma_t^2$$

De este modo, la heterocedasticidad condicional en la varianza del proceso, es equivalente a la existencia de autocorrelación entre los choques aletorios de la serie.

Dada esta breve explicación, consideremos que se toman ahora los residuales de la serie e_t, y nótese que si todas las autocorrelaciones en la serie original, y_t, se hubiesen tomado en cuenta en el planteamiento del modelo, esto se vería reflejado en el pronóstico del mismo, i.e., en su media condicional, lo cual implicaría que los residuales fuesen incorrelacionados y con media cero. Sin embargo esto puede no ocurrir, para ello, existe la prueba ARCH de Engle, dicha prueba estará dada por la regresión:

$$e_t^2 = \alpha_0 + \alpha_1 e_{t-1}^2 + \ldots + \alpha_1 e_{t-m}^2 + u_t$$

Donde u_t es un ruido blanco, además la hipótesis nula para está prueba es

$$\mathscr{H}_0 : \alpha_0 = \alpha_1 = \ldots = \alpha_m = 0$$

Dicha hipótesis, establece la no existencia de heterosedasticidad condicional para un periodo de m residuos. Por otro lado la hipótesis alternativa es

$$\mathscr{H}_1 : \text{existe } i \in \{1\ldots, m\} \text{ tal que } \alpha_i \neq 0$$

La implementación de esta prueba, se realizó en **R**, mediante el uso de la función *ArchTest* la cual, en este caso, recibe dos parámetros, un vector de residuales (al cuadrado) y el número de rezagos a considerar durante la prueba. Una vez realizado el contraste de hipótesis la función devuelve como resultado, una lista de objetos que incluye el p-value de la prueba.

Estos test se llevaron a cabo con el objetivo de descartar la posibilidad de que los residuales obtenidos por los modelos propuestos tengan efectos del modelo ARCH, lo que significa que no tendría sentido ajustar dicho modelo a los residuales cuadrados, además de que una de las características deseables de los modelos propuestos es que tales residuales sean homosedásticos, es decir, que tengan varianza constante.

Primeramente se realizó el test a los residuales cuadrados de todos lo modelos simultaneamente utilizando diez resagos. Filtrando los resultados de tal forma que sólo se muestren los p-value's de cada test se obtiene los riguientes resultados:

```
> Aptest<-apply(res2,2,function(x)unname(Arch Test(x,lags = 10)$p.value))
> Aptest
    res12     res22     res32     res42
0.9999978 0.9999984 0.9970371 0.9930856
```

Con lo anterior se puede concluir que para todos los casos se obtuvo un p-value mucho mayor a 0.05, por lo que no se rechaza la hipóteis nula, es decir, no se rechaza que los residuales cuadrados tengan efectos del modelo ARCH.

Posteriormente, se volvió a efectuar el test, pero en este caso se trataron a los residuales de cada modelo individualmente, esto con el fin se observar los resultados del test pero en esta ocasión considerando una cantidad distinta de rezagos, en particular, se consideraron los casos de uno hasta 18 rezagos, de esta forma se mostrarán 18 p-value's por lo que también usó la función summary sobre el resultado de tal forma que sea más fácil interpretarlos. Se obtuvieron los siguientes resultados:

- Modelo 1: ARIMA(0, 1, 1)

```
> a1test<-sapply(1:18,function(i)as.numeric(ArchTest(res12,lags = i)$p.value))
> summary(a1test)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.03372 0.31633 0.99401 0.72513 0.99995 1.00000
```

- Modelo 2: ARIMA(1, 1, 2)

```
> a2test<-sapply(1:18,function(i)as.numeric(ArchTest(res22,lags = i)$p.value))
> summary(a2test)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.02021 0.87014 0.99443 0.77095 0.99997 1.00000
```

- Modelo 3: SARIMA$(2,1,0)_5$

```
> a3test<-sapply(1:18,function(i)as.numeric(ArchTest(res32,lags = i)$p.value))
> summary(a3test)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.001107 0.013904 0.903093 0.623184 0.979426 0.997037
```

- Modelo 4: SARIMA$(1,1,0)_8$

```
> a4test<-sapply(1:18,function(i)as.numeric(ArchTest(res42,lags = i)$p.value))
> summary(a4test)
    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.01637 0.14585 0.96320 0.67026 0.99192 0.99506
```

Dado que todas las pruebas arrojaron una mediana superior a 0.9, entonces se puede concluir que en general no se rechaza $\mathscr{H}_0$ por lo que se considera que no existen efectos del modelo ARCH en los residuales al cuadrado de los modelos propuestos.

Lo anterior tiene sentido pues graficando las ACF y las PACF de los residuales cuadrado se puede apreciar que prácticamente ningún rezago supera las bandas de significancia.

