



UNIVERSIDAD DE LA REPÚBLICA
FACULTAD DE CIENCIAS ECONÓMICAS Y DE ADMINISTRACION
PASANTÍA PARA OBTENER EL TÍTULO DE LICENCIADO EN ESTADÍSTICA

MODELIZACIÓN DE LA SEIRE DIARIA DE LLAMADAS ENTRANTES AL
CALL CENTER HOLA ITAÚ

por

GASTÓN CEA

TUTOR: EC. SILVIA RODRIGUEZ

MONTEVIDEO
URUGUAY
NOVIEMBRE 2012

Página de Aprobación

FACULTAD DE CIENCIAS ECONÓMICAS Y ADMINISTRACIÓN

El tribunal docente integrado por los abajo firmantes aprueba la Monografía:

Título

.....
.....

Autor

.....
.....

Tutor

.....

Carrera

.....

Cátedra

.....

Puntaje

.....

Tribunal

Profesor..... (nombre y firma).

Profesor..... (nombre y firma).

Profesor..... (nombre y firma).

FECHA.....

RESUMEN

Este trabajo de pasantía consiste en el análisis y modelización de la serie diaria de llamadas entrantes al Call Center Hola Itaú de Banco Itaú Uruguay S.A. Para modelizar este proceso de llegada de llamadas se siguió la metodología de Box y Jenkins (1976) mediante un ajuste en tres etapas: identificación, estimación y diagnóstico. La serie de datos presentó algunos datos faltantes, los cuales fueron estimados mediante el “Enfoque de Outlier Aditivo” para estimación de datos faltantes, Gomez et al. (1993). Se analizaron, con especial énfasis, las posibles estacionalidades presentes en la serie y el efecto calendario sobre la misma. La estacionalidad principal elegida y tratada de forma estocástica fue la estacionalidad semanal. Se probó una posible estacionalidad secundaria, tratada de forma determinística: la estacionalidad anual, resultando no significativa para el modelo planteado; por otro lado la estacionalidad mensual fue descartada en el inicio a través de un análisis descriptivo. Para el tratamiento del efecto calendario se definieron siete variables *dummy* las cuales identificaron los días: feriados, noche buena y los cinco efectos de feriados que fueron definidos. Este tratamiento del efecto calendario surgió de análisis descriptivos (que se detallan en el cuerpo del documento) y con la ayuda del responsable del Call Center Hola Itaú: Diego Castillo. Se incluyeron diez outliers de tipo AO y un Outlier de tipo TC según la clasificación expuesta en Hillmer et al. (1983). Finalmente, fue identificado un modelo *SARIMA* $(1\ 0\ 0)(0\ 1\ 1)_5$ con una única estacionalidad estocástica (semanal), al que se le adicionan regresores por efecto calendario e intervenciones. Luego se analizó el modelo como herramienta predictiva en el corto plazo (5 días hábiles), con un error promedio de aproximadamente 10%.

DESCRIPTORES

Modelización, serie diaria, Call Center, Box y Jenkins, outlier, estacionalidad, efecto calendario, SARIMA, predicción, error promedio.

TABLA DE CONTENIDO

0. <u>INTRODUCCION</u>	7
1. <u>MARCO METODOLÓGICO</u>	8
1.1 ANOVA DE UN FACTOR.....	8
1.2 MODELOS ARMA & SARIMA.....	11
1.3 ESTACIONALIDAD.....	15
1.4 EFECTO CALENDARIO	18
1.5 METODOLOGÍA DE BOX – JENKINS	21
2. <u>RESULTADOS</u>	27
2.1 ESTACIONALIDAD.....	27
2.2 EFECTO CALENDARIO	35
2.3 TRATAMIENTO PARA DATOS FALTANTES.....	38
2.4 AJUSTE A UN MODELO SARIMA.....	41
2.5 EL MODELO COMO HERRAMIENTA DE PREDICCIÓN. 53	
3. <u>CONCLUSIÓN</u>	60
4. <u>BIBLIOGRAFÍA</u>	61
5. <u>ANEXO</u>	63

ÍNDICE DE ILUSTRACIONES, CUADROS Y GRÁFICOS

ÍNDICE DE ILUSTRACIONES

Figura	Título de la Figura	Página
2.1	Esquema para la metodología Box y Jenkins	41

ÍNDICE DE CUADROS

Cuadro	Título de Cuadro	Página
1.1	Identificación para órdenes autorregresivos y de medias móviles	22
2.1	Estadísticos descriptivos para prueba ANOVA	31
2.2	Prueba de homogeneidad de varianzas para prueba ANOVA	32
2.3	Prueba de comparación de medias Post-hoc	33
2.4	Promedio de llamadas entrantes por tipo de día	39
2.5	Estimaciones iniciales para datos faltantes	39
2.6	Datos descriptivos de la serie de llamadas entrantes	42
2.7	Coefficientes del modelo identificado	46
2.8	Datos descriptivos de los residuos del modelo estimado	48
2.9	Pruebas de incorrelación	50
2.10	Pruebas de normalidad	52
2.11	Datos descriptivos de los residuos en valor absoluto	52
2.12	Predicciones a un paso para las últimas tres semanas completas de la serie	54
2.13	Predicciones a cinco pasos para las últimas tres semanas completas de la serie	55
2.14	Predicciones a un paso para semanas con feriados	56
2.15	Predicciones a cinco pasos para semanas con feriados	57
2.16	Predicciones a cinco pasos a partir de la observación n°38	58
2.17	Predicciones fuera de la muestra a siete pasos	59

ÍNDICE DE GRÁFICOS

Gráfico	Título de Gráfico	Página
2.1	Llamadas entrantes Junio 2006, Setiembre 2007 y Marzo 2008	27
2.2	Diagramas de caja mensuales	28
2.3	Diagramas de caja diarios	30
2.4	Diagramas de caja diarios sin observaciones alejadas	31
2.5	Efectos prospectivos y retrospectivos de los feriados	36
2.6	Serie diaria de llamadas entrantes al Call Center Hola Itaú desde el 10 de abril de 2006 hasta el 8 de abril de 2009	42
2.7	Autocorrelograma y Autocorrelograma Parcial de la serie de datos	43
2.8	Autocorrelograma y Autocorrelograma Parcial de la primera diferencia estacional	44
2.9	Residuos del modelo identificado	45
2.10	Gráfico comparativo entre serie original y serie modelada	47
2.11	Residuos del modelo estimado	48
2.12	Gráfico comparativo entre histograma de residuos y densidad de distribución Normal (0, 100)	49
2.13	Autocorrelogramas de los residuos y de los residuos al cuadrado del modelo	50
2.14	Gráfico Q-Q Normal de los residuos	51
2.15	Predicciones fuera de la muestra a siete pasos	59

0. INTRODUCCIÓN

El objetivo de la pasantía consistió en poder modelizar adecuadamente el proceso de llegada de llamadas al Call Center Hola Itaú de Banco Itaú Uruguay S.A., de forma tal de poder predecir en el corto plazo, considerando como corto plazo el lapso de una semana hábil (5 días).

Ese modelo servirá de herramienta para poder establecer, de forma más eficiente, la dotación de personal necesaria en el Call Center y así poder afrontar de mejor manera el cumplimiento de los objetivos que este sector se traza año a año.

La base de datos fue brindada por el banco y consiste en 778 observaciones, desde el 10 de abril de 2006 hasta el 8 de abril de 2009. Para cada observación (momento del tiempo) se tienen los siguientes datos: cantidad de llamadas entrantes, fecha (día, mes, año) y día de la semana.

Este tipo de series diarias, presentan características muy particulares, las que hacen que su modelización sea distinta a la de series de mayor agregación temporal. Por ejemplo, es necesario analizar las diferentes estacionalidades, algunas de las cuales pueden presentarse simultáneamente; así como también el efecto calendario sobre la serie.

Para abordar la modelización del proceso de llegada de llamadas se siguió la metodología propuesta por Box y Jenkins, y así se planteó un ajuste en tres etapas (identificación, estimación y diagnóstico) a un modelo SARIMA, con especial énfasis en analizar el efecto calendario y posibles estacionalidades múltiples teniendo en cuenta la frecuencia de la serie.

Por último se analizó el modelo planteado como herramienta de predicción a uno y a cinco pasos. Utilizándolo, en este sentido, para diferentes momentos de la serie de datos.

1. MARCO METODOLÓGICO

1.1 ANOVA DE UN FACTOR

A continuación se presentan modelos y nociones teóricas de la prueba ANOVA de un factor y las pruebas post hoc. *¹

El análisis de varianza ANOVA de un factor sirve para comparar varios grupos en una variable cuantitativa. La hipótesis nula en la prueba ANOVA sugiere que las medias poblacionales son iguales. Para poner a prueba la hipótesis nula de igualdad de medias, se obtiene el estadístico F que refleja el grado de similitud entre las medias poblacionales. Dicho estadístico es de la forma:

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{n\sigma \frac{2}{Y}}{\bar{S}_Y^2}$$

El numerador del estadístico es una estimación de la varianza poblacional, pero está basada en la variabilidad existente dentro de cada grupo. Si las medias poblacionales son iguales, las medias muestrales no serán significativamente distintas a cierto nivel de confianza, existiendo entre ellas tan sólo diferencias atribuibles al azar. En este caso la estimación $\hat{\sigma}_1^2$ (basada en las diferencias entre las medias poblacionales) reflejará el mismo grado de variación que la estimación $\hat{\sigma}_2^2$ (basada en las diferencias entre las puntuaciones individuales) y por consiguiente el estadístico F tomará un valor próximo a 1. Por otro lado, si las medias son significativamente distintas, la estimación $\hat{\sigma}_1^2$ reflejará un mayor grado de variación que la estimación $\hat{\sigma}_2^2$ por lo que el estadístico F tomará un valor significativamente mayor a 1. Es decir, cuanto mayor sean las diferencias entre las medias, mayor será el valor que adquiera el estadístico F . Si suponemos que las poblaciones muestreadas se distribuyen Normal y sus varianzas son iguales, entonces el estadístico F se distribuye según el modelo de Probabilidad F de Fisher-Snedecor. De esta forma si, por ejemplo, el nivel crítico asociado al estadístico F es menor a 0,05 es posible rechazar la hipótesis de igualdad de medias con un nivel de confianza del 95%.

*¹ Consultar en Bibliografía: María Ángeles Cea D`Ancona (2004).

Al rechazar esta hipótesis, surge la interrogante de; ¿en dónde se encuentran las diferencias que permiten rechazar la hipótesis nula con cierto nivel de confianza?, ¿hay sólo una media que difiere de las demás?. Para resolver este tipo de interrogantes es necesario aplicar pruebas post hoc o a posteriori, estas comparaciones permiten controlar la probabilidad de cometer errores de tipo I al tomar decisiones sobre la igualdad de medias entre los grupos comparados.

Si se asume varianzas iguales entre las muestras comparadas es posible aplicar las siguientes pruebas post hoc:

DMS: Diferencia Mínima Significativa, prueba basada en la distribución t de Student. Este método propuesto por Fisher no ejerce ningún control sobre la tasa de error, cada comparación se realiza utilizando el nivel de significación establecido por lo que la tasa de error puede alcanzar $1 - (1 - \alpha)^k$, siendo α el nivel de significación y k el número de comparaciones a realizar.

Bonferroni: el método de Bonferroni también está basado en la distribución t de Student además de basarse en la desigualdad de Bonferroni. El método controla la tasa de error dividiendo el nivel de significación entre el número de comparaciones, cada comparación se evalúa utilizando un nivel particular de significación $\alpha_i = \alpha/k$.

Sidak: es similar a la prueba de Bonferroni, pero controla la tasa de error evaluando cada comparación con un nivel de significación: $\alpha_i = 1 - (1 - \alpha)^{1/k}$, rechazando así la hipótesis de igualdad de medias de forma menos rígida que el método de Bonferroni.

Scheffé: este método se basa en la distribución F y es utilizado para efectuar únicamente comparaciones por pares. Es considerado un método muy conservador debido a que es muy exigente para rechazar la hipótesis de igualdad de medias.

R-E-G-W F: Método de Ryan, Einot, Gabriel y Welsch basado en la distribución F. Consiste en un método por pasos, ordena de forma ascendente las k medias por su tamaño y efectúa todas las comparaciones posibles entre pares de medias. De este modo la media más pequeña y la más grande están separadas por k escalones, la media más pequeña y la segunda más grande por k - 1 escalones y siguiendo esa misma lógica se observa que dos medias adyacentes en el ordenamiento están separadas por 2 escalones. El número de escalones existente entre dos medias comparadas condiciona el nivel de significación de cada comparación, siendo éste mayor cuanto mayor la separación entre las medias a comparar dentro del ordenamiento mencionado. De esta forma el método efectúa la comparación utilizando un estadístico F y un nivel de significación $\alpha_i = 1 - (1 - \alpha)^{r/k}$, siendo r el número de escalones que separa a las dos medias comparadas. Si bien es un método muy potente, no es apropiado cuando los grupos tienen tamaños significativamente distintos.

R-E-G-W Q: Método de Ryan, Einot, Gabriel y Welsch basado en la distribución del rango studentizado. Método similar al R-E-G-W F pero basado en otra distribución, posee similares características.

S-N-K: Método de Student-Neuwman-Jeuls basado en la distribución del rango studentizado. Al igual que los métodos R-E-G-W F y R-E-G-W Q también se basa en el ordenamiento de medias por tamaño, pero a diferencia de éstos el nivel de significación para cada comparación de dos medias es siempre α , sin importar el número de pasos que las separe. Cuanto mayor el número de escalones que separa dos medias dentro del ordenamiento, mayor la diferencia mínima necesaria para considerar que las medias difieren significativamente.

Tukey: Diferencia honestamente significativa de Tukey. Equivale a utilizar el método S-N-K con el número de escalones equivalente al número de medias. Por lo tanto todas las comparaciones son referidas a la misma diferencia mínima.

Tukey-b: Este método consiste en considerar como diferencia mínima el valor medio entre la diferencia honestamente significativa de Tukey y la diferencia mínima obtenida con el método S-N-K para el caso en que el número de escalones que separan a dos medias es igual a 2.

Duncan: Método de comparación por pasos basado en la distribución del rango studentizado. Controla la tasa de error utilizando un nivel de significación $\alpha_i = 1 - (1 - \alpha)^{r-1}$ para el conjunto de medias separadas por r pasos (o escalones). Cuanto mayor el número de escalones que separa dos medias dentro del ordenamiento, mayor la diferencia mínima necesaria para considerar que las medias difieren significativamente.

Gabriel: Se basa en la distribución del módulo máximo studentizado. Es muy potente cuando se comparan grupos del mismo tamaño pero ocurre lo contrario cuando los tamaños de los grupos a comparar difieren significativamente.

Existen otras pruebas post hoc recomendadas para la aplicación en casos en que no se asume varianzas iguales entre las muestras comparadas, algunas de ellas son:

Games-Howell: Es un método similar al de Tukey. Se basa en la distribución del rango studentizado y en un estadístico T en el que, tras estimar las varianzas poblacionales suponiéndolas distintas, se corrigen los grados de libertad mediante la ecuación Welch. Es considerado el método que mejor controla la tasa de error, dentro de los métodos aconsejables para la aplicación de comparaciones en casos en que no se asume varianzas iguales y de los métodos disponibles para aplicación en el software SPSS.

C de Dunnett: Método idéntico al de Games-Howell excepto por la forma de corregir los grados de libertad de la distribución del rango studentizado. Es un método más conservador que el de Games-Howell

1.2 MODELOS ARMA & SARIMA

Proceso Estocástico

Un proceso estocástico es una familia de variables aleatorias. Los procesos estocásticos se clasifican como de tiempo discreto, o de tiempo continuo; según el tipo de espacio probabilístico del que se trate. También pueden clasificarse como de espacio discreto, o de espacio continuo; dependiendo de si el espacio de estados es finito o infinito numerable.

Definición de Serie de Tiempo

Una serie de tiempo es la realización de un proceso estocástico. Es el resultado de observar los valores de una variable registrados en varios intervalos de tiempo. Las series de tiempo pueden ser discretas o continuas, según lo sea su conjunto de observaciones, dependiendo de si éste es discreto o si se registra continuamente.

Modelos Autorregresivos

Definimos un modelo como autorregresivo cuando la variable endógena de un período t es explicada por las observaciones de ella misma correspondientes a períodos anteriores, añadiéndose una innovación.

Este tipo de procesos se abrevian con la sigla AR a la que se le incluye el orden del modelo al que se haga mención. El orden del modelo expresa el número de observaciones retrasadas de la serie analizada, tenidas en cuenta para la elaboración de la ecuación del modelo.

La expresión genérica de un modelo autorregresivo AR (p) se denota del siguiente modo:

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p} + a_t$$

Siendo $\phi_1, \phi_2, \dots, \phi_p$ los coeficientes del modelo, los cuales pueden ser estimados a través del método de mínimos cuadrados, el método de máxima verosimilitud o el método de los momentos. ϕ_0 es la constante del modelo y a_t representa un proceso de Ruido Blanco (RB) gaussiano.

Un proceso puede denominarse como de Ruido Blanco gaussiano, si y sólo si, cumple con las siguientes características:

- $E(a_t) = 0 \quad \forall t$
- $E(a_t^2) = \sigma^2 \quad \forall t$
- $E(a_t, a_s) = 0 \quad \forall t \neq s$
- $a_t \approx Normal$

Modelos de Medias Móviles

Un modelo de los denominados de medias móviles es aquel que explica el valor de una determinada variable en un período t en función de un término independiente y una combinación de procesos RB correspondientes a períodos precedentes.

Estos modelos son abreviados con la sigla MA, incluyéndosele a ésta el orden del modelo entre paréntesis. De este modo un modelo de medias móviles MA (q) con q términos de error se denota de la siguiente forma:

$$Y_t = \theta_1 + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \dots + \theta_q a_{t-q} + a_t$$

Donde $\theta_1, \theta_2, \dots, \theta_q$ son los coeficientes del modelo, θ_1 la constante y a_t representa un proceso de Ruido Blanco gaussiano.

Modelos Autorregresivos de Medias Móviles

Los modelos ARMA son modelos mixtos, son la conjunción de los procesos Autorregresivos: AR, y procesos de Medias Móviles: MA. El orden de estos modelos se denota (p,q); siendo p el orden de la parte autorregresiva y q el de la parte de medias móviles. De esta forma el modelo ARMA (p,q) se denota:

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} + \dots + \theta_q a_{t-q}$$

los coeficientes del modelo $\phi_1, \phi_2, \dots, \phi_p, \theta_1, \theta_2, \dots, \theta_q$, corresponden a la parte autorregresiva y la parte de medias móviles, respectivamente. C es la constante del modelo y a_t representa un proceso de Ruido Blanco gaussiano.

Modelos Autorregresivos Integrados de Medias Móviles

Los modelos ARIMA se construyen a partir de los modelos ARMA, teniendo en cuenta que para que la serie en estudio sea estacionaria en media se le tendrán que aplicar diferencias. Un modelo ARIMA (p,q,d) es un modelo ARMA (p,q) al que se le aplican d diferencias. De esta forma el modelo ARIMA (p,q,d) se denota:

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_p X_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} + \dots + \theta_q a_{t-q}$$

Donde $X_t = (1-L)^d \cdot Y_t$

Estacionariedad

Un proceso estocástico se dice estacionario si: diferentes funciones de distribución conjuntas respecto a intervalos de tiempo de igual longitud, son equivalentes.

Considerando que $t, t+1, t+2, \dots, t+k$ reflejan períodos sucesivos:

$$F(Y_t, Y_{t+1}, \dots, Y_{t+k}) = F(Y_{t+m}, Y_{t+1+m}, \dots, Y_{t+k+m})$$

Esta definición de estacionariedad se conoce como estacionariedad fuerte. El concepto de estacionariedad fuerte se diferencia de su condición menos estricta, la estacionariedad débil. Decimos que un proceso estocástico es débilmente estacionario si:

Las esperanzas matemáticas de las variables aleatorias son constantes en el tiempo:

$$E[Y_t] = E[Y_{t+m}] \quad \forall m$$

Las varianzas de las variables aleatorias son finitas y constantes en el tiempo:

$$Var[Y_t] = Var[Y_{t+m}] \neq \infty \quad \forall m$$

Las covarianzas entre dos variables aleatorias son constantes en distintos períodos de tiempo:

$$Cov[Y_t, Y_s] = Cov[Y_{t+m}, Y_{s+m}] \quad \forall m$$

De estas definiciones se desprende que, la estacionariedad fuerte implica estacionariedad débil pero no ocurre lo mismo en sentido contrario. Aunque si las distribuciones conjuntas se distribuyen de forma Normal, entonces la estacionariedad débil sí implica estacionariedad estricta y viceversa.

Modelos SARIMA

Se define un modelo SARIMA (Seasonal Autoregressive Integrated Moving Average), como un modelo ARIMA que además incorpora efectos de la estacionalidad.

Estos modelos ARIMA estacionales se denotan de la forma SARIMA (p d q)(P D Q)_s, en la que P, D, y Q son parámetros análogos a los parámetros p, d, q, para con su componente estacional. A su vez s toma el valor de la periodicidad de la serie.

Resumiendo se tiene que:

p: número de parámetros autorregresivos.

d: número de diferenciaciones para que la serie sea estacionaria.

q: número de parámetros de medias móviles.

P: número de parámetros autorregresivos en la parte estacional.

D: número de diferenciaciones para que la serie sea estacionaria, en la parte estacional.

Q: número de parámetros de medias móviles en la parte estacional.

s: periodicidad de serie (s=12 serie mensual, s=4 serie trimestral)

1.3 ESTACIONALIDAD

Desde el punto de vista de las series cronológicas, estacional, puede ser definido como todo patrón que se repite periódicamente en el mismo período de tiempo todas las semanas, todos los meses, todos los años, etc.. Para series con observaciones mensuales Box-Jenkins (1976, p. 303-04) definen estacionalidad como: “*la relación existente entre las observaciones para el mismo mes en años sucesivos*” y destacan: “*el efecto estacional implica que una observación de un mes en particular, por ejemplo abril, está relacionada con las observaciones de los abriles previos*” *². Esta definición puede ser aplicada análogamente a series de mayor frecuencia, por ejemplo, series diarias. En este tipo de series es que surge lo que en la literatura de series cronológicas se denomina estacionalidades simultáneas. De esta forma podemos definir, por ejemplo, tres tipos de estacionalidades: la relación existente entre las observaciones para el mismo mes en años sucesivos, la relación existente entre las observaciones para la misma semana en meses sucesivos y, la relación existente entre las observaciones para el mismo día de la semana en semanas sucesivas.

En Espasa et al. (1996, p. 3) se plantea una pregunta sobre si las técnicas habituales que se aplican a series mensuales, trimestrales, etc (series de baja frecuencia), pueden ser directamente aplicadas a series diarias o si resulta conveniente desarrollar técnicas específicas para estas series que tengan en cuenta su problemática particular. Para resolver esta cuestión los autores distinguen, en una primera instancia, los aspectos característicos de series de baja frecuencia que facilitan el desarrollo de metodologías sistematizadas de tratamiento, como la metodología ARIMA, desarrollada por Box y Jenkins (1976):

- a) *Existencia de un único ciclo estacional,*
- b) *cuyo periodo parece estar perfectamente determinado y*
- c) *para el que esquemas sencillos de naturaleza estocástica o determinística parecen adecuados.”*

Luego destacan los siguientes aspectos sobre las series diarias (de alta frecuencia):

- a) *Existencia de varios ciclos estacionales sobrepuestos, siendo los más comunes el semanal, mensual y anual.*
- b) *Aparición de ciclos de período variable por irregularidades del calendario, como la existencia de años bisiestos, las distintas duraciones de los meses o el efecto de la presencia de distinto número de fines de semana en los diferentes meses.*

*² Según traduce J.A. Caracena (2002, p. 1) .

- c) *Exigencia de combinación de esquemas determinísticos y estocásticos para captar mejor las estacionalidades.*
- d) *Los esquemas determinísticos, cuando son necesarios para un determinado efecto cíclico, suelen ser variables en función de otro ciclo o de variables meteorológicas.*
- e) *Dependencia importante, y con frecuencia de naturaleza no lineal, de variables exógenas como las meteorológicas, y muy especialmente de la temperatura, lluvia, luminosidad y viento.*
- f) *Complejo efecto de calendario en lo que se refiere a fiestas, periodos vacacionales, etc. ”*

Y, en base a todo lo anterior, concluyen respecto a las series de alta frecuencia: *“resulta difícil la búsqueda sistemática de modelos tomando como base los planteamientos comúnmente utilizados y se revela de gran interés el planteamiento de estrategias específicas de modelización para este tipo de series”*

Para la modelización es posible recurrir a dos tipos de esquemas base: determinístico o estocástico, aunque las características específicas de algunas series obligan al uso de esquemas mixtos de éstos dos.

Esquemas determinísticos:

La modelización de un ciclo de período s que tenga una evolución constante en el tiempo puede realizarse con s variables artificiales, de la forma:

$$Y_t = w_1 \partial_{1t} + w_2 \partial_{2t} + \dots + w_s \partial_{st}$$

Donde w_i son constantes y la suma de los mismos valdrá 0, ∂_{it} son variables artificiales que toman el valor 1 cuando t se corresponde al momento estacional i y 0 para otro caso.

Esquemas estocásticos:

La modelización de una estacionalidad estable de período s mediante esquemas estocásticos generalmente se hace mediante la aplicación a la serie original del operador suma:

$$U_{s-1} = (1 + L + L^2 + \dots + L^{s-1})$$

Modelización de distintas estacionalidades simultáneas

Lo más común en series de alta frecuencia es la presencia de más de una estacionalidad, por lo que son varios los esquemas de modelización a combinar. Planteándose así dos problemas fundamentales: el solapamiento de armónicos de frecuencias similares

provenientes de los distintos operadores suma y la aplicación de armónicos correspondientes a periodicidades no presentes en la serie.

Por todo esto es que la recomendación señalada en Espasa et al. (1996, p. 9) parte de la selección de una de las estacionalidades como principal, dándole prioridad al elegir su esquema (que la mayor parte de las veces será estocástico). Y luego las estacionalidades residuales de otros períodos se recogen de forma más precisa mediante conjuntos de esquemas determinísticos.

1.4 EFECTO CALENDARIO *³

El efecto producido en una determinada serie por los feriados asignados al calendario varía dependiendo del tipo de calendario y de serie que se trate. Es por esto que el tratamiento del efecto calendario (EC) es particularmente complejo, no sólo habrá que considerar la cantidad de feriados a analizar sino también la ubicación dentro de calendario de los mismos, así como las diferencias de impacto derivadas de la importancia de cada feriado, y factores sociológicos asociados al comportamiento de los individuos y colectivos sociales que llevan a ciertas generalidades en cada feriado o conjunto de feriados.

Por otro lado, si no se considera el EC dentro del tratamiento de cualquier serie de datos que se vean afectados por éste, los resultados de cualquier modelización sobre esos datos correrán el riesgo de ser inexactos, los residuos de dicho modelo contendrán los errores derivados de la no consideración del EC.

En Revuelta, J. M. (2001) se divide al EC en dos grandes grupos: (1) **fiestas o eventos especiales** y (2) **períodos vacacionales**. Dentro del primer grupo se incluye períodos cuyo efecto sobre el fenómeno de estudio sea claro y presente una corta duración temporal y dentro del segundo grupo incluye efectos con una dinámica más extensa y compleja. Revuelta destaca también que es de gran importancia considerar los diferentes tipos de fiestas o eventos especiales, distinguiendo la ponderación del efecto que cada uno conlleva. Para esta clasificación resulta fundamental el conocimiento de la serie a tratar por parte del analista, para poder asignar toda la heterogeneidad en los comportamientos de las fiestas (o eventos especiales) a la serie en cuestión. También destaca los atributos que afectan al impacto de las fiestas (o eventos especiales), atributos tales como: el día de la semana en que tienen lugar, el momento del mes o el momento del año, pudiendo variar en cada caso, los efectos que la fiesta (o evento especial) tenga sobre la serie de datos a analizar.

Cabe señalar que el efecto sobre el nivel de una determinada serie no se restringe al día específico en que la fiesta o el evento especial tiene lugar, sino que puede tener efectos retroactivos y/o prospectivos (afectando a períodos indefinidos de días previos y/o posteriores).

El EC es un fenómeno cambiante respecto al espacio y al tiempo, por la sencilla razón de que cada calendario depende de su contexto temporal, político, geográfico y social. En muchos casos las fiestas (o eventos especiales) y los períodos vacacionales son designados institucionalmente y se basan, generalmente, en hechos presentes o pasados que son considerados de relevancia social.

*³ Consultar en Bibliografía: Revuelta, J. M. (2001).

Por lo tanto, a continuación destacamos algunas generalidades del EC correspondientes al calendario actual que rige en la República Oriental del Uruguay, teniendo en cuenta también calendarios anteriores y sus diferencias con éste.

Considerando las agrupaciones y clasificaciones sugeridas por Revuelta y expuestas anteriormente, podemos distinguir dos grandes tipos de fiestas en Uruguay: (1) **feriados no laborables** y (2) **feriados laborables**. Los feriados no laborables son 5 a lo largo de cada año y la fecha en que corresponde la aplicación del feriado es inamovible. El artículo 18 de la Ley 12.590 del 23 de diciembre de 1958 establece sobre los feriados no laborables: "*Los días 1º de enero, 1º de mayo, 18 de julio, 25 de agosto y 25 de diciembre de cada año, todo trabajador percibirá remuneración como si trabajara; y en caso de trabajar recibirá doble paga*".

Por otro lado, los feriados laborables son 10 a lo largo del año y tienen la particularidad de ser móviles, pueden ser establecidos en una fecha distinta a la que corresponde según el día de la semana en que estos acontezcan y según el año en cuestión. Otro aspecto a tener en cuenta es que esta movilización de los feriados depende de decisiones institucionales que pueden modificarse año a año, ha habido años en los que los feriados laborables se mantuvieron fijos y otros en los que no. Estos aspectos deben ser considerados a los efectos de identificar los feriados, los efectos que éstos tienen sobre el nivel de la serie de estudio y las estacionalidades que la serie presenta.

Para tomar en cuenta los atributos de las fiestas, habrá que considerar características de la sociedad uruguaya. Respecto a los atributos semanales, cabe señalar que los días domingo no son laborables y, para muchos trabajadores, los días sábado tampoco lo son o, en caso de serlo, su horario laboral es parcial. Esta particularidad hace que el efecto que la fiesta pueda tener sobre la serie de datos a analizar dependa, en muchos casos, del día de la semana en que la fiesta tenga lugar. A modo de ejemplo, en lo que refiere a los atributos mensuales, una característica a señalar y que puede afectar a muchas series de datos económicos, consiste en que el momento de cobro mensual para los trabajadores asalariados, en la mayoría de los casos, es a principios de cada mes calendario. Y respecto a los atributos anuales habrá que tener en cuenta que meses incluyen cada estación del año. A grandes rasgos podrían agruparse del siguiente modo: enero, febrero y marzo constituyen el verano; abril, mayo y junio el otoño; julio, agosto y setiembre el invierno; y octubre, noviembre y diciembre la primavera. Este hecho tiene fundamental importancia en series de datos que presentan estacionalidades anuales, tal es el caso de series de datos turísticos, de vestimenta, etc.

Períodos vacacionales

Existen dos grandes períodos vacacionales dentro del calendario uruguayo, estos son: Semana Santa y Carnaval. Ambos períodos afectan a un importante porcentaje de la población en todos los rubros abarcando todas las edades, y la duración en los dos casos es de entre dos días y una semana.

En una segunda categoría se encuentran los períodos de vacaciones estudiantiles, estos son: vacaciones de verano, vacaciones de invierno y vacaciones de primavera. El primer período tiene lugar aproximadamente entre el 15 de diciembre y el 15 de marzo, dependiendo del instituto de enseñanza y del año que se considere. Las vacaciones de invierno constan de dos semanas de duración y se ubican, generalmente, dentro del mes de julio. Las vacaciones de primavera tienen entre dos días y una semana de duración, dependiendo del instituto de enseñanza que se trate, habitualmente se designan dentro del mes de setiembre.

Una característica específica de estos períodos vacacionales es que los mismos son móviles, pudiendo tener lugar en diferentes meses según el año del que se trate, lo que hace que para algunas series sea necesario un tratamiento especial al no presentarse una estacionalidad estricta.

1.5 METODOLOGÍA DE MODELIZACIÓN DE BOX Y JENKINS

Box y Jenkins (1976) desarrollaron un cuerpo metodológico con la finalidad de modelizar y predecir mediante modelos dinámicos de series temporales. La modelización consta de tres etapas: identificación, estimación y diagnóstico.

Identificación

Dentro de la primera etapa habrá que verificar la estacionariedad del modelo. Existen transformaciones que pueden aplicársele a los efectos de lograr que el modelo sea estacionario. A modo de ejemplo, si el modelo no es estacionario en media puede aplicársele alguna diferencia y lograr que de ese modo sí lo sea, o si no lo es en varianza, podría aplicársele una transformación logarítmica a los mismos efectos. Una vez identificado el modelo estacionario se podrá estimar luego un modelo en base a las observaciones previas. Será necesario identificar también las intervenciones a aplicar en la serie temporal original de forma de considerar outliers y otras variables de regresión que contribuyan a la modelización de la serie. Si la frecuencia de la serie con la que se trabaja es mayor a la frecuencia mensual habrá que definir la estacionalidad principal y, de existir, estacionalidades secundarias. Para concluir esta primera etapa se identificará uno o más modelos AR, MA, ARIMA o SARIMA que puedan ajustar la serie, para luego poder estimar los parámetros de dichos modelos y finalmente diagnosticar los resultados obtenidos y seleccionar el modelo.

El cuadro que se presenta a continuación expone resumida y esquemáticamente algunas formas de identificación para los órdenes autorregresivos y de medias móviles. Estas interpretaciones también son válidas para identificar los órdenes de la parte estacional del modelo.

Cuadro 1.1
Identificación para órdenes autorregresivos y de medias móviles

PROCESO	FAS	FACP
AR(p)	decaimiento exponencial	toma el valor 0 luego del rezago p
MA(q)	toma el valor 0 luego del rezago q	decaimiento exponencial
ARMA(p,q)	decaimiento exponencial	decaimiento exponencial

Identificación de Outliers

Outliers son observaciones incluidas dentro de la serie que no provienen del mismo proceso generador de datos que las demás, por lo que deben considerarse de forma diferenciada; incluyendo en el modelo variables específicas para cada uno de estos datos y así asignarlos de forma determinística.

En Hillmer et al. (1983) se consideran algunos tipos diferentes de outliers, entre los que destacamos: el outlier de tipo aditivo (Additive Outlier, AO), el cambio de nivel (Level Shift, LS) y el cambio temporal (Temporary Change, TC).

Un outlier aditivo (AO) es un suceso que afecta una serie en un solo instante temporal ($t = t_0$). Definimos el cambio de nivel (LS) como el suceso que afecta una serie en un periodo, con un efecto permanente sobre la misma. Y se define el cambio temporal (TC) como aquel suceso que tiene un impacto inicial sobre la serie, decayendo el efecto sobre la misma de forma exponencial.

Enfoque Outlier Aditivo para Estimación de Datos Faltantes

Gomez et al. (1993) plantearon esta metodología para la imputación de datos faltantes. La estrategia consiste en asignar valores de forma arbitraria para los datos faltantes y luego estimar los mismos como outliers de tipo AO, por ejemplo, a través del método de máxima verosimilitud. De esta forma el modelo finalmente adoptado será el que corrija el Outlier Aditivo mediante su coeficiente multiplicador y la estimación del dato faltante será la suma del valor asignado de forma arbitraria más la estimación del coeficiente AO.

Estimación

Dentro de esta etapa habrá que estimar los parámetros del/los modelo/s identificado/s en la etapa anterior. La estimación de dichos parámetros se podrá realizar a través de diferentes métodos de estimación. Los más aplicados, en general, son: el Método de Máxima Verosimilitud y Método de Mínimos Cuadrados. Una vez estimados los parámetros, éstos podrán ser testeados mediante pruebas de significación. A continuación se presentan características del Método de Máxima Verosimilitud que es el que fue escogido para estimar los parámetros del modelo planteado más adelante en el trabajo.

Método de Máxima Verosimilitud

El método consiste en considerar como estimador a aquel estadístico que maximice la función de verosimilitud. Éste estimador suele abreviarse de la forma *MLE* (maximum likelihood estimator). La función de verosimilitud se denota: $L(\theta | x) = L(x_1, x_2, \dots, x_n, \theta)$. Para calcular el máximo de la verosimilitud, es necesario determinar los valores para los cuales la derivada de la verosimilitud se anula. Por definición la verosimilitud es un producto de probabilidades o de densidades, por lo tanto para expresarla en forma aditiva, comenzamos por sustituir la verosimilitud por su logaritmo. Al ser el logaritmo una función creciente, es equivalente maximizar $\log[L(x_1, x_2, \dots, x_n, \theta)]$ o $L(x_1, x_2, \dots, x_n, \theta)$. Una vez determinado el valor de θ para el cual la derivada se anula, hay que asegurarse con la ayuda de la segunda derivada que el punto en cuestión es realmente un máximo.

Al trabajar con estimaciones basadas en el Método de Máxima Verosimilitud, es posible hacerlo en su versión condicionada o exacta, como se explica en Gras, J. A. (2001, p. 214)”:

“Al proceder con este método podemos condicionar la función de verosimilitud a los valores iniciales, o bien no hacerlo y llevar a cabo una estimación de la verosimilitud exacta. Cuando se aplica el principio de máxima verosimilitud con el enfoque no condicional se obtienen las funciones de verosimilitud exactas. Por su parte, cuando el principio se aplica con el enfoque condicional se obtienen funciones de verosimilitud condicionadas. Bajo el enfoque condicional, la maximización de la función de verosimilitud coincide con los estimadores mínimos cuadráticos.”

En caso que de la etapa de identificación surja más de un modelo posible, existen diferentes criterios de selección, como los son el criterio bayesiano y el criterio de Akaike, que basándose en predicciones dentro de la muestra, establecen métodos de elección. El criterio bayesiano propone elegir como el modelo que mejor ajusta, aquel que posee el menor indicador BIC: $BIC = \log\left(\frac{SCR}{n}\right) + [\log(n) + 2(1 + p + q)/n]$

El criterio propuesto por Akaike recomienda como el modelo que mejor ajusta, aquel que posee el menor indicador AIC:

$$AIC = \log\left(\frac{SCR}{n}\right) + 2(1 + p + q)/n$$

Siendo n el número de observaciones de la muestra de la serie temporal, p el parámetro de la parte autorregresiva del modelo y q el parámetro de la parte de medias móviles.

Diagnóstico

En la última etapa habrá que verificar los resultados del modelo elegido. Inicialmente se deberá verificar que los residuos del modelo siguen un proceso Ruido Blanco Gaussiano, para ello habrá que comprobar incorrelación, homoscedasticidad, media marginal igual a cero y normalidad. Existen diversos criterios y pruebas aplicables con la finalidad de efectuar dichas comprobaciones.

El estadístico Q de Box-Pierce es calculado como la suma ponderada de los cuadrados de una secuencia de autocorrelación.

$$Q_{BP} = n \sum_{k=1}^h \hat{\rho}_k^2$$

Donde n es el tamaño de la muestra, $\hat{\rho}_k$ es la autocorrelación de la muestra en el resago k , y h es el número de rezagos de la prueba.

Si se verifica que los errores del modelo provienen de un proceso ruido blanco, entonces el estadístico de Box-Pierce se distribuye aproximadamente como una distribución chi-cuadrado con $h - v$ grados de libertad, donde h es el número de rezagos utilizados en la estadística y v es el número de parámetros de ajuste. Este test es conocido como una prueba de maleta. Otra prueba de incorrelación es el acrónimo de prueba Ljung-Box, que es una versión del test de Box-Pierce. Al plantear la prueba de Box-Pierce, al igual que lo que sucede si se plantea la prueba de Ljung-Box, se reúne suficiente evidencia estadística para rechazar la hipótesis nula de incorrelación a un nivel de confianza del 95% (por ejemplo), si al calcular el p-valor correspondiente a la distribución Chi cuadrado establecida, éste no supera el valor 0,05.

Una posibilidad para verificar que las varianzas no siguen ninguna regularidad, es comprobar que el autocorrelograma de los residuos al cuadrado del/los modelo/s planteado/s no toma valores significativamente distintos de cero. De esta forma se estaría obteniendo cierta evidencia de homocedasticidad.