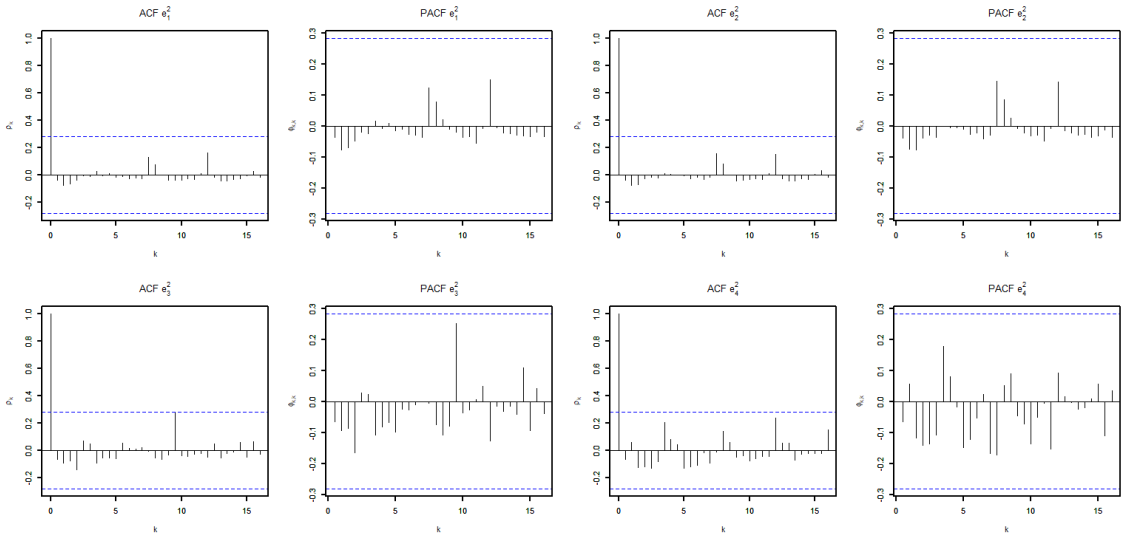

Figura 12: ACF y PACF de los residuales cuadrados.

4.4. Pruebas para los pronósticos

4.4.1. Prueba de Diebold y Mariano

Con el fin de saber, cual modelo tenía la mayor precisión al momento de generar predicciones, se decidio implementar esta prueba, ya que como se comento con anterioridad descartar alguno de estos modelos basados, únicamente, por su nivel AIC no parecía ser buena idea. La parte teórica de la prueba implementada se expone a continuación.

Primeramente, se tomo la decisión utilizar la función de pérdida.

$$g(x) = |x|$$

sobre los residuales obtenidos en cada modelo, esta elección fue tomada, debido a que los residuales siempre resultan ser números en el intervalo $(-1, 1)$, y por ende, si se aplicará la función de pérdida conocida como "el error cuadrático medio", disminuiria de manera no proporcional la magnitud de ambos residuales. por tanto, resulta ser una mejor opción utilizar el "error absoluto medio". De está modo, la prueba que se realizo a cada par de modelos, fue la siguiente Sean $\{e_t^i\}_{t\in T}, \{e_t^j\}_{t\in T}$, los residuales de los modelos i, j, para $i, j \in 1, \ldots, 4$ e $i \neq j$, entonces, se define el estadistico d_t, para $t \in T$ de la siguiente manera

$$d_t = \left|e_t^j\right| - \left|e_t^i\right|$$

De lo cual se tiene que el estadístico para la prueba será

$$DM = \frac{\bar{d}}{\sqrt{\mathbb{V}(\bar{d})}}$$

En donde

$$\bar{d} = \frac{1}{k}\sum_{t\in T_k} d_t$$

Y T_k denotará, al subconjunto de los primeros k elementos de T, considerando el orden establecido al principio de este trabajo. Por otro lado, la hipótesis nula de esta prueba es

$$\mathscr{H}_0 : \left|e_t^j\right| = \left|e_t^i\right| \quad \forall t \in T_k$$

que es equivalente a

$$\mathscr{H}_0 : \mathbb{E}(d_t) = 0 \quad \forall t \in T_k$$

Bajo dicha hipótesis, el estadístico $DM \sim N(0, 1)$, lo que nos permite rechazarla siempre que $DM > 1.96$, para un nivel de confianza del 95 %, o bajo el criterio del p-value, es decir siempre que este sea menor al 5 %, rechazar la hipótesis nula, en pocas palabras, implica que los modelos tienen distinta precisión al momento de pronosticar, pero hay una variada cantidad de hipotesis alternativas, que cumplen esto, las cuales se exponen a continuación

$$\mathscr{H}_1 = \begin{cases} \text{Ambos modelos tienen distinta precision} \\ \quad \text{El modelo } i \text{ tiene mayor precisión} \\ \quad \text{El modelo } j \text{ tiene mayor precisión} \end{cases}$$

Todas estas opciones, serán contrastadas en las líneas siguientes.