Para corroborar que la media de los residuos no sea significativamente mayor a cero en valor absoluto, se podrá corroborar que el intervalo de confianza al 95% (o al nivel de confianza que se desea asignar) contenga al valor cero. O lo que es lo mismo, que el

intervalo de confianza para la media posea un extremo de signo positivo y el otro negativo.

Para verificar que los residuos posean distribución normal es posible aplicar pruebas genéricas de normalidad, como los son: la prueba de Kolmogorov-Smirnov y la de Shapiro-Wilk.

La prueba de Kolmogorov-Smirnov es una prueba no paramétrica que es capaz de testear si dos muestras independientes corresponden a una misma distribución. Para aplicar esta prueba como prueba de normalidad simplemente se compara la serie que se quiere testear contra una surgida de la distribución Normal, de este modo se estaría aplicando la prueba de Kolmogorov-Smirnov para una muestra utilizando como distribución teórica a la distribución Normal. La prueba de K-S de una muestra es una dócima de bondad de ajuste. Esto es, se interesa en el grado de acuerdo entre la distribución de un conjunto de valores de la muestra y alguna distribución teórica específica. Determina si razonablemente puede pensarse que las mediciones muestrales provengan de una población que tenga esa distribución teórica. En la prueba se compara la distribución de frecuencia acumulativa de la distribución teórica con la distribución de frecuencia acumulativa observada. Se determina el punto en el que estas dos distribuciones muestran la mayor divergencia.

Hipótesis:

H₀: La distribución observada se ajusta a la distribución teórica.

$F(x) = F_i(x)$ para todo x .

H₁: La distribución observada no se ajusta a la distribución teórica.

El Test de Shapiro–Wilk plantea como hipótesis nula que una muestra x_1, \dots, x_n proviene de una población normalmente distribuida. Es considerado uno de los test más potentes para el contraste de normalidad, especialmente para muestras pequeñas.

El estadístico del test es:

$$W = \frac{\left(\sum_{i=1}^n a_i x_{(i)}\right)^2}{\left(\sum_{i=1}^n x_i - \bar{x}\right)^2}$$

donde

$x_{(i)}$ es el número que ocupa la i -ésima posición en la muestra;

\bar{x} es la media muestral;

las constantes a_i se calculan:

$$(a_1, \dots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}}$$

donde

$$m = (m_1, \dots, m_n)^T$$

siendo m_1, \dots, m_n los valores medios del estadístico ordenado, de variables aleatorias independiente muestradas de distribuciones normales y V es la matriz de covarianzas de ese estadístico de orden.

De esta forma la hipótesis nula se rechazará, a un nivel de confianza establecido por ejemplo en 95%, si el estadístico W es suficientemente pequeño para que el p-valor correspondiente sea menor a 0,05.

La etapa de diagnóstico incluye la evaluación de predicciones del modelo dentro de la muestra. Se deberá comprobar que las predicciones del modelo elegido sean satisfactorias, esta consideración dependerá en gran medida de lo que el analista considere como satisfactorio. A su vez, en caso de haber considerado más de un modelo en la fase de Identificación y de que éstos hayan sido diagnosticados de forma favorable, es posible elegir el modelo que mejor ajusta a la serie en base a las predicciones.

2. RESULTADOS

Las series diarias de llamadas entrantes en un Call Center son series cuya agregación temporal es el día, en el caso particular de los Call Center cuya operativa es de lunes a viernes; éstos son los únicos días a tomar en cuenta. Por lo tanto, después de cada viernes siempre hay un lunes, de forma ininterrumpida.

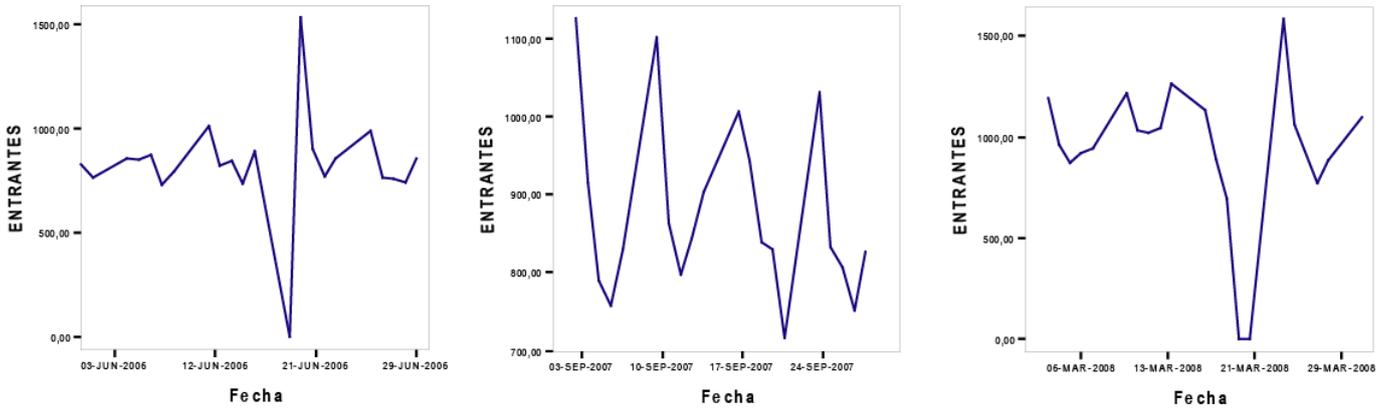
La modelización fue implementada a través del software **R**, programa abierto de uso libre. Para poder concretar la modelización hizo falta considerar las distintas estacionalidades presentes en la serie y el Efecto Calendario, a su vez fue necesario identificar y tratar outliers y datos faltantes en la serie. Para luego poder identificar y estimar un modelo adecuado para la serie diaria de llamadas entrantes al Call Center. Como último paso se realizaron pruebas de diagnóstico para analizar la bondad de ajuste del modelo y se corroboró el modelo como herramienta de predicción a corto plazo (a 5 pasos).

2.1 ESTACIONALIDAD

Las posibles estacionalidades presentes en la serie a tratar y dentro del período considerado son: la estacionalidad semanal (según el día de la semana que se trate), la estacionalidad mensual (según el momento del mes en el que se esté) y la estacionalidad anual (según el mes del año que se trate).

La experiencia empírica y el conocimiento de la persona responsable del sector Hola Itaú: Diego Castillo, sobre la evolución de llamadas indica que dentro de las posibles estacionalidades la más fuerte es la semanal, luego la anual y por último la estacionalidad mensual. Esta última de existir sería extremadamente débil, debajo se muestra, a modo de ejemplo, la evolución de llamadas dentro de tres meses tomados al azar de la muestra:

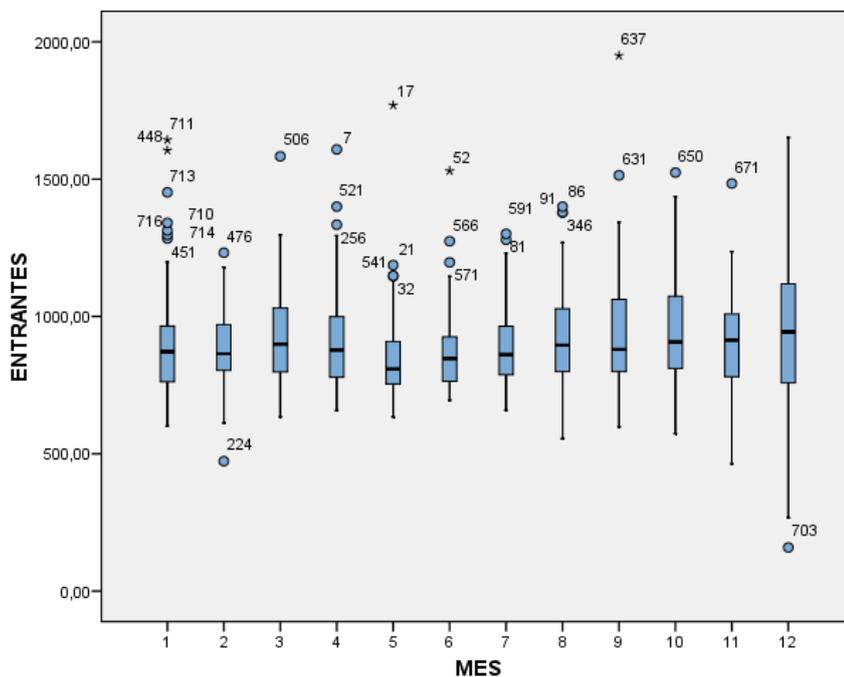
Gráfico 2.1
Llamadas entrantes Junio 2006, Setiembre 2007 y Marzo 2008



Observando los gráficos se distingue claramente una estacionalidad semanal y reforzamos la idea inicial de la inexistencia de una fuerte estacionalidad mensual, por lo que esta última no fue considerada para la modelización.

Para obtener información respecto a una posible estacionalidad anual efectuamos gráficos de caja para cada uno de los meses tomando los datos de toda la muestra a lo largo del período abril 2006 – abril 2009.

Gráfico 2.2
Diagramas de caja mensuales



Gráficos agrupados por mes desde Enero (mes 1) a Diciembre (mes 12)

Como información adicional tenemos que el mes que presentó menor promedio de llamadas entrantes fue mayo, con un promedio de 861 llamadas y el que presentó el mayor promedio de llamadas fue octubre, con un promedio de 960 llamadas. Por lo que el rango de los diferentes meses es menor a 100 llamadas. Toda esta información nos da indicios de que la estacionalidad anual sería, en todo caso, de baja magnitud.

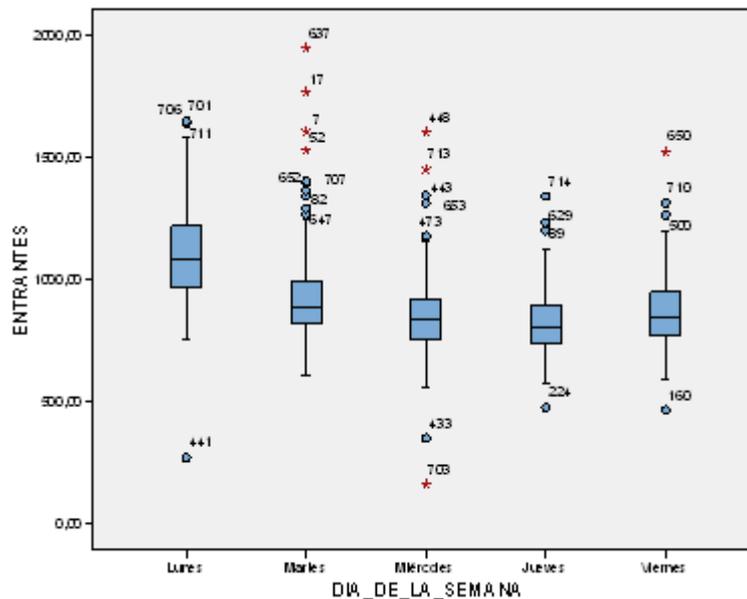
ANOVA de un factor

Esta técnica fue aplicada exclusivamente para el análisis de la estacionalidad semanal. No fue aplicada sobre las otras posibles estacionalidades debido a que obtuvimos ciertos indicios de su debilidad o inexistencia, por otro lado la experiencia empírica destaca una fuerte presencia de estacionalidad semanal dentro de la serie de datos que se analiza.

Como ya fue expuesto: la técnica ANOVA tiene por objetivo analizar si las diferentes medias que adquieren las categorías de una variable, según los valores de otra, presentan diferencias significativas entre sí. La hipótesis nula de esta prueba establece que todas las medias de las categorías coinciden y que de existir diferencias, éstas se deben a errores de tipo aleatorio. A su vez, la técnica permite comparar las medias de las categorías dos a dos, es decir una categoría contra otra. En caso de que existan diferencias entre alguna o algunas de las categorías esta utilidad permite identificar dónde se producen esa o esas diferencias.

Se analizó si los promedios que toma la serie para cada uno de los días de la semana presentan diferencias significativas, para así poder identificar una posible estacionalidad semanal. Como primer paso, previo a la prueba ANOVA, se efectuó un análisis exploratorio para identificar la existencia de datos alejados de la media dentro de cada día, que puedan distorsionar el valor de la misma.

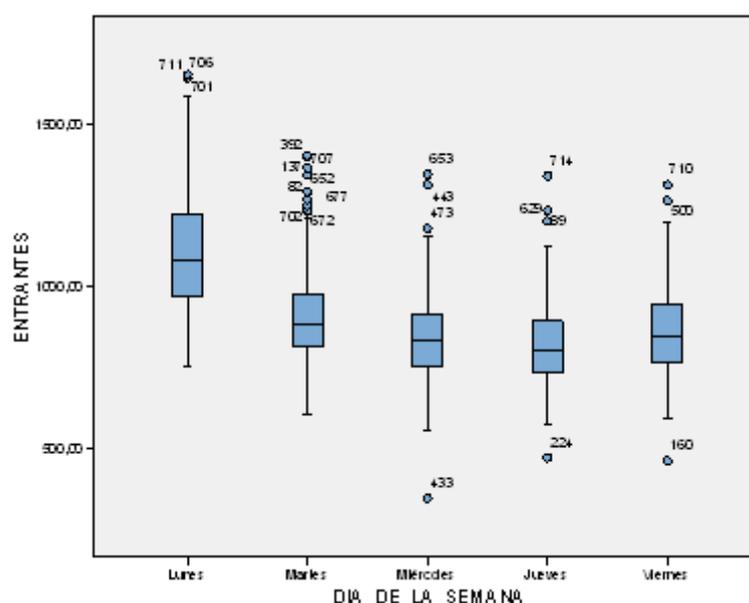
Gráfico 2.3
Diagramas de caja diarios



Se puede observar que entre los días lunes hubo una observación que se alejó claramente del resto, esta observación es la única que se aleja a más de tres desvíos estándar de la media y corresponde a un 24 de diciembre; día en el que el Call Center no opera en horario completo. Dentro de los días martes hubo 4 observaciones que distan a más de 3 longitudes de caja de la tercera bisagra de Tukey. De esas cuatro observaciones; tres corresponden a días siguientes a feriados. Los días miércoles hubo 2 observaciones a más de 3 longitudes de caja de la tercera bisagra de Tukey y los dos casos corresponden a días posteriores a feriados, mientras que hubo una observación a más de 3 longitudes de caja de la primera bisagra de Tukey, la cual corresponde a un 24 de diciembre. En total se identificaron 9 posibles observaciones alejadas de sus respectivas medias; lo cual significa un elevado porcentaje del total como para ser considerado atípico, pero de estas 9 observaciones hay 2 que corresponden a días en los que la operativa no fue completa (24 de diciembre) y 5 en los que si bien fueron martes o miércoles, corresponden a días posteriores a períodos sin operativa; hecho característico de los días lunes. Por lo que pese a que es un porcentaje considerable, 9 datos se excluyeron del análisis, el cual se realizó con 724 de las 733 observaciones disponibles.

Así quedaron los gráficos de caja, sin considerar las 9 observaciones excluidas:

Gráfico 2.4
Diagramas de caja diarios sin observaciones alejadas



Cuadro 2.1
Estadísticos descriptivos para prueba ANOVA

ENTRANTES	N	Media	Desv. Típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
LUNES	140	1101,21	183,16	15,48	1070,60	1131,82	752,00	1651,00
MARTES	143	915,19	153,04	12,80	889,89	940,49	606,00	1403,00
MIÉRCOLES	149	847,91	140,13	11,48	825,23	870,60	348,00	1345,00
JUEVES	146	820,44	131,33	10,87	798,96	841,92	473,00	1341,00
VIERNES	146	861,75	144,24	11,94	838,16	885,35	463,00	1313,00
Total	724	907,43	180,73	6,72	894,25	920,62	348,00	1651,00

Observando los principales estadísticos descriptivos para estos 724 datos se puede distinguir que el día que posee más datos relevados es el miércoles y el que tiene menos es el lunes, día que presenta la mayor desviación. El día que posee mayor promedio de llamadas entrantes es el lunes seguido del martes, viernes, miércoles y por último el jueves; que es el que posee el menor promedio de llamadas entrantes. Se puede observar que los días con mayor volumen de llamadas son los que mayor desviación presentan, en el mismo orden: lunes, martes, viernes, miércoles y por último jueves. Se distingue que el mínimo se alcanzó un día miércoles y no un jueves, esto podría ser un indicio de que esa observación debió ser considerada como atípica, pero la primera bisagra de Tukey para los días miércoles es menor que para los días jueves: por lo tanto el hecho de que el mínimo se haya alcanzado un día miércoles se explica por el tipo de

distribución que posee este día (con mayor variabilidad y “colas” más pesadas) y no porque sea una observación atípica. Por otro lado el máximo corresponde a un día lunes, hecho que a priori era esperable.

Cuadro 2.2
Prueba de homogeneidad de varianzas para prueba ANOVA

ENTRANTES			
Estadístico de Levene	gl1	gl2	Sig.
4,403	4	719	,002

Observando el nivel de significación de la prueba de Levene podemos rechazar, con un nivel de confianza del 95%, la hipótesis que establece que las varianzas para los distintos días de la semana son homogéneas. Esto se debe a que el nivel de significación es menor a 0,05. La conclusión se corresponde con la observación inicial de que los días con mayor volumen de llamadas entrantes presentan mayor desviación.

Dado que se rechazó homogeneidad de varianzas, como ya fue explicado dentro del Marco Teórico, una forma de comparar medias es a través de la prueba Post- hoc de Games-Howell.

Cuadro 2.3
Prueba de comparación de medias Post-hoc

(I) DÍA DE LA SEMANA	(J) DÍA DE LA SEMANA	Diferencia de medias (I-J)	Error típico	Sig.	Intervalo de confianza al 95%	
					Límite superior	Límite inferior
Lunes	Martes	186,03(*)	20,09	,000	130,87	241,18
	Miércoles	253,30 (*)	19,27	,000	200,36	306,24
	Jueves	280,78(*)	18,91	,000	228,80	332,75
	Viernes	239,46 (*)	19,55	,000	185,77	293,16
Martes	Lunes	-186,03 (*)	20,09	,000	-241,18	-130,87
	Miércoles	67,28(*)	17,19	,001	20,08	114,47
	Jueves	94,75 (*)	16,79	,000	48,65	140,85
	Viernes	53,44(*)	17,50	,021	5,39	101,48
Miércoles	Lunes	-253,30 (*)	19,27	,000	-306,24	-200,36
	Martes	-67,28(*)	17,19	,001	-114,47	-20,08
	Jueves	27,47	15,81	,412	-15,92	70,87
	Viernes	-13,84	16,56	,919	-59,30	31,62
Jueves	Lunes	-280,78(*)	18,91	,000	-332,75	-228,80
	Martes	-94,75 (*)	16,79	,000	-140,85	-48,65
	Miércoles	-27,47	15,81	,412	-70,87	15,92
	Viernes	-41,32	16,14	,081	-85,63	3,00
Viernes	Lunes	-239,46(*)	19,55	,000	-293,16	-185,77
	Martes	-53,44(*)	17,50	,021	-101,48	-5,39
	Miércoles	13,84	16,56	,919	-31,62	59,30
	Jueves	41,32	16,14	,081	-3,00	85,63

(*) La diferencia de medias es significativa al nivel .05.

Esta prueba muestra que la diferencia de medias entre el lunes y cualquier otro día es significativa para un nivel de confianza del 95%, lo mismo sucede con el martes. Mientras que las diferencias dos a dos de medias, que existen entre los días miércoles, jueves y viernes no son significativas. Por lo tanto existen diferencias estadísticamente significativas entre las medias de los distintos días de la semana.

Tratamiento de estacionalidades simultáneas

Se siguió la recomendación expuesta en Espasa et al. (1996) respecto a la modelización de distintas estacionalidades simultáneas. Con todos los datos obtenidos y sumados a la información otorgada por Diego Castillo, seleccionamos a la estacionalidad semanal

como la principal estacionalidad presente en la serie y seleccionamos una única posible estacionalidad secundaria: la estacionalidad anual, no considerando la estacionalidad mensual.

De esta forma, y siguiendo el esquema propuesto en Espasa et al. (1996) tratamos a la estacionalidad principal con un esquema estocástico y a la estacionalidad secundaria con un esquema determinístico.

Las variables indicatrices del esquema determinístico, elegido para tratar a la estacionalidad anual, quedaron definidas de la siguiente manera:

$$Enero = En_t = \begin{cases} 1 & \forall t=\text{día dentro del mes de enero} \\ 0 & \text{otro caso} \end{cases} \quad Febrero = Fb_t = \begin{cases} 1 & \forall t=\text{día dentro del mes de febrero} \\ 0 & \text{otro caso} \end{cases}$$

$$Marzo = Mz_t = \begin{cases} 1 & \forall t=\text{día dentro del mes de marzo} \\ 0 & \text{otro caso} \end{cases} \quad Abril = Ab_t = \begin{cases} 1 & \forall t=\text{día dentro del mes de abril} \\ 0 & \text{otro caso} \end{cases}$$

$$Mayo = My_t = \begin{cases} 1 & \forall t=\text{día dentro del mes de mayo} \\ 0 & \text{otro caso} \end{cases} \quad Junio = Jn_t = \begin{cases} 1 & \forall t=\text{día dentro del mes de junio} \\ 0 & \text{otro caso} \end{cases}$$

$$Julio = Jl_t = \begin{cases} 1 & \forall t=\text{día dentro del mes de julio} \\ 0 & \text{otro caso} \end{cases} \quad Agosto = Ag_t = \begin{cases} 1 & \forall t=\text{día dentro del mes de agosto} \\ 0 & \text{otro caso} \end{cases}$$

$$Setiembre = St_t = \begin{cases} 1 & \forall t=\text{día dentro del mes de setiembre} \\ 0 & \text{otro caso} \end{cases} \quad Octubre = Oc_t = \begin{cases} 1 & \forall t=\text{día dentro del mes de octubre} \\ 0 & \text{otro caso} \end{cases}$$

$$Noviembre = Nv_t = \begin{cases} 1 & \forall t=\text{día dentro del mes de noviembre} \\ 0 & \text{otro caso} \end{cases}$$

2.2 EFECTO CALENDARIO

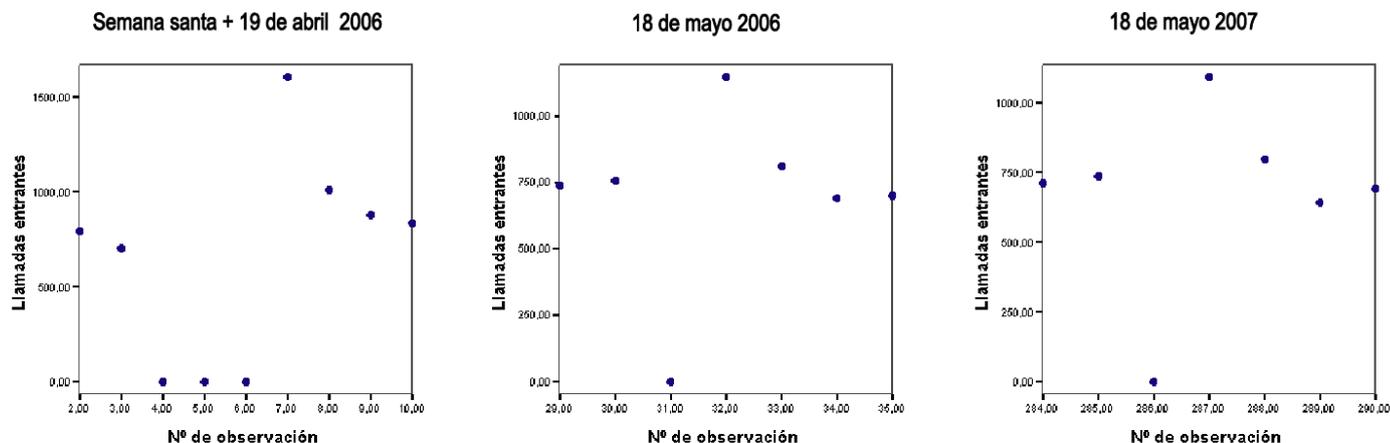
El conocimiento empírico del analista es fundamental para identificar los efectos del calendario sobre la serie que hace falta destacar y contemplar.

Para el caso de las series de llamadas entrantes a un call center es habitual que a mayor número de días sin servicio, se presente mayor cantidad de llamadas entrantes en el día de retorno al servicio. Este fenómeno hace que se produzcan ciertas estacionalidades, por ejemplo, que los días lunes sean generalmente los que más llamadas reciben, en comparación al resto de los días hábiles. Y, del mismo modo, cuando se produce un feriado, esto repercute en el nivel de llamadas entrantes al siguiente día hábil posterior. A su vez, no todos los feriados producen los mismos efectos; debido a que en cada calendario existen distintos tipos de feriado.

Para el caso particular de la serie a tratar, todos los feriados (laborables y no laborables) tienen el mismo efecto sobre el día en que ocurren: el call center no tiene operativa y por lo tanto no se reciben llamadas. Teniendo en cuenta la dificultad de caracterizar particularidades de cada feriado debido a la pequeña muestra disponible (se cuenta con un máximo de 3 observaciones, una por año, de cada feriado), y en base a experiencia empírica dentro del funcionamiento del Call Center, con el aval de Diego Castillo (responsable del área Hola Itaú): realizamos el supuesto de que los efectos retrospectivos y prospectivos de todos los feriados son iguales, sin importar de que fiesta o período vacacional se trate, ni si es móvil o no lo es, ni tampoco distinguimos el momento del mes o el día de la semana en que éste ocurre, con la salvedad de que cuando existe un período de más de un día feriado contiguo el efecto prospectivo es de mayor magnitud.

Para contrastar las hipótesis de nuestro supuesto y obtener herramientas que ayuden a interpretar el efecto calendario dentro de la serie de llamadas, tomamos tres feriados de ejemplo elegidos al azar y observamos que ocurrió en los días previos y posteriores a estos.

Gráfico 2.5 Efectos prospectivos y retrospectivos de los feriados



Las observaciones nº 4 y 5 corresponden a semana santa 2006, la observación nº 6 corresponde al feriado del 19 de abril de 2006 corrido al 17 de abril. La observación 31 corresponde al feriado del 18 de mayo de 2006 corrido al 22 de mayo. La observación nº 286 corresponde al feriado de 18 de mayo de 2007 corrido al 21 de mayo

Observando los gráficos vemos que luego de cada feriado se produce un fuerte impacto en el nivel de llamadas entrantes el cual disminuye en el día siguiente para prácticamente extinguirse al tercer día posterior al feriado. También se observa que, de existir, el efecto que el feriado tiene sobre el día previo es de baja magnitud y es nulo en los días anteriores a éste. Por otro lado puede verificarse que cuando existe más de un día consecutivo sin operativa, el efecto prospectivo es aún mayor.

Todo esto apoya las hipótesis que sustentan nuestros supuestos, por lo tanto definimos el efecto calendario para la serie a tratar del siguiente modo:

Feriado: se definió como feriado todo día, distinto de los sábados y domingos, en que el call center no tenga operativa. De esta forma si un día es feriado la cantidad de llamadas entrantes será igual a cero. Esta definición se basa en que el call center no presenta operativa en ningún día feriado, sea éste laborable o no laborable y sin importar la fecha en que tenga lugar, ya sea que haya sido movido o no. Una particularidad de la serie es que todos los días en que el número de llamadas entrantes fue cero, se corresponden con días feriados. Por lo tanto un día es feriado, si y solo si, el número de llamadas entrantes es cero.

Efecto feriado 0: este efecto refleja la repercusión que tiene un feriado o un período de días feriados, en la operativa del día hábil que antecede a este suceso.

Efecto feriado 1: se define como el efecto que un feriado simple tiene sobre la operativa del primer día hábil posterior a éste.

Efecto feriado 2: efecto que el feriado tiene sobre la operativa del segundo día posterior a éste.

Doble efecto feriado 1: este doble efecto refiere a la repercusión que un período de más de un día feriado tiene sobre el siguiente día con operativa.

Doble efecto feriado 2: efecto que provoca un período de más de un día feriado sobre la operativa del segundo día posterior al período.

Veinticuatro de diciembre: todo día veinticuatro de diciembre que tenga operativa, en estos días la operativa parcial.

Se crearon variables *dummy* que indicaran para cada día en particular si se está en presencia de algún efecto feriado de los definidos, de ese modo se produjeron las variables: FERIADO, EFECTO FERIADO 0, EFECTO FERIADO 1, EFECTO FERIADO 2, DOBLE EFECTO FERIADO 1 y DOBLE EFECTO FERIADO 2 y VEINTICUATRO DE DICIEMBRE. Debajo se exponen las variables de forma analítica con su universo de valores posibles:

$$Feriados = FE_t = \begin{cases} 1 & \forall t = \text{día feriado} \\ 0 & \text{otro caso} \end{cases}$$

$$\text{Efecto Feriado 0} = EF0_t = \begin{cases} 1 & \forall t = \text{día previo a un feriado o período de feriados} \\ 0 & \text{otro caso} \end{cases}$$

$$\text{Efecto Feriado 1} = EF1_t = \begin{cases} 1 & \forall t = \text{día posterior a un feriado} \\ 0 & \text{otro caso} \end{cases}$$

$$\text{Efecto Feriado 2} = EF2_t = \begin{cases} 1 & \forall t = \text{segundo día posterior a un feriado} \\ 0 & \text{otro caso} \end{cases}$$

$$\text{Doble Efecto Feriado 1} = DEF1_t = \begin{cases} 1 & \forall t = \text{día posterior a un período de feriados} \\ 0 & \text{otro caso} \end{cases}$$

$$\text{Doble Efecto Feriado 2} = DEF2_t = \begin{cases} 1 & \forall t = \text{segundo día posterior a un período de feriados} \\ 0 & \text{otro caso} \end{cases}$$

$$24 \text{ de Diciembre} = NB_t = \begin{cases} 1 & \forall t = \text{día 24 de diciembre} \\ 0 & \text{otro caso} \end{cases}$$

Y la expresión analítica del efecto calendario definido es de la forma:

$$w_{-1}EF0_t + w_0FE_t + w_1EF1_t + w_2EF2_t + w_3DEF1_t + w_4DEF2_t + w_5NB_t$$

Donde w_i representa los coeficientes del efecto calendario definido

2.3 TRATAMIENTO PARA DATOS FALTANTES

A los efectos de poder modelizar la serie completa con las 778 observaciones, fue necesario asignarle un valor numérico a cada dato faltante. En total hubo **5 datos faltantes** los cuales se encontraban a lo largo de la serie, el primer dato faltante se registró en la observación número 14 y el último en la observación número 508, dentro de un orden ascendente por fecha.

Posibles soluciones

Un método para estimar estos datos consistiría en tomar la serie desde el principio hasta el primer dato faltante y para esta serie partida plantear un modelo y aplicar una predicción hacia adelante para asignarle así un dato estimado al primer dato faltante, luego continuar hasta el siguiente dato faltante, volver a plantear un modelo para la nueva serie partida y volver a predecir, en este caso, el segundo dato faltante; siguiendo esa lógica hasta el quinto y último dato sin valor asignado. Otra posibilidad sería aplicar el mismo razonamiento pero partiendo la serie desde el primer día posterior al último dato faltante hasta el final y aplicar la misma metodología pero con predicciones hacia atrás. Estas metodologías se hicieron de muy difícil aplicación ya que el último dato faltante se encuentra cercano al final de la serie y el primero se encuentra extremadamente cercano al dato inicial, lo que provoca que las modelizaciones sean muy erráticas y aún más las predicciones.

Estrategia elegida

La estrategia utilizada se denomina “Enfoque Outlier Aditivo para Estimación de Datos Faltantes” y está planteada en Gomez et al. (1993). Como fue explicado en el marco metodológico, esta estrategia consiste en asignar de forma arbitraria los valores para los datos faltantes y luego estimar los mismos como outliers de tipo AO, lo que será realizado a través del método de máxima verosimilitud. La suma del valor arbitrario más la estimación del coeficiente AO, será entonces la estimación del dato faltante.

Una posibilidad para plantear los puntos de partida arbitrarios de cada dato faltante es asignarles a todos la media de la muestra total. Si bien esta estimación es habitual en modelizaciones diarias por su fácil aplicación, no sería muy conveniente debido a que: como ya fue expuesto, existen estacionalidades y efecto calendario que pueden afectar sensiblemente cada observación. Tomando en cuenta las particularidades de la serie a tratar se calcularon los promedios de forma individual para, los días lunes, martes, miércoles, jueves o viernes en que no existan distorsiones por efecto calendario y, por otro lado se consideraron los promedios de los días en que se presentó algún efecto

calendario con el conocimiento previo de que en ninguno de los casos el dato faltante correspondía a un veinticuatro de diciembre. Considerando así la estacionalidad principal y el efecto calendario definidos.

De esta forma se obtuvieron las siguientes medias:

Cuadro 2.4
Promedio de llamadas entrantes por tipo de día

DIAS Y EFECTOS FERIADO	N	Media
Lunes	127	1088,81
Martes	130	908,39
Miércoles	128	835,99
Jueves	133	811,68
Viernes	129	862,45
Efecto feriado 0	25	804,60
Efecto feriado 1	23	1136,39
Efecto feriado 2	23	962,70
Doble efecto feriado 1	7	1388,14
Doble efecto feriado 2	8	1018,37

Siendo N el número de casos disponibles para cada categoría

En el cuadro expuesto se observan, a simple vista y sin aplicación de métodos comparativos, las diferencias existentes en los promedios entre los distintos días de la semana y entre los distintos días que han sufrido alguna distorsión por efecto calendario.

La estrategia elegida le asigna a cada dato faltante el valor del promedio del día de la semana que a este le corresponda, y si el dato faltante se ve afectado por alguno de los efectos de calendario ya definidos, se le aplica el valor promedio para cada efecto.

Aplicando redondeo para asignar los valores libres de decimales, se obtuvieron las siguientes estimaciones:

Cuadro 2.5
Estimaciones iniciales para datos faltantes

Datos Faltantes	Día de la semana	Efecto Feriado	Estimación
27-APR-2006	Jueves	Ninguno	812
28-APR-2006	Viernes	Efecto feriado 0	805
21-FEB-2007	Miércoles	Doble Efecto Feriado 1	1388
07-DEC-2007	Viernes	Ninguno	862
26-MAR-2008	Miércoles	Ninguno	836

Una vez obtenido un modelo definitivo, tendremos nuevas estimaciones para estos datos faltantes (que serán tratados como outliers AO), y podremos verificar que tan buenas

resultaron nuestras estimaciones iniciales respecto a las estimaciones obtenidas finalmente observando cuan pequeños resultan los valores de los coeficientes.

Las expresiones analíticas de las 5 variables indicatrices de datos faltantes son de la forma:

$$FL1_t = \begin{cases} 1 & \forall t = \text{primer dato faltante} \\ 0 & \text{otro caso} \end{cases} \quad FL2_t = \begin{cases} 1 & \forall t = \text{segundo dato faltante} \\ 0 & \text{otro caso} \end{cases} \quad FL3_t = \begin{cases} 1 & \forall t = \text{tercero dato faltante} \\ 0 & \text{otro caso} \end{cases}$$

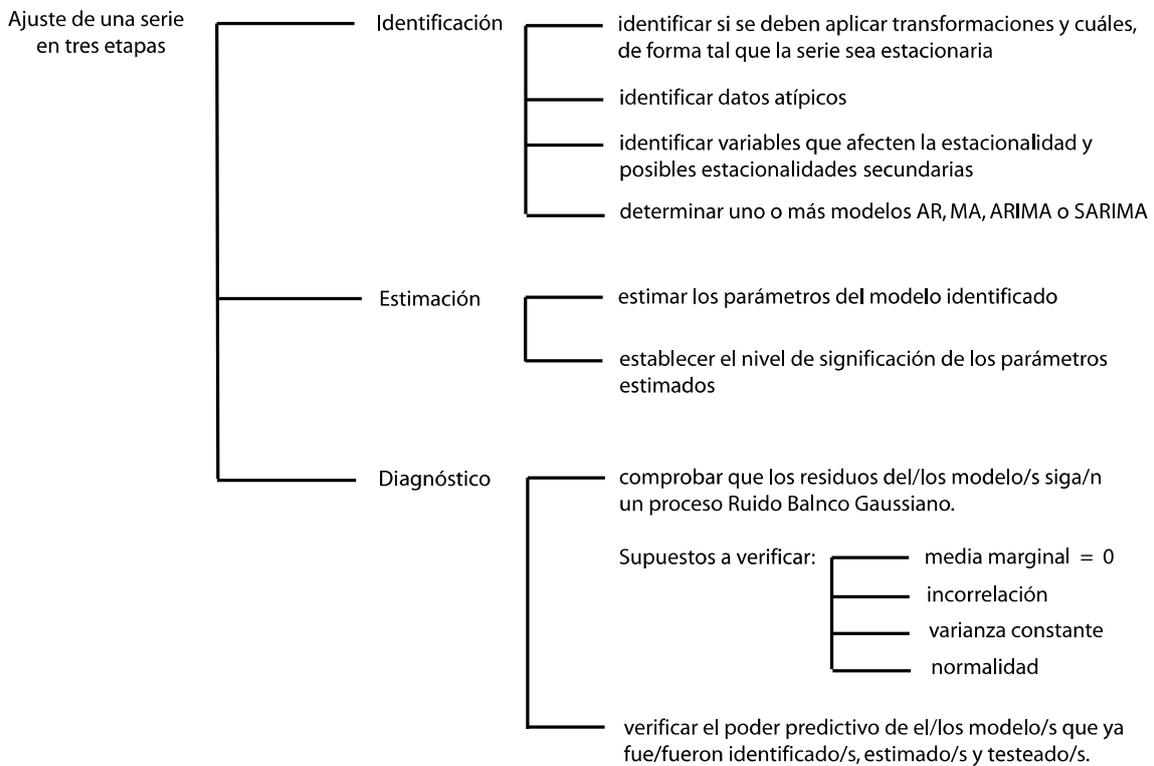
$$FL4_t = \begin{cases} 1 & \forall t = \text{cuarto dato faltante} \\ 0 & \text{otro caso} \end{cases} \quad FL5_t = \begin{cases} 1 & \forall t = \text{quinto dato faltante} \\ 0 & \text{otro caso} \end{cases}$$

2.4 AJUSTE DE LA SERIE A UN MODELO SARIMA (p, d, q) (P, D, Q)s

El ajuste fue planteado en tres etapas siguiendo la metodología propuesta por Box y Jenkins (1976), tal como muestra el esquema, suponiendo que los residuos ε_t del modelo siguen un proceso Ruido Blanco Gaussiano con media igual a cero y varianza constante, y suponiendo incorrelación entre ellos.

Figura 2.1

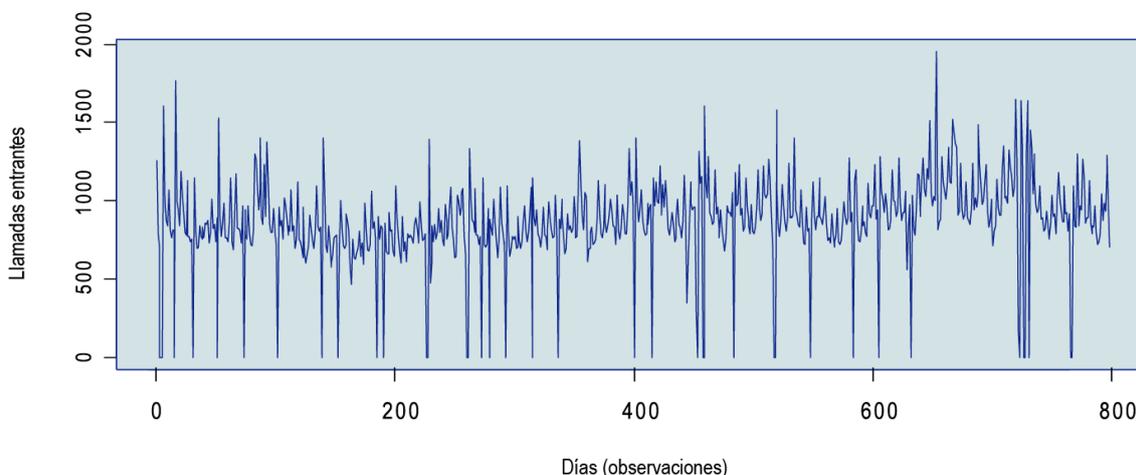
Esquema para la metodología Box y Jenkins



Identificación

Inicialmente se efectuó un análisis descriptivo de la serie de llamadas entrantes. En el gráfico de la serie se muestra la evolución de las llamadas entrantes, en volumen, desde el 10 de abril de 2006 al 8 de abril de 2009, con un total de 778 observaciones, de las cuales cinco correspondían a datos faltantes a los que le fueron asignados valores según las estimaciones de puntos de partida explicadas anteriormente. Se observan 40 datos a lo largo de la serie, los cuales toman el valor cero, estos datos corresponden a días feriados (como ya fue explicado anteriormente).