Los resultados obtenidos a partir del código en **R** utilizado para la implementación de la prueba de Diebold y Mariano arrojan las medianas de los p-values obtenidos en el contraste de hipótesis realizado para cada 2 modelos, utilizando un horizonte desde 1 hasta 10 periodos. Los modelos que se encuentran en las filas corresponden a i, mientras que los que están en las columnas, a j

4.4.2. Aplicación de los modelos propuestos

Ambos modelos tienen distinta precisión

En esta prueba puede observarse que las medianas de los p-values obtenidos son considerablemente cercanas a 1. Entonces, la probabilidad de equivocarse, dado que se rechaza la hipótesis nula $\mathscr{H}_0$ es muy alta. Lo anterior sugiere que la probabilidad de tomar una decisión correcta al no rechazar $\mathscr{H}_0$ es baja. En este caso, se tomó la decisión de aceptar la hipótesis nula, ya que, para continuar

con el análisis era necesario tomar una decisión, y de acuerdo a las pruebas de hipótesis, aceptar $\mathscr{H}_0$ es lo más favorable. Dado lo anterior, puede concluirse que todos los modelos poseen el mismo poder predictivo. Por lo cual, resulta innecesario contrastar con las hipótesis alternativas que indican que uno de los modelos tiene mayor precisión que el otro, ya que estas situaciones son un caso particular de la hipótesis alternativa que plantea que todos los modelos tienen distinto poder predictivo, la cual fue rechazada.

El resultado siguiente se obtiene al realizar la prueba para un horizonte de predicción igual a 10 períodos.

```
        M2        M3        M4
M1 0.9496802 0.9695739 0.8787355
M2 0.0000000 0.9740251 0.8684626
M3 0.0000000 0.0000000 0.8885896
```

Lo anterior es consistente con los resultados obtenidos para un horizonte de 1 a 10 periodos de predicción, lo que significa que sin importar cuantas predicciones consideremos, con un máximo de 10 periodos, no se alterará el resultado.

```
       M2        M3        M4
M1 0.944974 0.9720232 0.8910257
M2 0.000000 0.9778832 0.8959688
M3 0.000000 0.0000000 0.8851792
```

De este modo, puede decirse que los pronósticos generados por los cuatro modelos, de forma estadística, no son distintos uno de otro, y por ende, no es posible tomar una decisión respecto a cual es el mejor modelo, bajo la prueba de Diebold y Mariano. Por esta razón se considera pertinente utilizar un criterio distinto para poder seleccionar el modelo más adecuado para la Serie de Tiempo.

4.5. Ajuste con valores observados

Como ya se mencionó, la diferencia entre el valor real de la variable aleatoria y la estimación realizada, da como resultado el valor residual, es decir, $e_t = P_t - \hat{P}_t$ donde $\hat{P}_t = \mathbb{E}_{I_{t-1}}(P_t)$. Por lo tanto, el valor ajustado, $\hat{P}_t$ de la serie de tiempo para los datos se obtiene de la resta de los valores observados menos el residual.

De esta forma se puede obtener, de manera explícita, el pronóstico $\hat{P}_t$, para el mismo horizonte de tiempo en que se cuenta con datos observados, es decir, desde finales de 1996 hasta el segundo semestre del 2019, y compararlos con los mismos, de forma gráfica, como se muestra a continuación.

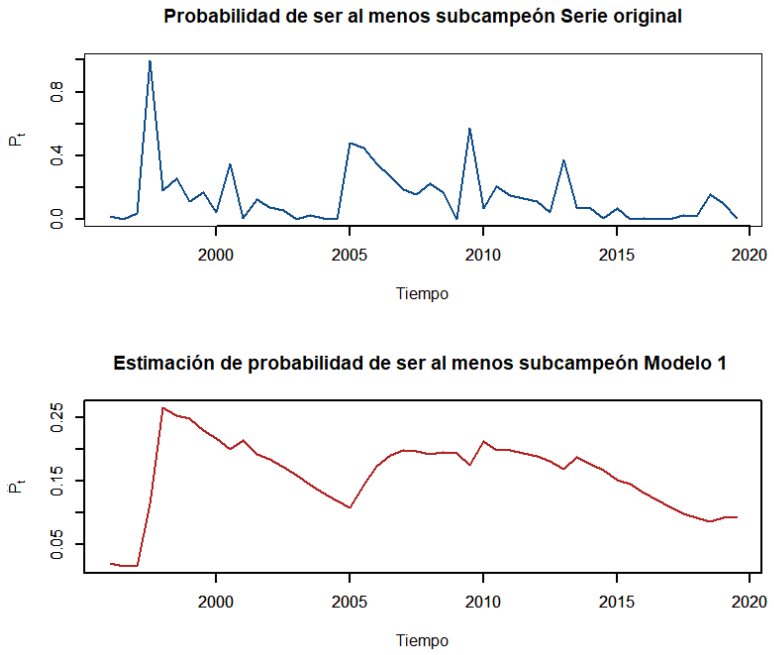

Figura 13: Estimación ARIMA(0,1,1)