Gráfico 2.6
Serie diaria de llamadas entrantes al Call Center Hola Itaú
desde el 10 de abril de 2006 hasta el 8 de abril de 2009



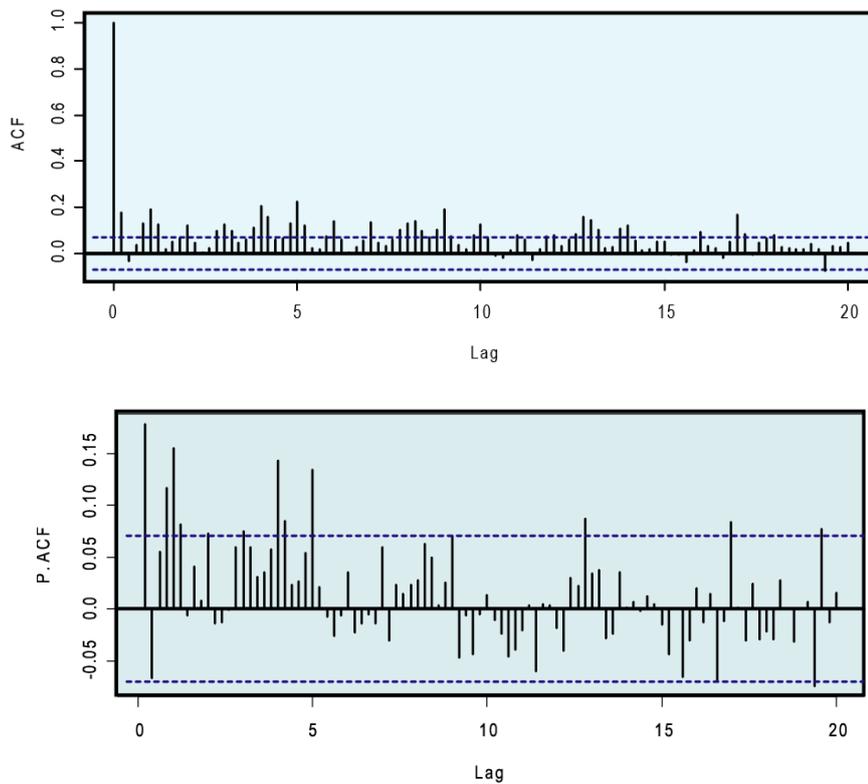
Para exponer los datos descriptivos de esta serie que aparecen a continuación, fueron excluidas las 40 observaciones correspondientes a días feriados, debido a que las mismas no corresponden a días en los que el call center haya tenido operativa.

Cuadro 2.6
Datos descriptivos de la serie de llamadas entrantes

LLAMADAS ENTRANTES		
N	Válidos	738
	Perdidos	0
Media		912,67
Desv. típ.		197,39
Varianza		38934,68
Mínimo		159,00
Máximo		1950,00
Percentiles	5	676,85
	25	780,00
	50	879,50
	75	1000,25
	95	1294,15

A través del análisis descriptivo se observa que, sin considerar los días feriados y aplicando redondeo para no considerar ningún decimal, la serie posee una media de 913 llamadas entrantes, una mediana de 879 llamadas y una varianza de 38.935 correspondiente a una desviación típica de 197. El máximo de llamadas entrantes fue de 1.950 y se registró el día 23 de setiembre de 2008 y el mínimo correspondió al 24 de diciembre de 2008, día en que la operativa es parcial y donde se registraron 159 llamadas entrantes.

Gráfico 2.7
Autocorrelograma y Autocorrelograma Parcial de la serie de datos

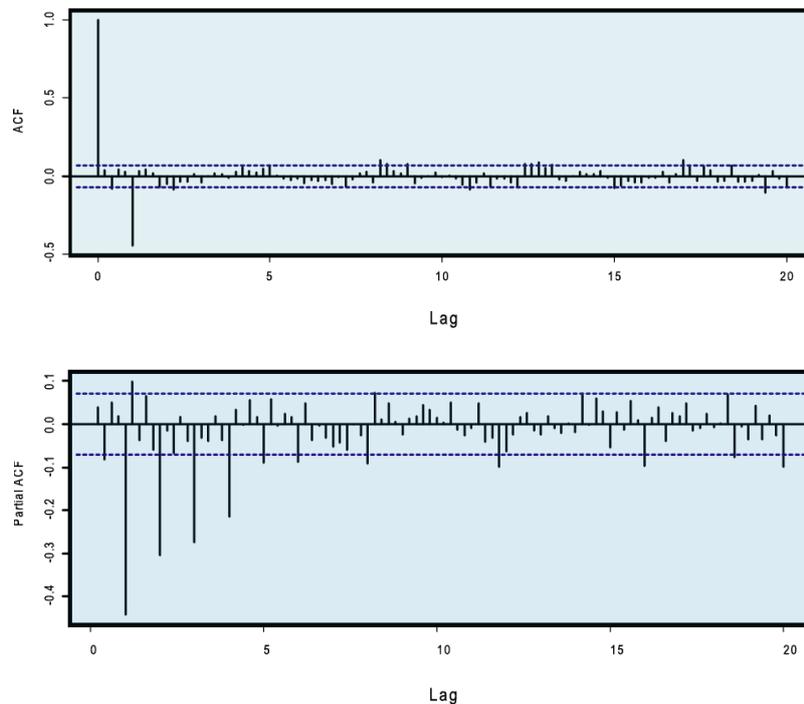


A partir de los gráficos del autocorrelograma y el autocorrelograma parcial se obtuvo la siguiente interpretación:

El gráfico del autocorrelograma posee decaimiento exponencial mientras que el gráfico del autocorrelograma parcial toma los primeros valores cercanos a cero luego del primer rezago. Estas características se podrían interpretar como correspondientes a un modelo AR(1).

Para obtener más herramientas de interpretación, se efectuaron los mismos gráficos para la primera diferencia estacional.

Gráfico 2.8
Autocorrelograma y Autocorrelograma Parcial de la primera diferencia estacional



Observando estos dos últimos gráficos se obtuvo la siguiente interpretación:

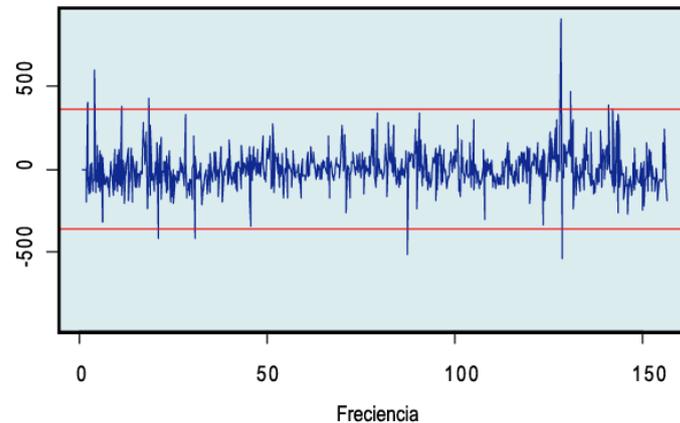
El gráfico del autocorrelograma toma valores cercanos a cero luego del primer rezago estacional y el gráfico del autocorrelograma parcial posee un claro decaimiento exponencial en la parte estacional, características que podrían interpretarse como correspondientes a un modelo MA(1) para la primera diferencia estacional.

De esta forma, a partir de las dos interpretaciones mencionadas y sin considerar los rezagos determinísticos, se identificó el siguiente modelo: $SARIMA (1\ 0\ 0)(0\ 1\ 1)_5$.

Identificación de Outliers

Si se observa el gráfico de la serie de tiempo original (presentado en la página 39) podemos observar que no se distingue ningún cambio de nivel evidente, por lo que resulta posible descartar la existencia de outliers de tipo LS (cambio de nivel). Para analizar la existencia de otro tipo de outliers, se realizó el ejercicio de verificar cuantos datos dentro de la serie de residuos del modelo identificado eran mayores, en valor absoluto, a tres veces el desvío estándar del modelo. Se obtuvieron así doce posibles outliers:

Gráfico 2.9
Residuos del modelo identificado



De esos doce posibles outliers tenemos que en nueve casos los residuos del modelo presentaron signo opuesto en t y en $t+1$, analizando uno a uno los casos se concluyó que estos nueve outliers corresponde, a nuestro criterio, clasificarlos como outliers de tipo AO. Luego tenemos dos casos en los que, si bien los residuos del modelo presentaron el mismo signo en t y en $t+1$, el residuo en $t+1$ era muy cercano a cero, por lo que estos otros dos casos también fueron clasificados como outliers de tipo AO. Por último observamos un posible outlier que sí podría clasificarse como de tipo TC tal que:

$$TC_t(x) = \begin{cases} 1 & \text{si } t=i \\ x & \text{si } t=i+1 \\ x^2 & \text{si } t=i+2 \\ x^3 & \text{si } t=i+3 \\ \vdots & \\ 0 & \forall t < i \end{cases}$$

donde i corresponde al momento en que se registra el outlier. Se analizó cual sería la propagación de este posible outlier comprobando si el modelo mejora, teniendo en cuenta el criterio de mínimo AIC e incluyendo en cada caso: un outlier TC($x=0,2$), lo que implica un rápido decaimiento y por lo tanto que el efecto se diluye rápidamente, un TC($x=0,5$) o considerándolo como un outlier AO. Como resultado se obtuvo que la mejor especificación para dicho outlier sería de tipo TC(0,2). *⁴

Estimación

Para estimar los parámetros del modelo se eligió aplicar el método de máxima verosimilitud. Todas las variables *dummy*, que fueron creadas según se expuso anteriormente dentro de éste capítulo, fueron consideradas para determinar el modelo completo. Se comprobó mediante pruebas empíricas que los coeficientes resultantes de incluir las *dummy* que determinaban la estacionalidad secundaria no eran significativos

*⁴ Ver páginas 66 a 69 en ANEXO

(asumiendo un nivel de significación de 0,05). Lo mismo sucedió con la variable indicatriz de Efecto Feriado 0 ($EF0_t$) y con uno de los outliers identificados $*^5$. Por lo que ninguna de estas variables fue incluida, simplificando sensiblemente la modelización al descartar 14 variables *dummy*. De este modo fueron considerados, además de los parámetros del modelo identificado, los datos faltantes, once outliers y todas las variables *dummy* de efecto calendario excepto $EF0_t$.

La expresión analítica del modelo elegido es de la forma:

$$\begin{aligned}
 Y_t = & 0,58*Y_{t-1} + Y_{t-5} - 0,58*Y_{t-6} - 0,93*\varepsilon_{t-5} + \varepsilon_t \dots \\
 & - 953,25*FE_t + 215,93*EF1_t + 57,01*EF2_t + 358,61*DEF1_t + 123,68*DEF2_t, \dots \\
 & - 874,42*NB + 71,62*FL1_t + 44,58*FL2_t + 371*FL3_t - 33,94*FL4_t - 1,25*FL5_t, \dots \\
 & + 395,71*AO1_t + 676,19*AO2_t + 474,64*AO3_t + 410,64*AO4_t - 313,87*AO5_t, \dots \\
 & - 344,68*AO6_t - 448,38*AO7_t + 929,97*AO8_t + 375*AO9_t + 293,54*AO10_t + 99,51*TC(0,2)
 \end{aligned}$$

Cuadro 2.7
Coefficientes del modelo identificado

	Coefficiente	Desvío estándar	Estadístico t
AR 1	0,58	0,03	18,20
S. MA 1	-0,93	0,02	-52,53
FEt (Feriado)	-953,25	17,89	-53,29
NBt (Noche buena)	-874,42	61,31	-14,26
EF1t (Efecto feriado 1)	215,93	23,91	9,03
EF2t (Efecto feriado 2)	57,01	20,78	2,74
DEF1t (Doble efecto feriado 1)	358,61	39,70	9,03
DEF2t (Doble efecto feriado 2)	123,68	34,79	3,55
FL1t (dato faltante 1)	71,62	96,58	0,74
FL2t (dato faltante 2)	44,58	96,75	0,46
FL3t (dato faltante 3)	371,00	92,69	4,00
FL4t (dato faltante 4)	-33,94	122,31	-0,28
FL5t (dato faltante 5)	-1,25	86,97	-0,01
AO1 (Outlier 1)	395,71	94,11	4,20
AO2 (Outlier 2)	676,19	88,82	7,61
AO3 (Outlier 3)	474,64	88,14	5,38
AO4 (Outlier 4)	410,64	85,86	4,78
AO5 (Outlier 5)	-313,87	88,24	-3,56
AO6 (Outlier 6)	-344,68	88,11	-3,91
AO7 (Outlier 7)	-448,38	85,87	-5,22
AO8 (Outlier 8)	929,97	85,72	10,85
AO9 (Outlier 9)	375,00	85,89	4,37
AO10 (Outlier 10)	293,54	86,00	3,41
TC(0,2) (Outlier 11)	99,51	24,61	4,04

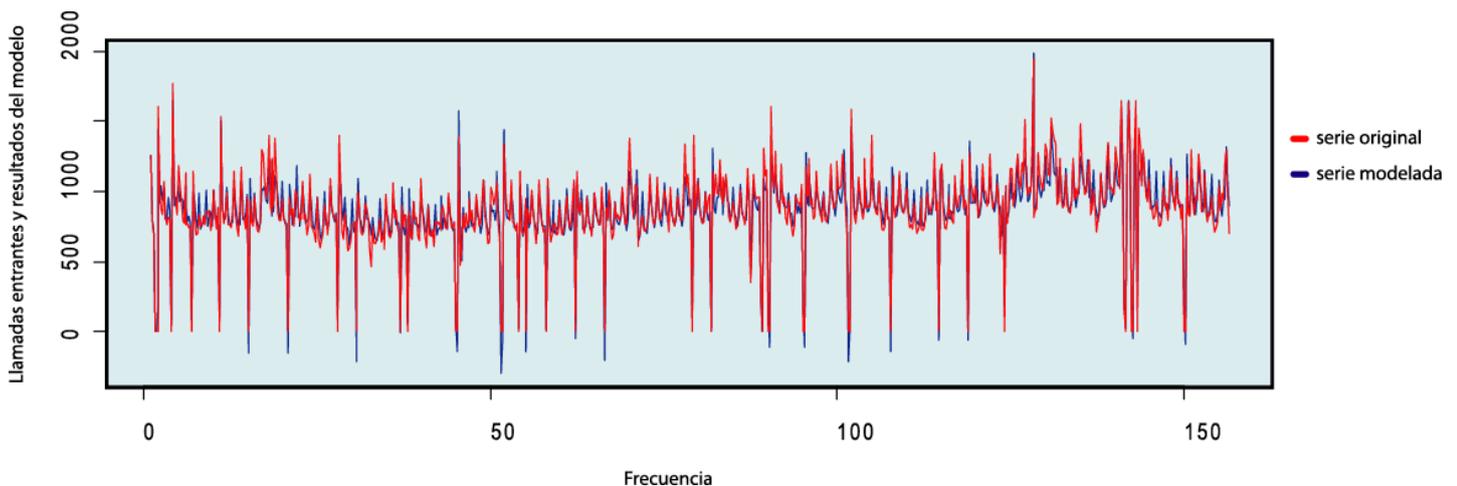
*⁵ Ver páginas 66 a 69 en ANEXO

Arriba se exponen los coeficientes resultantes de la estimación a través del método de máxima verosimilitud en el modelo identificado, así como también los coeficientes a considerar de verificarse alguno de los efectos feriado definidos, de estar en presencia de alguno de los outliers ya identificados o si se tratase de algún dato faltante. Además se detallan los desvíos estándar de cada una de estas estimaciones. Para comprobar el nivel de significación de éstos parámetros se calculó para cada uno el valor del estadístico t asumiendo un nivel de significación de 0,05.

Con el modelo definitivo planteado, todos los valores estimados para las variables determinantes del modelo quedan significativos a un nivel de 0,05, excepto por cuatro de los cinco datos faltantes, lo que era esperable debido a las estimaciones previas para los puntos de partida.

A continuación se expone un gráfico comparativo entre el modelo estimado y la serie original.

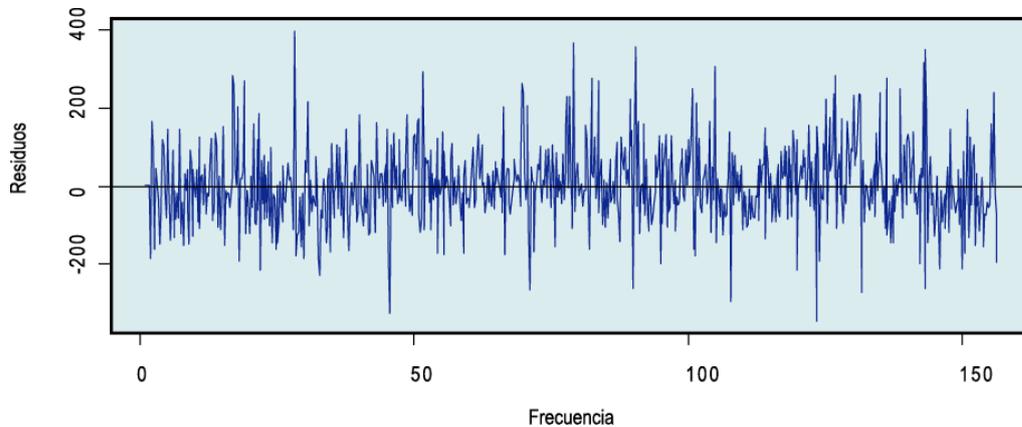
Gráfico 2.10
Gráfico comparativo entre serie original y serie modelada



Diagnóstico

Para iniciar el último paso correspondiente a la metodología de modelización propuesta por Box y Jenkins se procedió a analizar los residuos del modelo. Abajo se muestra un gráfico con la evolución de los residuos del modelo, un cuadro con los datos descriptivos de éstos y el histograma junto con la densidad de una distribución normal.

Gráfico 2.11
Residuos del modelo estimado



Cuadro 2.8
Datos descriptivos de los residuos del modelo estimado

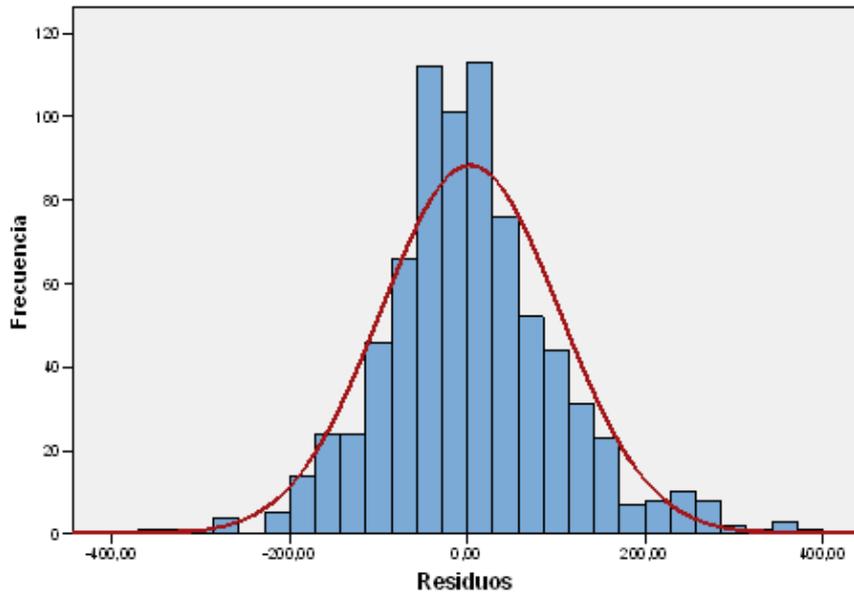
Residuos del modelo	Media		3,08
	Intervalo de confianza para la media al 95%	Límite inferior	-3,98
		Límite superior	10,14
	Mediana		-1,86
	Desv. típ.		100,34
	Mínimo		-344,20
	Máximo		394,98
	Asimetría		,40
	Curtosis		1,22

datos descriptivos calculados en el software SPSS, los mismos difieren levemente si son calculados en el software R por efectos de redondeo

A partir de los datos descriptivos calculados, se observa que la media de los residuos es cercana a cero e incluso el intervalo de confianza para la media al 95% contiene al cero dentro del mismo. El desvío estándar es de 100 y el máximo error en valor absoluto fue de 395 llamadas. Observando que el estimador de asimetría es cercano a cero puede detectarse que la distribución de los residuos es simétrica respecto a la media y observando el valor del estadístico de Curtosis se distingue que la distribución de los residuos es relativamente leptocúrtica, más empinada en comparación a la distribución normal.