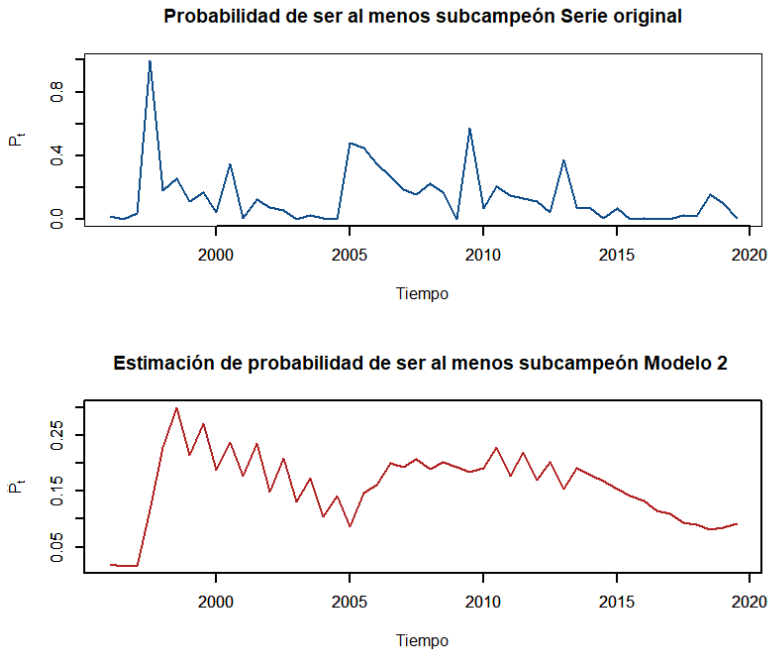

Figura 14: Estimación ARIMA(1,1,2)

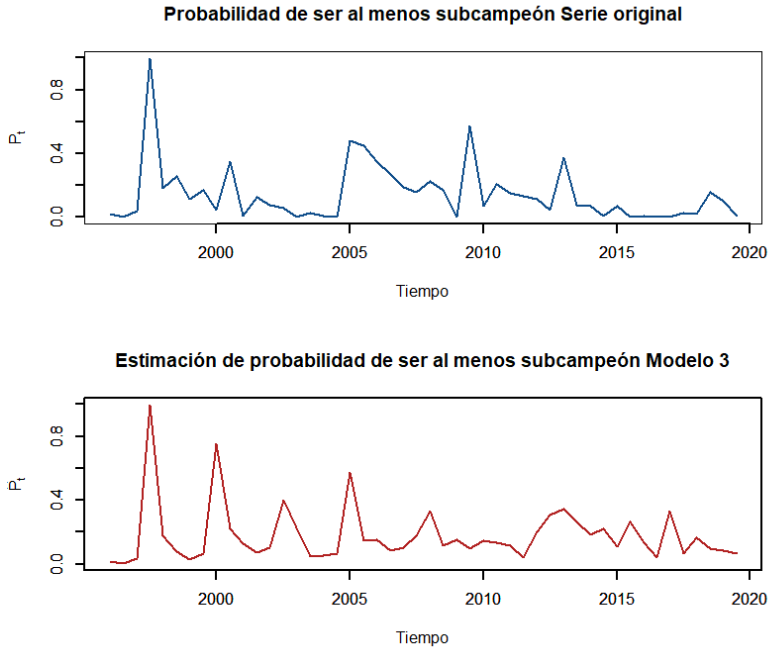

Figura 15: Estimación SARIMA$(2,1,0)_5$

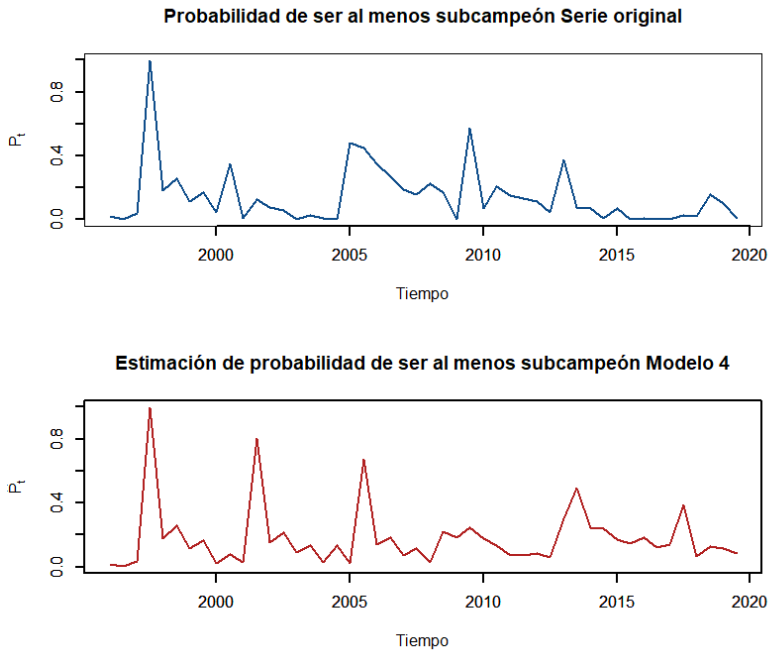

Figura 16: Estimación SARIMA$(1,1,0)_8$

Se puede apreciar, de manera gráfica, que los modelos estacionales se ajustan mejor a la serie original, ya que poseen la característica de reflejar de buena manera el comportamiento cíclico de alzas en la probabilidad, que se observan cada cierto periodo en la serie original. Lo anterior se corrobora graficando los pronósticos a un horizonte de 10 periodos, como se expone en la siguiente sección.