Abajo se presenta un gráfico comparativo entre el histograma de los residuos y la densidad de una distribución Normal (0, 100), donde se pueden apreciar gráficamente, todos los comentarios surgidos a partir de los datos descriptivos.

Gráfico 2.12
Gráfico comparativo entre histograma de residuos y densidad de distribución Normal (0, 100)



Pruebas de Incorrección

El primer paso dentro de la etapa de diagnóstico consiste en verificar que la media marginal de los residuos sea igual a cero. Con los datos descriptivos calculados anteriormente se observó que la media no era significativamente distinta de cero y que el intervalo de confianza para la media al 95% contenía al cero. El siguiente aspecto a verificar es la incorrección de los residuos del modelo, para ello efectuamos dos pruebas; la prueba de Box-Pierce y la de Ljung-Box, ambas ya explicadas en el Marco Teórico. Estas pruebas fueron efectuadas hasta el rezago quince y sus resultados fueron:

Cuadro 2.9 Pruebas de incorrelación

test de Box-Pierce

DISTRIBUCIÓN ASINTÓTICA	VALOR	p-valor
χ^2 (quince g ^o de libert.)	25,23	0,047

test de Ljung-Box

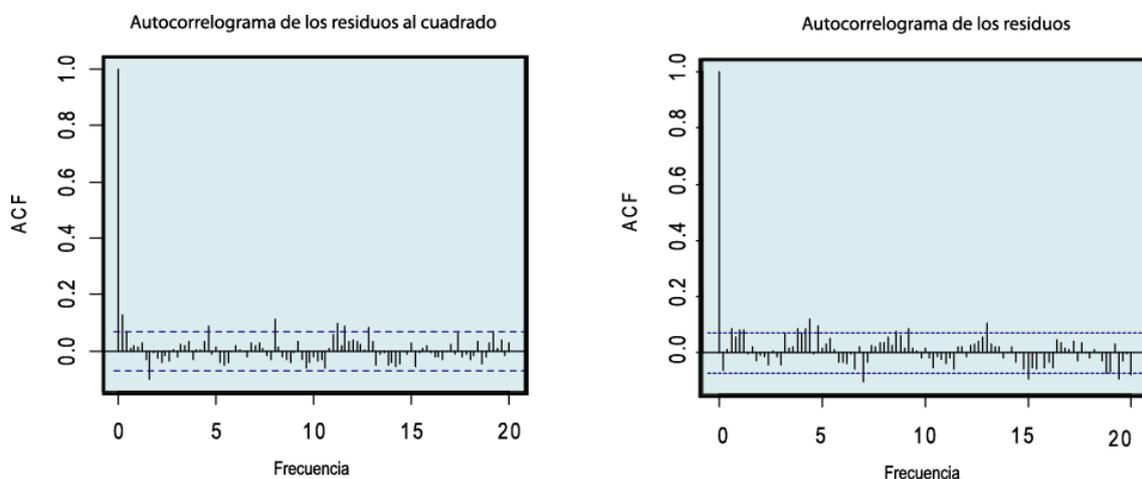
DISTRIBUCIÓN ASINTÓTICA	VALOR	p-valor
χ^2 (quince g ^o de libert.)	25,47	0,044

Observando el hecho de que los p-valores de ambas pruebas resultan menores a 0,05, se verifica que no es posible reunir suficiente evidencia estadística para rechazar la hipótesis nula de incorrelación a un nivel de confianza del 95%, aunque sí sería posible hacerlo con un nivel de confianza del 96% o mayor.

Testeo de homogeneidad de varianzas

Para testear si las varianzas de los residuos siguen alguna tendencia observamos que el autocorrelograma de los residuos al cuadrado no tome valores significativamente distintos de cero. Abajo se muestra el Autocorrelograma de los residuos del modelo elevados al cuadrado, conjuntamente con el autocorrelograma de los residuos.

Gráfico 2.13
Autocorrelogramas de los residuos y de los residuos al cuadrado del modelo

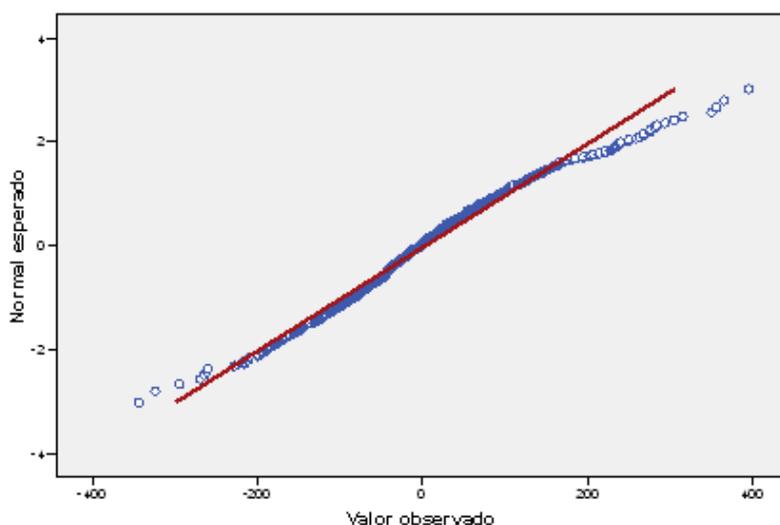


Observando el Aurocorrelograma de los residuos al cuadrado vemos que una gran proporción de las correlaciones entre éstos no son significativamente distintas de cero, por lo que podemos suponer que la varianza no sigue ninguna tendencia, o lo que es lo mismo; se mantiene constante.

Testeos de normalidad

A través de los estimadores de asimetría y curtosis se verificó que la distribución de los residuos es lo suficientemente simétrica y relativamente leptocúrtica. A través del gráfico comparativo entre la distribución de los residuos y una distribución Normal con iguales media y varianza, se expuso cómo visualmente ambas distribuciones no difieren de forma significativa. Abajo se expone el gráfico comparativo denominado Q-Q plot.

Gráfico 2.14
Gráfico Q-Q Normal de los residuos



En el Q-Q plot se exponen de forma comparativa, con una línea roja el gráfico esperado bajo distribución Normal y, en círculos azules los valores observados en los residuos del modelo. A través de esta otra comparación visual tampoco se observan diferencias marcadas entre ambos gráficos. Como pruebas de normalidad fueron elegidas la propuesta por Kolmogorov y Smirnov y la de Shapiro y Wilk.

Cuadro 2.10
Pruebas de normalidad

test de Kolmogorov-Smirnov

Valor del estadístico	Grados de libertad	p-valor
0,058	778	1,45e ⁻⁰⁰⁶

test de Shapiro-Wilk

Valor del estadístico	Grados de libertad	p-valor
0,981	778	1,75e ⁻⁰⁰⁸

Si bien en ambas pruebas se reúne suficiente evidencia estadística para rechazar la hipótesis nula de normalidad, estos resultados pueden ser atribuidos al alto número de casos con el que se efectuaron las pruebas. Por estas razones es que se desestimaron los resultados de las pruebas de Shapiro-Wilk y Kolmogorov-Smirnov, suponiendo válidas las impresiones visuales obtenidas anteriormente en los gráficos comparativos.

Análisis descriptivo de los residuos en valor absoluto

A los efectos de analizar la magnitud de los errores del modelo se efectuó un análisis específico considerando el valor absoluto de los residuos, de forma tal que no influyan las compensaciones entre errores de subestimación y sobreestimación.

Cuadro 2.11
Datos descriptivos de los residuos en valor absoluto

	Válidos	778
	Perdidos	0
Media		75,15
Desv. típ.		66,51
Mínimo		0,10
Máximo		394,98
Percentiles	5	4,28
	10	11,12
	25	25,83
	50	54,38
	75	104,66
	90	163,64
	95	214,32

Observando los datos descriptivos que surgen del análisis es posible corroborar que, en promedio, el modelo estima con un error de 75 llamadas; lo que representa un error aproximado del 8% respecto al promedio general de llamadas entrantes (913). A su vez se observa que el desvío estándar del error es de 66,51, que el máximo error fue de 395 llamadas y que el 90% de los errores del modelo son menores a 164 llamadas entrantes, lo que representa un porcentaje cercano al 18% del promedio general de la serie original.

2.5 EL MODELO COMO HERRAMIENTA DE PREDICCIÓN

La última etapa de la fase de diagnóstico consiste en verificar el poder predictivo del modelo estimado. En esta etapa se podrá evaluar la eficacia de la metodología para predecir en el corto plazo (una semana hábil) el volumen de llamadas entrantes al call center. Para efectuar las predicciones fue necesario partir la serie, luego plantear el modelo para la serie partida y por último establecer las predicciones de las observaciones que fueron excluidas de la serie, a este proceso se le denomina predicción dentro de la muestra. Las predicciones fueron evaluadas mediante el criterio de menor error relativo.

Predicción a un paso

Se efectuaron predicciones a un paso para cada uno de los días de la última, penúltima y antepenúltima semana completa de la serie. La matriz de regresiones utilizada en todos los casos se compone íntegramente por ceros, debido a que ninguna de las variables de regresión que intervienen en el modelo aplica para los últimos datos de la serie. Debajo se exponen los resultados de las predicciones a un paso.

Cuadro 2.12
Predicciones a un paso para las últimas tres semanas completas de la serie

Última Semana Completa

	Predicción	Error estándar	Observación	Error de Predicción	Porcentaje de Error
Lunes	988	100,34	1041	53	5,09 %
Martes	808	100,29	870	62	7,13 %
Miércoles	728	100,23	987	259	26,24 %
Jueves	836	100,33	933	97	10,40 %
Viernes	854	100,26	1192	338	28,36 %

Penúltima Semana Completa

	Predicción	Error estándar	Observación	Error de Predicción	Porcentaje de Error
Lunes	1027	100,41	968	59	6,10 %
Martes	772	100,50	799	27	3,38 %
Miércoles	693	100,47	716	23	3,21 %
Jueves	682	100,44	738	56	7,59 %
Viernes	745	100,38	788	43	5,46 %

Antepenúltima Semana Completa

	Predicción	Error estándar	Observación	Error de Predicción	Porcentaje de Error
Lunes	1061	100,60	1132	71	6,27 %
Martes	876	100,54	859	17	1,98 %
Miércoles	732	100,56	786	54	6,87 %
Jueves	723	100,50	842	119	14,13 %
Viernes	812	100,44	833	21	2,52 %

El promedio de los errores para las predicciones exhibidas fue de 8,98 %, ninguno de los errores fue mayor a 29% y los errores más grandes, en términos relativos, no se presentaron repetidas veces en los mismos días de la semana.

Predicción a cinco pasos

Para las mismas observaciones que se efectuaron las predicciones a un paso se realizaron también predicciones a cinco pasos de las tres semanas de forma íntegra. De manera que los días lunes quedaron con predicciones a un paso, los martes con predicciones a dos pasos y de forma análoga hasta los viernes que tuvieron predicciones a cinco pasos. Debajo se exponen los resultados de las mismas.

Cuadro 2.13
Predicciones a cinco pasos para las últimas tres semanas completas de la serie

Última Semana Completa

	Predicción	Error estándar	Observación	Error de Predicción	Porcentaje de Error
Lunes	987	100,34	1041	54	5,19 %
Martes	835	115,91	870	35	4,02 %
Miércoles	765	120,67	987	222	22,49 %
Jueves	764	122,22	933	169	18,11 %
Viernes	813	122,74	1192	379	31,80 %

Penúltima Semana Completa

	Predicción	Error estándar	Observación	Error de Predicción	Porcentaje de Error
Lunes	1027	100,41	968	59	6,10 %
Martes	863	115,68	799	64	8,01 %
Miércoles	787	120,26	716	71	9,92 %
Jueves	781	121,72	738	43	5,83 %
Viernes	827	122,19	788	39	4,95 %

Antepenúltima Semana Completa

	Predicción	Error estándar	Observación	Error de Predicción	Porcentaje de Error
Lunes	1060	100,60	1132	72	6,36 %
Martes	891	115,82	859	32	3,73 %
Miércoles	806	120,36	786	20	2,54 %
Jueves	790	121,80	842	52	6,17 %
Viernes	838	122,26	833	5	0,60 %

El promedio de los errores para las predicciones realizadas fue de 9,05 %, ninguno de los errores fue mayor a 32% y, al igual que lo ocurrido con las predicciones a un paso, los errores más grandes (en términos relativos) no se presentaron repetidas veces en los mismos días de la semana, aún cuando las predicciones fueron a una cantidad diferente de pasos para cada día de la semana.

Predicción a un paso con efecto calendario

Se realizaron predicciones a un paso para dos semanas distintas de la serie, una en la que el día lunes fue feriado y otra en la que tanto el día lunes como el martes fueron feriados consecutivos. De manera que una semana se vio afectada por los denominados

“efecto feriado 1” y “efecto feriado 2”, mientras que la otra lo hizo por el “doble efecto feriado 1” y el “doble efecto feriado 2”. Las matrices de regresiones en estos casos presentaron un uno en el efecto calendario que correspondió a cada predicción y ceros en el resto de los valores. A su vez para las predicciones correspondientes a la semana en la que únicamente el día lunes fue feriado; las matrices de regresiones tuvieron cuatro columnas menos, debido a que esta semana fue cronológicamente previa a cuatro de los outliers identificados. Mientras tanto para las predicciones correspondientes a la semana en la que tanto el día lunes como el martes fueron feriados; las matrices de regresiones presentaron cinco columnas menos, porque ésta semana fue además previa a uno de los datos faltantes que existieron en la serie. Debajo se exponen los resultados de las predicciones a un paso.

Cuadro 2.14
Predicciones a un paso para semanas con feriados

Predicciones a un paso con efecto calendario

	Predicción	Error estándar	Observación	Error de Predicción	Porcentaje de Error
Martes con Efecto Feriado 1	1033	96,57	1036	3	0,00 %
Miércoles con Efecto Feriado 2	849	96,49	820	29	0,03 %
Miércoles con Doble Efecto Feriado 1	1336	97,55	1178	158	13,41 %
Jueves con Doble Efecto Feriado 2	847	97,63	970	151	15,57 %

Los porcentajes de error se encontraron dentro de los niveles vistos en las anteriores predicciones que no incluían datos expuestos a efectos de calendario.

Predicción a cinco pasos con efecto calendario

Para las mismas dos semanas expuestas a efecto calendario se efectuaron predicciones a cinco pasos de las semanas completas. De esta manera para una semana: el “efecto feriado 1” se predijo en el segundo paso y el “efecto feriado 2” en el tercero, mientras que para la otra semana analizada: el “doble efecto feriado 1” se predijo en el tercer paso y el “doble efecto feriado 2” en el cuarto. Las matrices de regresiones presentaron unos en los efectos calendarios que correspondieron a cada predicción y ceros en el resto de los valores.

Cuadro 2.15
Predicciones a cinco pasos para semanas con feriados

Predicciones a cinco pasos con efecto calendario

	Predicción	Error estándar	Observación	Error de Predicción	Porcentaje de Error
Martes con Efecto Calendario 1	1150	108,92	1036	233	11,00 %
Miércoles con Efecto Calendario 2	908	112,22	820	88	10,73 %
Miércoles con Doble Efecto Calendario 1	1353	114,59	1178	175	14,86 %
Jueves con Doble Efecto Calendario 2	929	115,71	970	41	4,23 %

Nuevamente los porcentajes de error se encontraron dentro de los niveles vistos en las anteriores predicciones.

Predicción a cinco pasos en el inicio de la serie

Como última prueba predictiva dentro de la muestra, se efectuó una predicción a cinco pasos para los primeros datos de la serie de forma tal de verificar que los errores de predicción analizados no se correspondieran específicamente a un período de la serie en particular. De esta forma la serie se cortó luego de las primeras 37 observaciones, para predecir a cinco pasos concluyendo los primeros dos meses. Las predicciones se efectuaron de modo que para los días lunes y martes, las mismas fueran a cuatro y cinco pasos respectivamente, debido a que las anteriores predicciones para estos días fueron a uno y dos pasos. La matriz de regresiones presentó únicamente ocho columnas ya que los datos a predecir eran cronológicamente previos a nueve de los outliers identificados, a tres de los datos faltantes y a todos los veinticuatro de diciembre; a su vez uno de los outliers se correspondió con el único doble efecto feriado identificado dentro del período, por lo que tampoco se consideró. Adicionalmente dicha matriz se compuso íntegramente por ceros porque ningún regresor aplicaba para los datos a predecir.

Cuadro 2.16
Predicciones a cinco pasos a partir de la observación n°38

Octava y Novena Semana de la Serie

	Predicción	Error estándar	Observación	Error de Predicción	Porcentaje de Error
Miércoles	722	90,05	754	54	4,24 %
Jueves	657	105,91	829	116	20,75 %
Viernes	667	111,39	766	61	12,92 %
Lunes	738	113,42	856	31	13,79 %
Martes	749	114,19	847	61	11,57 %

Una vez más los porcentajes de error se encontraron dentro de los niveles vistos en las anteriores predicciones.

Con todas estas herramientas se podría concluir que el modelo es útil para predecir datos a cinco pasos con un porcentaje de error promedio de aproximadamente 10% , sin mejorar mayormente cuando las predicciones se aplican a un paso.

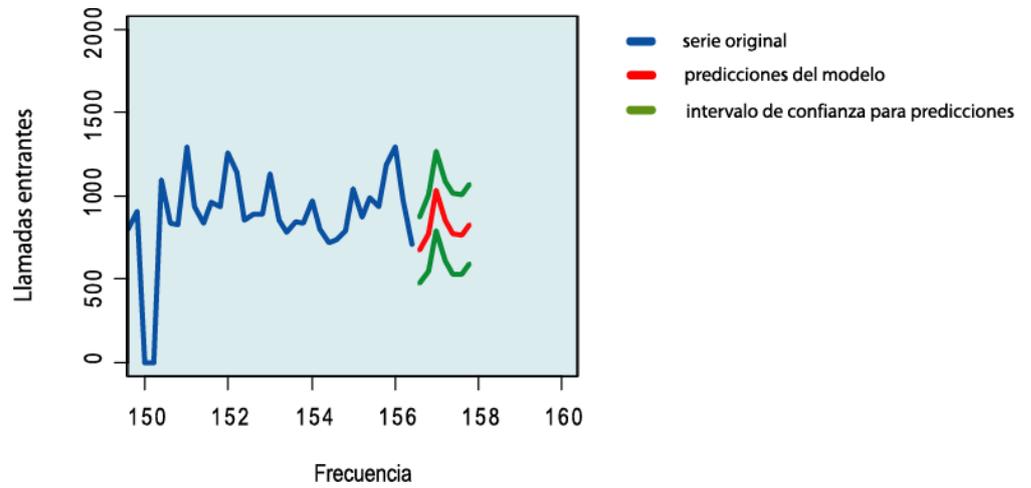
Predicción fuera de la muestra

Se aplicaron predicciones fuera de la muestra a siete pasos, estableciendo de esta forma predicciones para los dos últimos días de la semana que no se llega a completar con la serie original (la cual llega hasta un día miércoles) y, a su vez, estableciendo predicciones para la siguiente semana de forma completa. En este caso la matriz de regresiones quedó formada exclusivamente por ceros, ya que ninguna de las variables de regresión que intervienen en el modelo aplica para las predicciones efectuadas. Abajo se detallan los resultados de estas predicciones y un gráfico que muestra la supuesta evolución de la serie en base a dichas predicciones. Este gráfico muestra a su vez intervalos de confianza para las predicciones, los mismos fueron establecidos a un nivel de confianza del 95%.