4.6. Pronósticos de los distintos modelos

En los gráficos siguientes se muestra el pronóstico realizado, por cada uno de los modelos, para 10 periodos de predicción, además, se proporciona un intervalo de confianza para dichos pronósticos, y se habla un poco acerca de los mismos.

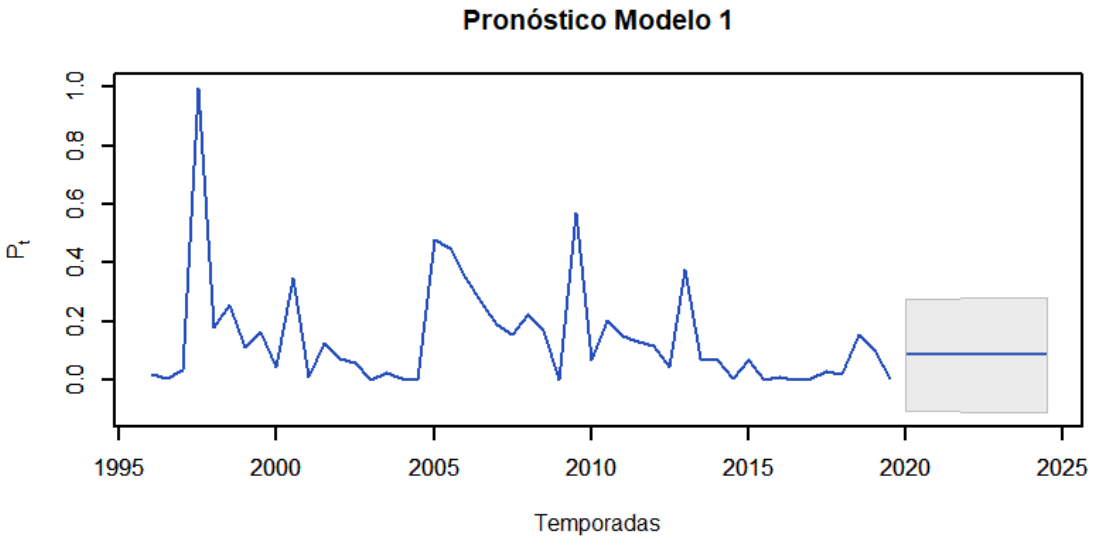

Figura 17: Pronóstico ARIMA(0,1,1)

En la siguiente matriz se muestran el extremo inferior del intervalo, el valor de la predicción, el extremo superior del intervalo, y la desviación estándar, en ese orden, para cada periodo de predicción del primer modelo propuesto, el cual es un ARIMA(0,1,1).

```
     L         P         U     Desv.Est.
-0.1054731 0.0851525 0.2757781 0.1906256
-0.1062996 0.0851525 0.2766046 0.1914521
-0.1071225 0.0851525 0.2774275 0.1922750
-0.1079420 0.0851525 0.2782470 0.1930945
-0.1087579 0.0851525 0.2790629 0.1939104
-0.1095705 0.0851525 0.2798755 0.1947230
-0.1103796 0.0851525 0.2806846 0.1955321
-0.1111855 0.0851525 0.2814905 0.1963380
-0.1119880 0.0851525 0.2822930 0.1971405
-0.1127873 0.0851525 0.2830923 0.1979398
```

Puede observarse que el intervalo contiene valores negativos, sin embargo, la predicción de las temporadas permanece positiva, de hecho, se mantiene constante en 0.0851525.