Cuadro 2.17
Predicciones fuera de la muestra a siete pasos

	Predicción	Error estándar
Predicción 1 (jueves)	672	100,65
Predicción 2 (viernes)	776	116,27
Predicción 3 (lunes)	1033	121,05
Predicción 4 (martes)	853	122,60
Predicción 5 (miércoles)	773	123,12
Predicción 6 (jueves)	769	123,84
Predicción 7 (viernes)	832	124,08

Gráfico 2.15
Predicciones fuera de la muestra a siete pasos



3. CONCLUSIÓN

El objetivo de la pasantía consistía en poder modelizar adecuadamente el proceso de llegada de llamadas al Call Center Hola Itaú, de forma tal de poder predecir en el corto plazo, considerando como corto plazo el lapso de una semana hábil (5 días). Se abordó la modelización siguiendo la metodología propuesta por Box y Jenkins, y así se planteó un ajuste en tres etapas a un modelo SARIMA, con especial énfasis en analizar el efecto calendario y posibles estacionalidades múltiples. Se identificó así un modelo SARIMA $(1\ 0\ 0)(0\ 1\ 1)_5$ con una única estacionalidad estocástica (semanal), al que se le adicionan regresores por efecto calendario e intervenciones. El modelo estimado fue el siguiente:

$$\begin{aligned} Y_t = & 0,58*Y_{t-1} + Y_{t-5} - 0,58*Y_{t-6} - 0,93*\varepsilon_{t-5} + \varepsilon_t \dots \\ & - 953,25 * FE_t + 215,93 * EF1_t + 57,01 * EF2_t + 358,61 * DEF1_t + 123,68 * DEF2_t \dots \\ & - 874,42 * NB + 71,62 * FL1_t + 44,58 * FL2_t + 371 * FL3_t - 33,94 * FL4_t - 1,25 * FL5_t \dots \\ & + 395,71 * AO1_t + 676,19 * AO2_t + 474,64 * AO3_t + 410,64 * AO4_t - 313,87 * AO5_t \dots \\ & - 344,68 * AO6_t - 448,38 * AO7_t + 929,97 * AO8_t + 375 * AO9_t + 293,54 * AO10_t + 99,51 * TC(0,2) \end{aligned}$$

Donde FE_t , $EF1_t$, $EF2_t$, $DEF1_t$, $DEF2_t$ y NB_t representan las variables indicatrices de efectos de calendario. $FL1_t$, $FL2_t \dots FL5_t$ representan las indicatrices correspondientes a cada dato faltante de la serie. $AO1_t$, $AO2_t \dots AO10_t$ corresponden a los outliers de tipo AO que fueron identificados y TC_t representa la variable correspondiente al outlier que fue identificado como de tipo TC.

Evaluando las predicciones de este modelo SARIMA (mediante el error relativo de las mismas), se obtuvo como resultado final de las pruebas efectuadas que el modelo estimado es bueno para predecir en el corto plazo (5 días hábiles) con una tolerancia de error promedio de aproximadamente 10%.

4. BIBLIOGRAFÍA

Shapiro, S. S. y Wilk, M. B. (1965) *An Analysis of Variance Test for Normality (complete samples)*. Biometrika.

Allen A. O. (1972) *Probability, Statistics, and Queueing Theory with Computer Science Applications*. Academic Press.

Ross S. M. (1972) *Probability Models*. Academic Press.

Box, G. E. y Jenkins, G. M. (1976) *Time Series Analysis: Forecasting and Control*. Holden-Day.

Hillmer, S. C. , Bell, W. R. y Tiao, G. C. (1983) *Modeling Considerations in the Seasonal Adjustment of Economic Time Series*. A. Zellener.

Canavos, George C. (1986) *Probabilidad y Estadística, Aplicaciones y Métodos*. McGraw-Hill.

Chang, I. H., Tiao, G. C. and C. Chen (1988) *Estimation of Time Series Parameters in the Presence of Outliers*. Technometrics.

Gómez, V., Maravall A. y Peña D. (1993) *Computing Missing Values in Time Series*. Universidad Carlos III, Madrid.

Espasa, A., Revuelta J. M. y Cancelo J. R. (1996) *Modelización Automática de Series Diarias de Actividad Económica*. Universidad Carlos III, Madrid.

Devore, Jay L. (2001/2004 segunda edición) *Probabilidad y Estadística para Ingeniería y Ciencia*. Thomson Editores.

Gras, Jaume Arnau (2001) *Diseños de Series Temporales: Técnicas de Análisis*. Edicions Universitat de Barcelona.

Revuelta, J. M. (2001): *Desarrollo de una Metodología Automática de Modelización de Series Diarias de Actividad Económica. Aplicación a series diarias de demanda eléctrica*. Tesis doctoral, Universidad Carlos III. Madrid.

Caracena, J. A. (2002) *Un Procedimiento Completo para la detección de estacionalidad en series Económicas*. CEMFI, España.

Casella, G y Berger, R. L. (2002 second edition) *Statistical Inference*. Duxbury.

Pardo, A. y Ruiz, M. A. (2002) *SPSS 11. Guía para el Análisis de Datos*. Interamericana de España S.A.

María Ángeles Cea D`Ancona (2004) Textos de apoyo para la materia Análisis Multivariante, Departamento de Sociología, Universidad Complutense de Madrid.

César Pérez López (2006) *Problemas resueltos de econometría*. Thomson Editores, Madrid

Lopes, R. H. C. y Reid, I., Hobson, P. R. (2007) *The two-dimensional Kolmogorov-Smirnov test*. XI International Workshop on Advanced Computing and Analysis Techniques in Physics Research.

Rafael de Arce y Ramón Mahía (2008) Publicaciones del Departamento de Economía Aplicada (U.D.I. Econometría e Informática), Universidad Autónoma de Madrid.

5. ANEXO

A continuación se exponen los códigos de programación en software R, así como las salidas del mismo programa durante la etapa de ajuste a un modelo SARIMA y las diferentes etapas de predicción (una vez definido el modelo).

```
##### IMPORTAR LOS ARCHIVOS .TXT #####

setwd("C:/Archivos de programa/R/R-2.13.1")
###instalar y cargar paquete TSA
feriados=read.table("feriados.txt",header=FALSE)
feriados=ts(feriados)
nochebuena=read.table("noche buena.txt",header=FALSE)
nochebuena=ts(nochebuena)
efecto_feriado0=read.table("pre_feriado.txt",header=FALSE)
efecto_feriado0=ts(efecto_feriado0)
efecto_feriado1=read.table("post_feriado.txt",header=FALSE)
efecto_feriado1=ts(efecto_feriado1)
efecto_feriado2=read.table("segundo_post_feriado.txt",header=FALSE)
efecto_feriado2=ts(efecto_feriado2)
doble_efecto_feriado1=read.table("doble_efecto_feriado.txt",header=FALSE)
doble_efecto_feriado1=ts(doble_efecto_feriado1)
doble_efecto_feriado=read.table("doble_efecto_feriado1.txt",header=FALSE)
doble_efecto_feriado=ts(doble_efecto_feriado)
doble_efecto_feriado2=read.table("doble_efecto_feriado2.txt",header=FALSE)
doble_efecto_feriado2=ts(doble_efecto_feriado2)
#####
#### OUTLIERS ####
#####
outlier7=read.table("7.txt",header=FALSE)
outlier7=ts(outlier7, frequency=5)
outlier17=read.table("17.txt",header=FALSE)
outlier17=ts(outlier17, frequency=5)
outlier52=read.table("52.txt",header=FALSE)
outlier52=ts(outlier52, frequency=5)
outlier89=read.table("89.txt",header=FALSE)
outlier89=ts(outlier89, frequency=5)
outlier101=read.table("101.txt",header=FALSE)
outlier101=ts(outlier101, frequency=5)
outlier150=read.table("150.txt",header=FALSE)
outlier150=ts(outlier150, frequency=5)
outlier433=read.table("433.txt",header=FALSE)
outlier433=ts(outlier433, frequency=5)
outlier637=read.table("637.txt",header=FALSE)
outlier637=ts(outlier637, frequency=5)
outlier638=read.table("638.txt",header=FALSE)
outlier638=ts(outlier638, frequency=5)
outlier650=read.table("650.txt",header=FALSE)
outlier650=ts(outlier650, frequency=5)
outlier701=read.table("701.txt",header=FALSE)
outlier701=ts(outlier701, frequency=5)
outlier706=read.table("706.txt",header=FALSE)
```

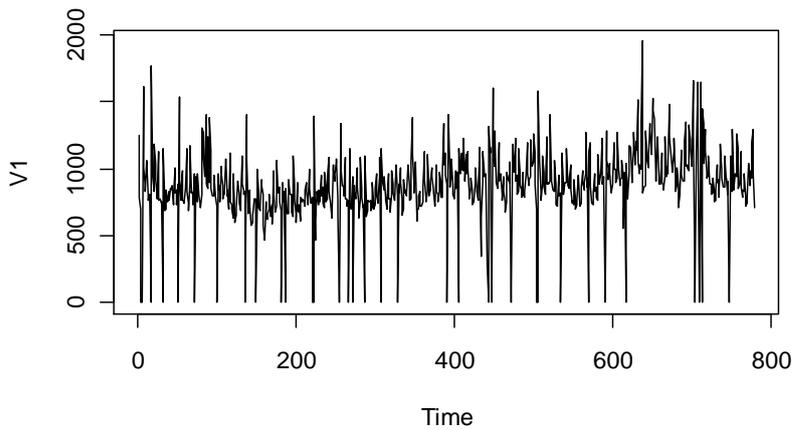
```

outlier706=ts(outlier706, frequency=5)
outlier706p=read.table("706`.txt",header=FALSE)
outlier706p=ts(outlier706p, frequency=5)
outlier706pp=read.table("706``.txt",header=FALSE)
outlier706pp=ts(outlier706pp, frequency=5)
#####
enero=read.table("enero.txt",header=FALSE)
enero=ts(enero)
febrero=read.table("febrero.txt",header=FALSE)
febrero=ts(febrero)
marzo=read.table("marzo.txt",header=FALSE)
marzo=ts(marzo)
abril=read.table("abril.txt",header=FALSE)
abril=ts(abril)
mayo=read.table("mayo.txt",header=FALSE)
mayo=ts(mayo)
junio=read.table("junio.txt",header=FALSE)
junio=ts(junio)
julio=read.table("julio.txt",header=FALSE)
julio=ts(julio)
agosto=read.table("agosto.txt",header=FALSE)
agosto=ts(agosto)
setiembre=read.table("setiembre.txt",header=FALSE)
setiembre=ts(setiembre)
octubre=read.table("octubre.txt",header=FALSE)
octubre=ts(octubre)
noviembre=read.table("noviembre.txt",header=FALSE)
noviembre=ts(noviembre)
faltante1=read.table("faltante1.txt",header=FALSE)
faltante1=ts(faltante1)
faltante2=read.table("faltante2.txt",header=FALSE)
faltante2=ts(faltante2)
faltante3=read.table("faltante3.txt",header=FALSE)
faltante3=ts(faltante3)
faltante4=read.table("faltante4.txt",header=FALSE)
faltante4=ts(faltante4)
faltante5=read.table("faltante5.txt",header=FALSE)
faltante5=ts(faltante5)

#####  IMPORTAR Y GRAFICAR LA SERIE  #####

serie=read.table("serie2.txt",header=FALSE)
serie=ts(serie)
plot(serie)
serie=ts(serie, frequency=5)
dserie=diff(serie,5,1)

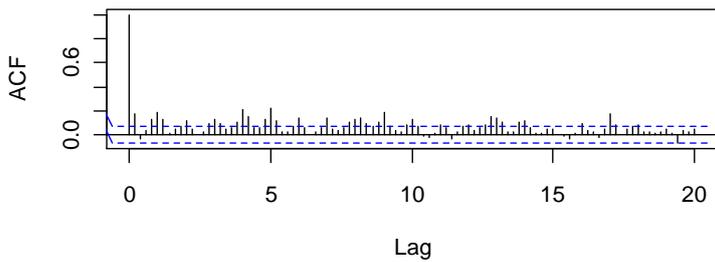
```



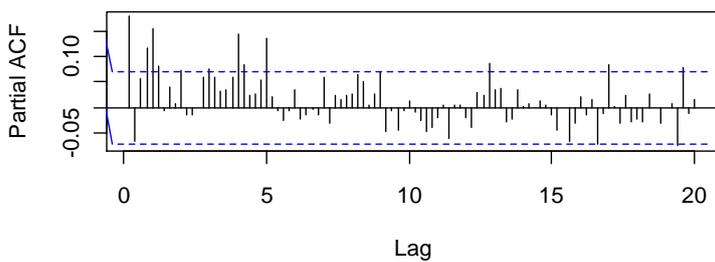
IDENTIFICACIÓN DEL MODELO

```
par(mfrow=c(2,1))
a=acf(serie, lag.max=100)
b=pacf(serie, lag.max=100)
```

V1

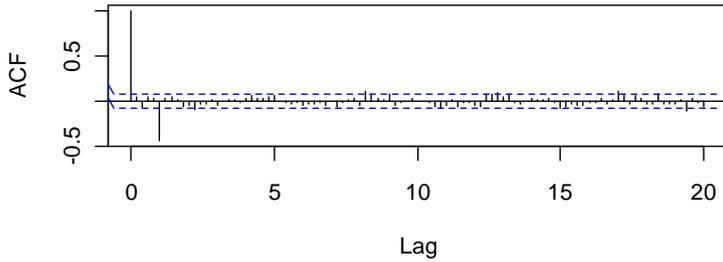


Series serie

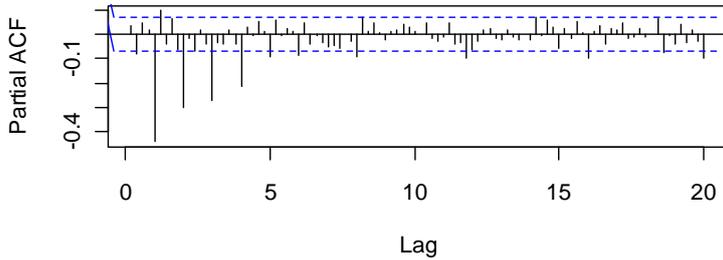


```
par(mfrow=c(2,1))
a=acf(dserie, lag.max=100)
b=pacf(dserie, lag.max=100)
```

V1



Series dserie



```
##### ESTIMACIÓN DEL MODELO #####

rega=cbind(feriados[1:778,],nochebuena[1:778,],efecto_feriado0[1:778,],efecto_
feriadol[1:778,],efecto_feriado2[1:778,],doble_efecto_feriadol[1:778,],doble_e
fecto_feriado2[1:778,],faltantel[1:778,],
faltante2[1:778,],faltante3[1:778,],faltante4[1:778,],
faltante5[1:778,],enero[1:778,],febrero[1:778,],marzo[1:778,],
abril[1:778,],mayo[1:778,],junio[1:778,],julio[1:778,], agosto[1:778,],
setiembre[1:778,], octubre[1:778,], noviembre[1:778,])

ma=arima(serie,
method="ML",order=c(1,0,0),seasonal=list(order=c(0,1,1),period=5),xreg=rega)
ma
Call:
arima(x = serie, order = c(1, 0, 0), seasonal = list(order = c(0, 1, 1), period = 5),
xreg = rega, method = "ML")

Coefficients:
ar1      sma1      rega1      rega2      rega3      rega4      rega5      rega6      rega7
0.4587  -0.9211  -944.7239 -856.0324 -4.4130  237.6172  76.2987  419.4025  130.1499
s.e.    0.0347   0.0188   23.0823   78.6668  23.6437   27.0035  25.0293  45.4072  42.1189
rega8   rega9   rega10  rega11  rega12  rega13  rega14  rega15
39.9364  21.6170 313.7127 -51.6015  5.7512 -46.1317 -70.1396 -26.0429
s.e.    118.1323 120.8812 114.2785 151.9543 107.9278 34.4877 37.2185 37.4132
rega16  rega17  rega18  rega19  rega20  rega21  rega22  rega23
-32.4101 -66.4066 -63.9894 -54.3673 -8.6919 14.1254 11.3555 -21.2502
s.e.    38.2930 38.8802 38.9731 38.4504 37.8735 37.7724 36.2556 35.4088

sigma^2 estimated as 14302: log likelihood = -4799.76, aic = 9651.52

t=ma$coef/sqrt(diag(ma$var.coef))
t
ar1      sma1      rega1      rega2      rega3      rega4      rega5
13.20358341 -49.11564892 -40.92845083 -10.88174944 -0.18664733  8.79949551  3.04837590
rega6      rega7      rega8      rega9      rega10      rega11      rega12
9.23647911  3.09006111  0.33806490  0.17882859  2.74516020 -0.33958567  0.05328735
rega13      rega14      rega15      rega16      rega17      rega18      rega19
-1.33762736 -1.88453716 -0.69608819 -0.84637154 -1.70797982 -1.64188947 -1.41396134
rega20      rega21      rega22      rega23
-0.22949705  0.37395995  0.31320582 -0.60013778
```

```
#####
##### BUSCO OUTLIERS

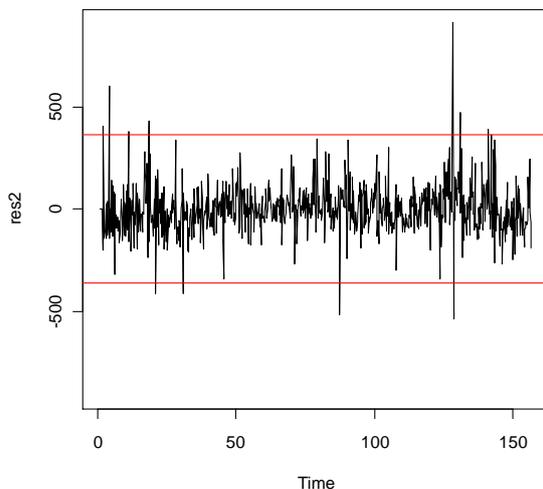
reg2=cbind(feriados[1:778,],nochebuena[1:778,],efecto_feriado1[1:778,]
efecto_feriado2[1:778,],doble_efecto_feriado1[1:778,],
doble_efecto_feriado2[1:778,],faltante1[1:778,],faltante2[1:778,],
faltante3[1:778,],faltante4[1:778,],faltante5[1:778,])
m2=arima(serie,method="ML",order=c(1,0,0),seasonal=list(order=c(0,1,1),period=
5),xreg=reg2)

m2
Call:
arima(x = serie, order = c(1, 0, 0), seasonal = list(order = c(0, 1, 1), period = 5),
      xreg = reg2, method = "ML")

Coefficients:
      ar1      sma1      reg21      reg22      reg23      reg24      reg25      reg26      reg27
 0.4847 -0.9221 -950.5603 -860.4899 231.2628 71.5855 409.0774 121.8669 51.2572
s.e. 0.0341 0.0194 21.2716 75.9114 26.7132 24.9368 45.0982 41.9954 117.5315
      reg28      reg29      reg210      reg211
 36.4959 314.3156 -43.9334 5.8498
s.e. 117.7340 113.9348 151.4597 107.5782

sigma^2 estimated as 14521: log likelihood = -4805.65, aic = 9639.3

res2=m2$residuals
plot(res2,ylim=c(-900,900))
abline(3*(14521^(0.5)), 0, col="red")
abline(-3*(14521^(0.5)), 0, col="red")
```



```
# ENCUENTRO 12 POSIBLES Y LOS CONSIDERO AO

reg22i=cbind(feriados[1:778,],nochebuena[1:778,],efecto_feriado1[1:778,],
efecto_feriado2[1:778,],doble_efecto_feriado1[1:778,],
doble_efecto_feriado2[1:778,],faltante1[1:778,],faltante2[1:778,],
faltante3[1:778,],faltante4[1:778,],faltante5[1:778,],
outlier7[1:778,],outlier17[1:778,],outlier52[1:778,],
outlier89[1:778,],outlier101[1:778,],outlier150[1:778,],
outlier433[1:778,],outlier637[1:778,],outlier638[1:778,],outlier650[1:778,],
outlier701[1:778,],outlier706[1:778,])
m22i=arima(serie,
method="ML",order=c(1,0,0),seasonal=list(order=c(0,1,1),period=5),xreg=reg22i)
```

```
t=m22i$coef/sqrt(diag(m22i$var.coef))
t
```

```
      ar1      sma1      reg22i1      reg22i2      reg22i3      reg22i4      reg22i5
18.05587280 -50.48551140 -53.05863387 -14.23789989  8.80581865  2.64840482  8.76555024
      reg22i6      reg22i7      reg22i8      reg22i9      reg22i10      reg22i11      reg22i12
 3.43972423  0.72835739  0.44928200  4.06806883 -0.26163096 -0.03882967  4.26841457
      reg22i13      reg22i14      reg22i15      reg22i16      reg22i17      reg22i18      reg22i19
 7.62002630  5.40392701  4.76773086 -3.51051434 -3.84855190 -5.21654128  9.17536938
      reg22i20      reg22i21      reg22i22      reg22i23
-1.23867276  4.37187318  3.38895551  2.86288418
```

```
# DESCARTO 1 POR NO SER SIGNIFICATIVO Y PRUEBO UN POSIBLE TC 0.2 O 0.5
#OUTLIER 706 TOMADO COMO DE TIPO AO
```

```
reg22=cbind(feriados[1:778,],nochebuena[1:778,],      efecto_feriado1[1:778,],
efecto_feriado2[1:778,], doble_efecto_feriado1[1:778,],
doble_efecto_feriado2[1:778,],      faltante1[1:778,],      faltante2[1:778,],
faltante3[1:778,], faltante4[1:778,], faltante5[1:778,], outlier7[1:778,],
outlier17[1:778,], outlier52[1:778,], outlier89[1:778,], outlier101[1:778,],
outlier150[1:778,],      outlier433[1:778,],      outlier637[1:778,],
outlier650[1:778,], outlier701[1:778,], outlier706[1:778,])
```

```
#OUTLIER 706 TOMADO COMO DE TIPO TC 0.2
```

```
reg22p=cbind(feriados[1:778,],nochebuena[1:778,],      efecto_feriado1[1:778,],
efecto_feriado2[1:778,],      doble_efecto_feriado1[1:778,],
doble_efecto_feriado2[1:778,],      faltante1[1:778,],      faltante2[1:778,],
faltante3[1:778,], faltante4[1:778,], faltante5[1:778,], outlier7[1:778,],
outlier17[1:778,], outlier52[1:778,], outlier89[1:778,], outlier101[1:778,],
outlier150[1:778,],      outlier433[1:778,],      outlier637[1:778,],
outlier650[1:778,], outlier701[1:778,], outlier706p[1:778,])
```

```
#OUTLIER 706 TOMADO COMO DE TIPO TC 0.5
```

```
reg22pp=cbind(feriados[1:778,],nochebuena[1:778,],      efecto_feriado1[1:778,],
efecto_feriado2[1:778,],      doble_efecto_feriado1[1:778,],
doble_efecto_feriado2[1:778,],      faltante1[1:778,],      faltante2[1:778,],
faltante3[1:778,], faltante4[1:778,], faltante5[1:778,], outlier7[1:778,],
outlier17[1:778,], outlier52[1:778,], outlier89[1:778,], outlier101[1:778,],
outlier150[1:778,],      outlier433[1:778,],      outlier637[1:778,],
outlier650[1:778,], outlier701[1:778,], outlier706pp[1:778,])
```

```
m22=arima(serie,
method="ML",order=c(1,0,0),seasonal=list(order=c(0,1,1),period=5),xreg=reg22)
m22p=arima(serie,
method="ML",order=c(1,0,0),seasonal=list(order=c(0,1,1),period=5),xreg=reg22p)
m22pp=arima(serie,
method="ML",order=c(1,0,0),seasonal=list(order=c(0,1,1),period=5),xreg=reg22p)
```

```
m22i
```

```
Call:
```

```
arima(x = serie, order = c(1, 0, 0), seasonal = list(order = c(0, 1, 1), period = 5),
      xreg = reg22i, method = "ML")
```

```
Coefficients:
```

```
      ar1      sma1      reg22i1      reg22i2      reg22i3      reg22i4      reg22i5      reg22i6      reg22i7
 0.578 -0.9297 -954.5068 -876.3639 211.4541 55.8529 349.0985 120.1286 70.6467
s.e.  0.032  0.0184  17.9897  61.5515  24.0130  21.0893  39.8262  34.9239  96.9946
      reg22i8      reg22i9      reg22i10      reg22i11      reg22i12      reg22i13      reg22i14      reg22i15
43.6474 378.5160 -32.1307 -3.3903 403.4079 679.6421 478.1081 410.9645
s.e.  97.1492 93.0456 122.8092 87.3133 94.5100 89.1916 88.4742 86.1971
      reg22i16      reg22i17      reg22i18      reg22i19      reg22i20      reg22i21      reg22i22      reg22i23
-310.9527 -340.3433 -449.7278 876.1657 -118.2237 376.9233 292.6588 252.9746
s.e.  88.5775 88.4341 86.2119 95.4911 95.4438 86.2155 86.3566 88.3636
```

```
sigma^2 estimated as 10219: log likelihood = -4670.16, aic = 9392.32
```