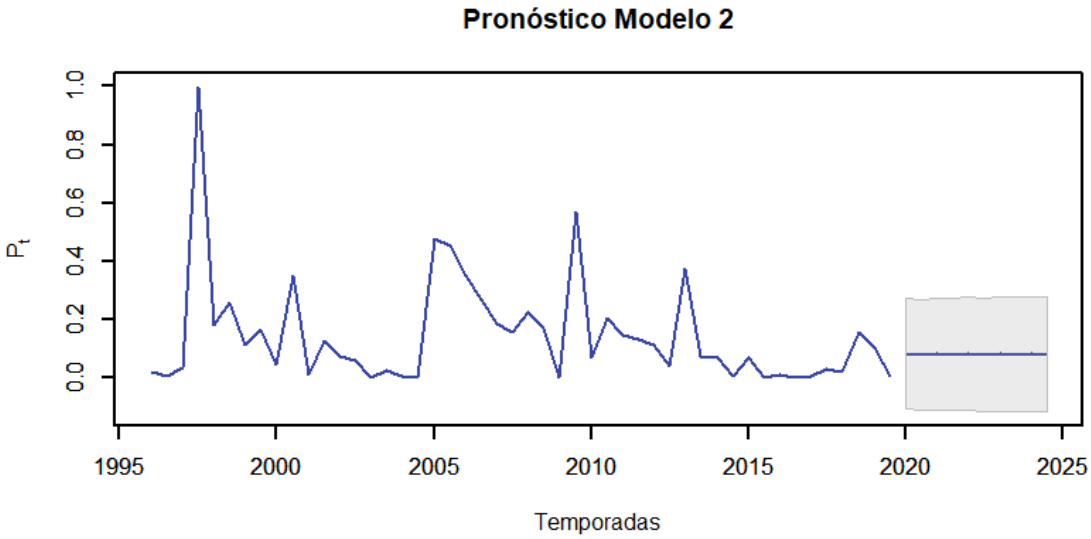

Figura 18: Pronóstico ARIMA(1,1,2)

En la siguiente matriz se muestran el extremo inferior del intervalo, el valor de la predicción, el extremo superior del intervalo, y la desviación estándar, en ese orden, para cada periodo de predicción del primer modelo propuesto, el cual es un ARIMA(1,1,2).

```
     L          P          U       Desv.Est.
-0.1073041 0.08252024 0.2723446 0.1898244
-0.1103961 0.07979718 0.2699905 0.1901933
-0.1095409 0.08235038 0.2742417 0.1918913
-0.1123541 0.07995644 0.2722669 0.1923105
-0.1116859 0.08220106 0.2760880 0.1938869
-0.1142556 0.08009645 0.2744485 0.1943521
-0.1137552 0.08206978 0.2778948 0.1958250
-0.1161121 0.08021954 0.2765512 0.1963316
-0.1157618 0.08195437 0.2796705 0.1977161
-0.1179321 0.08032775 0.2785876 0.1982598
```

Puede observarse que el intervalo contiene valores negativos, sin embargo, la predicción de las temporadas permanece positiva. Nótese que, a diferencia del modelo anterior, la predicción tiene ligeras variaciones, a pesar de eso, los cambios son mínimos, y casi imperceptibles, por lo que, a simple vista, el pronóstico parece ser una recta constante.

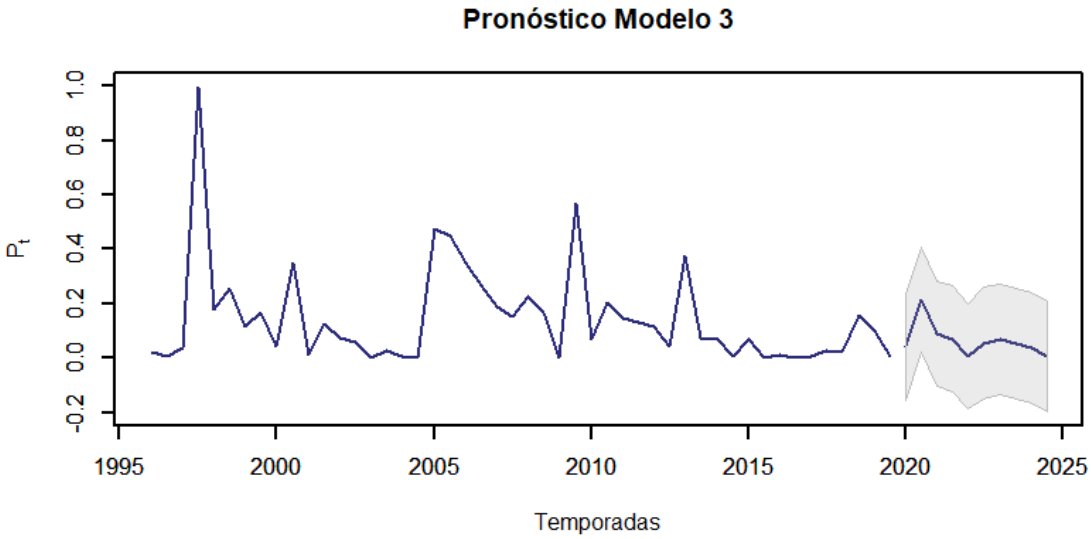

Figura 19: Pronóstico SARIMA$(2,1,0)_5$

En la siguiente matriz se muestran el extremo inferior del intervalo, el valor de la predicción, el extremo superior del intervalo, y la desviación estándar, en ese orden, para cada periodo de predicción del primer modelo propuesto, el cual es un SARIMA$(0,0,0)(2,1,0)_5$

```
      L          P         U      Desv.Est.
-0.15309044 0.041854212 0.2367989 0.1949446
 0.01988163 0.214826281 0.4097709 0.1949446
-0.10501286 0.089931787 0.2848764 0.1949446
-0.12544961 0.069495038 0.2644397 0.1949446
-0.18910757 0.005837076 0.2007817 0.1949446
-0.14817573 0.055526191 0.2592281 0.2037019
-0.13516059 0.068541333 0.2722433 0.2037019
-0.14878982 0.054912102 0.2586140 0.2037019
-0.16678954 0.036912383 0.2406143 0.2037019
-0.19913105 0.004570869 0.2082728 0.2037019
```

Puede observarse que el intervalo contiene algunos valores negativos, sin embargo, la predicción de las temporadas permanece positiva. Cabe resaltar que, en comparación con los modelos ARIMA expuestos anteriormente, este modelo presenta un compotamiento distinto, en donde se muestra un incremento veloz de la probabilidad, del primer al segundo periodo de predicción, para después decrecer hasta llegar a un valor muy cercano a cero, en el quinto periodo. Luego, incrementa de nuevo, de forma menos brusca, y decrece lentamente, alcanzando el valor más pequeño de todas las predicciones dadas por este modelo en un horizonte de 1 a 10 temporadas.

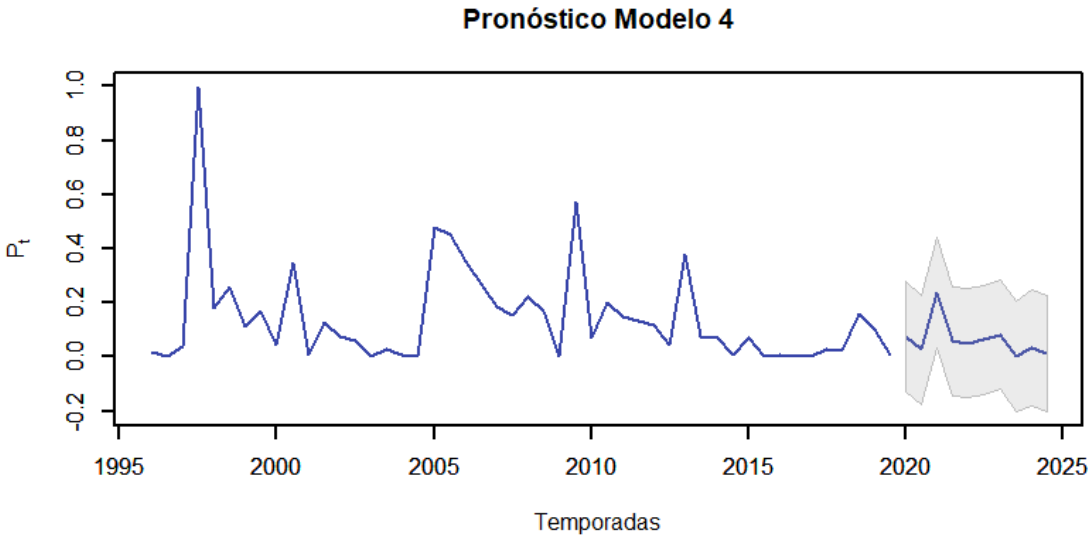

Figura 20: Pronóstico SARIMA$(1,1,0)_8$

En la siguiente matriz se muestran el extremo inferior del intervalo, el valor de la predicción, el extremo superior del intervalo, y la desviación estándar, en ese orden, para cada periodo de predicción del primer modelo propuesto, el cual es un SARIMA$(1,1,0)_8$

```
      L           P          U       Desv.Est.
-0.12637654 0.075455989 0.2772885 0.2018325
-0.17500698 0.026825549 0.2286581 0.2018325
 0.03602032 0.237852848 0.4396854 0.2018325
-0.14435404 0.057478491 0.2593110 0.2018325
-0.15050562 0.051326906 0.2531594 0.2018325
-0.13913196 0.062700567 0.2645331 0.2018325
-0.11975871 0.082073820 0.2839063 0.2018325
-0.19965824 0.002174287 0.2040068 0.2018325
-0.18153522 0.033847096 0.2492294 0.2153823
-0.20417839 0.011203927 0.2265862 0.2153823
```

Puede observarse que el intervalo contiene algunos valores negativos, sin embargo, la predicción de las temporadas permanece positiva. A diferencia de los modelos ARIMA expuestos anteriormente, este modelo presenta un compotamiento distinto, en donde se muestra un pequeño decremento, seguido de un aumento veloz en la probabilidad del segundo periodo de predicción, donde alcanza el valor más alto pronosticado por este modelo en un horizonte de 1 a 10 temporadas. Luego, con la misma fuerza decrece y se mantiene estable durante el cuarto, quinto y sexto periodo, para crecer ligeramente y luego decrecer. Finalmente el periodo nueve tiene un incremento pequeño, el cual vuelve a disminuir para el último periodo de predicción.

Nótese que los pronósticos de los modelos más sencillos tienden a la media de forma muy veloz, lo cual implica que no aportan demasiada información a futuro. Es por ello que se decide escoger entre aquellos modelos que pronostiquen resultados distintos.

5. Conclusiones

A pesar de contar con una cantidad limitada de datos, se lograron proponer cuatro modelos que describieran la probabilidad de que el Cruz Azul fuera, al menos, subcampeón, y tras realizar las pruebas de diagnóstico para los modelos propuestos, se llegó a la conclusión de que todos obtuvieron niveles parecidos, pues desde que se consideró proponer algún modelo, se verificó que los parámetros fueran significativos por medio de la prueba Z.

Por otra parte, para analizar la serie de los residuales, se realizó un análisis descriptivo, así como también se implementaron las pruebas de Ljung-Box y Dickey Fuller, es importante mencionar que por la prueba de Ljung-Box fue posible aceptar la hipótesis nula, la cual consiste en que estos residuos se distribuyen como ruido blanco y de esta manera, se pudieron validar los supuestos de estacionariedad para cada uno de los modelos propuestos.

Adicionalmente, para los criterios de información de Akaike y Bayesiano, se obtuvieron resultados que favorecieron a los modelos con menos parámetros. Teniendo en cuenta que lo anterior no implica que estos modelos ajusten mejor a los datos.

Para obtener un mejor análisis, se realizó la prueba para la parte de pronósticos de Diebold y Mariano, ya que ésta evalúa por medio de una función de pérdida, qué tan parecido es el poder predictivo, sin embargo no se pudo concluir si algún modelo tenía mejores capacidades predictivas que otro.

De esta manera, se generaron las gráficas asociadas a las estimaciones de cada modelo, con el propósito de compararlas respecto a la serie original, de esta manera fue posible apreciar que los patrones de estacionalidad que se presentan en la serie original fueron modelados de mejor manera con el último modelo propuesto, el cual está definido como un modelo SARIMA$(1,1,0)_8$, que aunque en un principio generaba estimaciones poco certeras, se observó que es el modelo que mejor replica la tendencia decreciente, así como el comportámiento de alzas periódicas que tiene la serie asociada a la probabilidad de que el Cruz Azul sea, al menos, sea subcampeón.

Finalmente, al obtener el gráfico correspondiente al pronóstico con horizonte de 10 temporadas, se confirmó que el modelo SARIMA$(1,1,0)_8$ es el mejor, bajo los criterios considerados, dado que arroja predicciones que tienen la misma conduca decreciente y periódica. Esta predicción permite verificar la hipótesis planteada al inicio de la investigación, es decir, bajo el modelo 4, se puede afirmar que la mala racha que el Cruz Azul ha tenido, continuará en un horizonte de 10 periodos de predicción, por lo tanto, se puede concluir que en este horizonte, la probabilidad de que el Cruz Azul sea, al menos, subcampeón será baja.

Referencias

[1] AKAIKE, H. (Dic.,1974) *A new look at the statistical model identification.* IEEE Transactions on Automatic Control 19. No.6 .Vol 19, pp. 716-723.

[2] BROOKS, D.(2018). *"Cruzazulear": el curioso verbo inventado en México por una mala racha del club Cruz Azul. BBC News.* Recuperado de `https://www.bbc.com/mundo/noticias-45253743` el 14 de mayo de 2020.

[3] CASELLA, G., BERGER, R. (2002) *Statistical Inference.* (pp. 397) (2^a ed.) Estados Unidos de América: Duxbury Advanced Series.

[4] CHATFIELD, C., XING, H. (2019) *The Analysis of Time Series: An Introduction with R.*(pp. 83-143) Estados Unidos de América: Chapman and Hall/CRC.

[5] ENDERS,W. (2015). *Applied Econometric Time Series.*(pp. 86-87)(4^a ed.) Estados Unidos de América: Wiley.

[6] ENGLE, R. (Jul.,1982) *Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation.* No.4 .Vol 50, pp. 987-1007.

[7] HAMILTON, J. (1994) *Time Series Analysis.* (pp.72-77) New Jersey: Princeton University Press.

[8] MAHADEVA, L., ROBINSON, P. (2009) *Prueba de raíz unitaria para ayudar a la construcción de un modelo* (pp. 29) (*a* ed.) Centro de Estudios Monetarios Latinoamericanos.

[9] Marín, J. y Velazquez, V. (2019). *Vamos a rescatar a Cruz Azul del corrupto Billy Álvarez.* [Video]. Disponible en `https://www.espn.com.mx/video/clip/_/id/6357272`. Recuperado el 12 de mayo de 2020.

[10] Redacción Vamos Cruz Azul (2020). *Un estudio matemático determinó las posibilidades de Cruz Azul para ganar el Clausura 2020.* Recuperado de https://vamoscruzazul.bolavip.com/ligamx/Un-estudio-matematico-determino-las-posibilidades-de-Cruz-Azul-para-ganar-el-Clausura-2020-20200518-0001.html el 14 de mayo de 2020.

[11] SHUMWAY, R., STOFFER, D. (2011) *Time Series Analysis and Its Application.*(pp. 141-154) (3^a ed.). Estados Unidos de América: Springer.

[12] BOX, G., JENKINS, G., REINSEL, G., LJUNG, G. (2015) *Time series analysis: forecasting and control* (pp. 21-46) (5^a ed.) New Jersey: Wiley Publishing.

[13] WEI,W. (2006) *Time Series Analysis. Univariate and Multivariate Methods.* (pp.18-20) (2^a ed.) Estados Unidos de América: Pearson.

$$\mathbb{P}(I_t > 0)$$