m22

Call:

arima(x = serie, order = c(1, 0, 0), seasonal = list(order = c(0, 1, 1), period = 5),
xreg = reg22, method = "ML")

Coefficients:

	ar1	sma1	reg221	reg222	reg223	reg224	reg225	reg226	reg227
	0.5771	-0.9299	-954.4230	-875.5547	211.6111	55.9715	349.5711	120.2110	70.7890
s.e.	0.0320	0.0183	18.0084	61.6335	24.0380	21.1149	39.8729	34.9645	97.1089
	reg228	reg229	reg2210	reg2211	reg2212	reg2213	reg2214	reg2215	reg2216
	43.8142	378.0971	-32.6801	-2.8773	402.8645	679.7386	477.9601	411.074	-310.8853
s.e.	97.2655	93.1773	122.9703	87.4298	94.6426	89.3156	88.5992	86.318	88.7013
	reg2217	reg2218	reg2219	reg2220	reg2221	reg2222			
	-340.5209	-449.5886	927.3635	377.1008	292.5550	252.8224			
s.e.	88.5580	86.3372	86.1739	86.3368	86.4799	88.4887			

sigma^2 estimated as 10240: log likelihood = -4670.93, aic = 9391.85

m22p

Call:

arima(x = serie, order = c(1, 0, 0), seasonal = list(order = c(0, 1, 1), period = 5),
xreg = reg22p, method = "ML")

Coefficients:

	ar1	sma1	reg22p1	reg22p2	reg22p3	reg22p4	reg22p5	reg22p6	reg22p7
	0.5784	-0.9322	-953.2512	-874.4164	215.9310	57.0061	358.6141	123.6813	71.6167
s.e.	0.0318	0.0177	17.8890	61.3063	23.9139	20.7823	39.7009	34.7905	96.5776
	reg22p8	reg22p9	reg22p10	reg22p11	reg22p12	reg22p13	reg22p14	reg22p15	
	44.6756	371.0028	-33.9434	-1.2536	395.7104	676.1875	474.6382	410.6401	
s.e.	96.7467	92.6895	122.3067	86.9680	94.1116	88.8251	88.1373	85.8632	
	reg22p16	reg22p17	reg22p18	reg22p19	reg22p20	reg22p21	reg22p22		
	-313.8713	-344.6831	-448.3760	929.9693	375.0044	293.5421	99.5124		
s.e.	88.2379	88.1126	85.8703	85.7194	85.8943	85.9970	24.6066		

sigma^2 estimated as 10130: log likelihood = -4666.87, aic = 9383.73

m22pp

Call:

arima(x = serie, order = c(1, 0, 0), seasonal = list(order = c(0, 1, 1), period = 5),
xreg = reg22pp, method = "ML")

Coefficients:

	ar1	sma1	reg22pp1	reg22pp2	reg22pp3	reg22pp4	reg22pp5	reg22pp6	reg22pp7
	0.5756	-0.9321	-956.4051	-875.7945	214.2016	59.8905	342.9456	115.2396	70.7020
s.e.	0.0318	0.0179	17.9191	61.4961	23.9397	20.7804	39.7594	34.8644	96.8299
	reg22pp8	reg22pp9	reg22pp10	reg22pp11	reg22pp12	reg22pp13	reg22pp14	reg22pp15	
	42.7605	381.7505	-29.6939	-5.1591	406.6073	677.9900	476.4089	410.7098	
s.e.	96.9984	92.9552	122.6880	87.2333	94.4007	89.1025	88.4100	86.1346	
	reg22pp16	reg22pp17	reg22pp18	reg22pp19	reg22pp20	reg22pp21	reg22pp22		
	-312.6694	-342.5417	-448.8802	930.4778	377.0984	292.0435	41.5887		
s.e.	88.5101	88.3862	86.1511	85.9916	86.1641	86.2726	11.4583		

sigma^2 estimated as 10171: log likelihood = -4668.41, aic = 9386.81

MENOR AIC EN EL MODELO CON OUTLIER 706 TOMADO COMO DE TIPO TC 0.2

PRUEBA DE SIGNIFICACIÓN CON EL MODELO ELEGIDO

t=m22p\$coef/sqrt(diag(m22p\$var.coef))

t

	ar1	sma1	reg22p1	reg22p2	reg22p3	reg22p4	reg22p5
	18.20146608	-52.53535369	-53.28703749	-14.26307835	9.02952645	2.74301690	9.03289738
	reg22p6	reg22p7	reg22p8	reg22p9	reg22p10	reg22p11	reg22p12
	3.55503317	0.74154558	0.46177954	4.00264203	-0.27752725	-0.01441456	4.20469279
	reg22p13	reg22p14	reg22p15	reg22p16	reg22p17	reg22p18	reg22p19
	7.61257177	5.38521165	4.78249344	-3.55710401	-3.91184589	-5.22154728	10.84899227
	reg22p20	reg22p21	reg22p22				
	4.36588006	3.41339863	4.04414168				

GRAFICO COMPARATIVO ENTRE SERIE MODELADA Y SERIE ORIGINAL

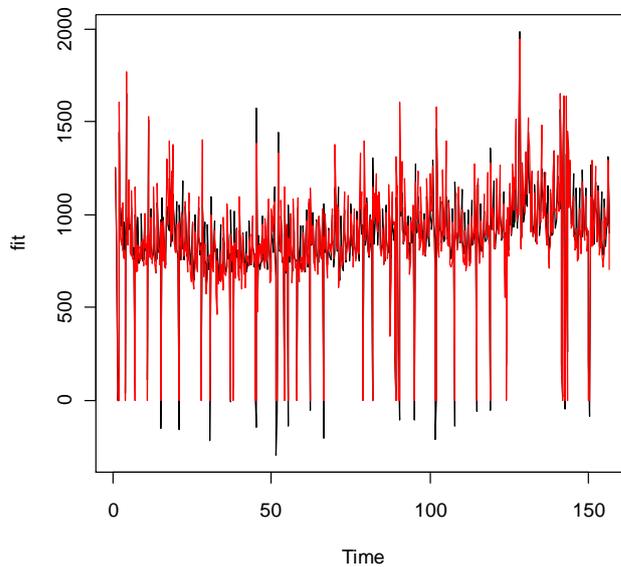
#CALCULAR LA SERIE DE RESIDUOS

res22=m22p\$residuals

#CALCULAR LA SERIE MODELADA

fit=serie[1:778,]- (res22)

```
#REALIZAR GRÁFICO COMPARATIVO
plot(fit)
lines(serie, col="red")
```

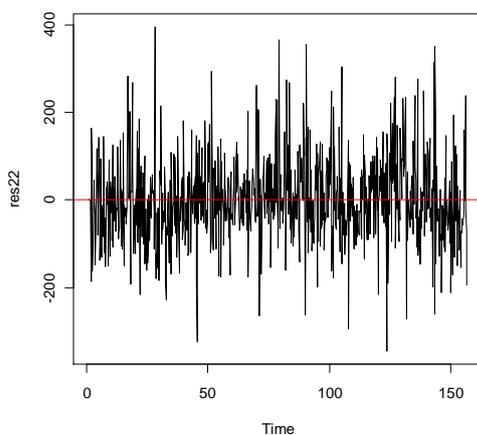


```
##### DIAGNÓSTICO #####
```

```
#EXPONER MEDIAS DE LA SERIE DE RESIDUOS Y LA SERIE DE RESIDUOS EN VALOR
ABSOLUTO
```

```
mean(res)
3.081948
res2=abs(res)
mean(res2)
75.14691
```

```
#GRAFICAR LA SERIE DE RESIDUOS
plot(res22)
abline(0, 0, col="red")
```



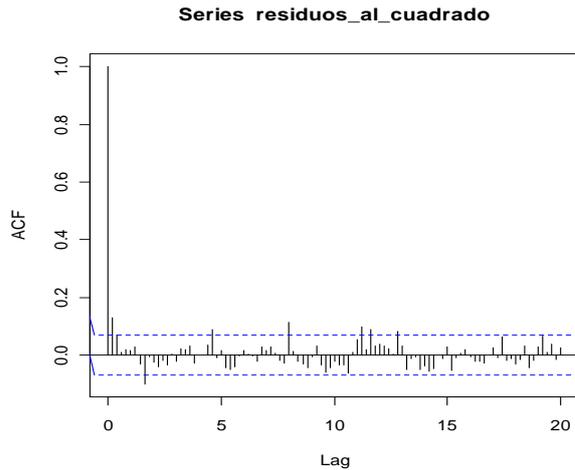
```
##### INCORRELACION
```

```
Box.test(res22,type="Box-Pierce", lag=15)
```

```
Box-Pierce test
data: res22
X-squared = 25.2319, df = 15, p-value = 0.04691
```

```
Box.test(res22,type="Ljung-Box", lag=15)
Box-Ljung test
data: res22
X-squared = 25.4735, df = 15, p-value = 0.04393
```

```
##### HOMOCEDEASTICIDAD
residuos_al_cuadrado=res22^2
al=acf(residuos_al_cuadrado, lag.max=100)
```



```
##### NORMALIDAD testada en SPSS
```

```
#####
##### PREDICCION DENTRO DE LA MUESTRA A 1 PASO #####
```

```
##### ULTIMA SEMANA #####
newreg <- matrix(c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0),
nrow = 1, ncol=22, byrow=TRUE, dimnames = list(c("row1"), c("C.1",
"C.2", "C.3", "C.4", "C.5", "C.6", "C.7", "C.8", "C.9", "C.10", "C.11",
"C.12", "C.13", "C.14", "C.15", "C.16", "C.17", "C.18", "C.19", "C.20",
"C.21", "C.22")))
reg01=cbind(feriados[1:770,],nochebuena[1:770,],
efecto_feriado1[1:770,],efecto_feriado2[1:770,],
doble_efecto_feriado1[1:770,],doble_efecto_feriado2[1:770,],
faltante1[1:770,],faltante2[1:770,],faltante3[1:770,],
faltante4[1:770,],faltante5[1:770,],outlier7[1:770,],
outlier17[1:770,],outlier52[1:770,],outlier89[1:770,],
outlier101[1:770,],outlier150[1:770,],outlier433[1:770,],
outlier637[1:770,],outlier650[1:770,],outlier701[1:770,], outlier706p[1:770,])
reg02=cbind(feriados[1:771,],nochebuena[1:771,], efecto_feriado1[1:771,],
efecto_feriado2[1:771,], doble_efecto_feriado1[1:771,],
doble_efecto_feriado2[1:771,], faltante1[1:771,], faltante2[1:771,],
faltante3[1:771,], faltante4[1:771,], faltante5[1:771,], outlier7[1:771,],
outlier17[1:771,], outlier52[1:771,], outlier89[1:771,], outlier101[1:771,],
outlier150[1:771,], outlier433[1:771,], outlier637[1:771,],
outlier650[1:771,], outlier701[1:771,], outlier706p[1:771,])
reg03=cbind(feriados[1:772,],nochebuena[1:772,], efecto_feriado1[1:772,],
efecto_feriado2[1:772,], doble_efecto_feriado1[1:772,],
doble_efecto_feriado2[1:772,], faltante1[1:772,], faltante2[1:772,],
```

```
faltante3[1:772,], faltante4[1:772,], faltante5[1:772,], outlier7[1:772,],
outlier17[1:772,], outlier52[1:772,], outlier89[1:772,], outlier101[1:772,],
outlier150[1:772,], outlier433[1:772,], outlier637[1:772,],
outlier650[1:772,], outlier701[1:772,], outlier706p[1:772,])
reg04=cbind(feriados[1:773,],nochebuena[1:773,], efecto_feriado1[1:773,],
efecto_feriado2[1:773,], doble_efecto_feriado1[1:773,],
doble_efecto_feriado2[1:773,], faltante1[1:773,], faltante2[1:773,],
faltante3[1:773,], faltante4[1:773,], faltante5[1:773,], outlier7[1:773,],
outlier17[1:773,], outlier52[1:773,], outlier89[1:773,], outlier101[1:773,],
outlier150[1:773,], outlier433[1:773,], outlier637[1:773,],
outlier650[1:773,], outlier701[1:773,], outlier706p[1:773,])
reg05=cbind(feriados[1:774,],nochebuena[1:774,], efecto_feriado1[1:774,],
efecto_feriado2[1:774,], doble_efecto_feriado1[1:774,],
doble_efecto_feriado2[1:774,], faltante1[1:774,], faltante2[1:774,],
faltante3[1:774,], faltante4[1:774,], faltante5[1:774,], outlier7[1:774,],
outlier17[1:774,], outlier52[1:774,], outlier89[1:774,], outlier101[1:774,],
outlier150[1:774,], outlier433[1:774,], outlier637[1:774,],
outlier650[1:774,], outlier701[1:774,], outlier706p[1:774,])
```

```
seriepart01=serie[1:770]
seriepart01=ts(seriepart01)
seriepart02=serie[1:771]
seriepart02=ts(seriepart02)
seriepart03=serie[1:772]
seriepart03=ts(seriepart03)
seriepart04=serie[1:773]
seriepart04=ts(seriepart04)
seriepart05=serie[1:774]
seriepart05=ts(seriepart05)
```

```
m01=arima(seriepart01,
method="ML",order=c(1,0,0),seasonal=list(order=c(0,1,1),period=5),xreg=reg01)
m02=arima(seriepart02,
method="ML",order=c(1,0,0),seasonal=list(order=c(0,1,1),period=5),xreg=reg02)
m03=arima(seriepart03,
method="ML",order=c(1,0,0),seasonal=list(order=c(0,1,1),period=5),xreg=reg03)
m04=arima(seriepart04,
method="ML",order=c(1,0,0),seasonal=list(order=c(0,1,1),period=5),xreg=reg04)
m05=arima(seriepart05,
method="ML",order=c(1,0,0),seasonal=list(order=c(0,1,1),period=5),xreg=reg05)
```

```
pred01=predict(m01,n.ahead=1,newxreg=newreg)
pred02=predict(m02,n.ahead=1,newxreg=newreg)
pred03=predict(m03,n.ahead=1,newxreg=newreg)
pred04=predict(m04,n.ahead=1,newxreg=newreg)
pred05=predict(m05,n.ahead=1,newxreg=newreg)
```

```
datos01=serie[771:771]
datos02=serie[772:772]
datos03=serie[773:773]
datos04=serie[774:774]
datos05=serie[775:775]
```

```
pred01
$pred
Time Series:
Start = 771
End = 771
Frequency = 1
```

```
row1
987.5448

$se
Time Series:
Start = 771
End = 771
Frequency = 1
[1] 100.3385
```

```
datos01
1041
```

```
pred02
$pred
Time Series:
Start = 772
End = 772
Frequency = 1
row1
808.3133
```

```
$se
Time Series:
Start = 772
End = 772
Frequency = 1
[1] 100.2859
```

```
datos02
870
```

```
pred03
$pred
Time Series:
Start = 773
End = 773
Frequency = 1
row1
727.5552
```

```
$se
Time Series:
Start = 773
End = 773
Frequency = 1
[1] 100.2294
```

```
datos03
987
```

```
pred04
$pred
Time Series:
Start = 774
End = 774
Frequency = 1
row1
835.6882
```

```

$se
Time Series:
Start = 774
End = 774
Frequency = 1
[1] 100.3261

```

```

datos04
933

```

```

pred05
$pred
Time Series:
Start = 775
End = 775
Frequency = 1
row1
853.9008

```

```

$se
Time Series:
Start = 775
End = 775
Frequency = 1
[1] 100.2608

```

```

datos05
1192

```

```

##### PENULTIMA SEMANA #####
newreg <- matrix(c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0),
nrow = 1, ncol=22, byrow=TRUE, dimnames = list(c("row1"), c("C.1",
"C.2", "C.3", "C.4", "C.5", "C.6", "C.7", "C.8", "C.9", "C.10", "C.11",
"C.12", "C.13", "C.14", "C.15", "C.16", "C.17", "C.18", "C.19", "C.20",
"C.21", "C.22")))
reg11=cbind(feriados[1:765,],nochebuena[1:765,], efecto_feriado1[1:765,],
efecto_feriado2[1:765,], doble_efecto_feriado1[1:765,],
doble_efecto_feriado2[1:765,], faltantel[1:765,], faltante2[1:765,],
faltante3[1:765,], faltante4[1:765,], faltante5[1:765,], outlier7[1:765,],
outlier17[1:765,], outlier52[1:765,], outlier89[1:765,], outlier101[1:765,],
outlier150[1:765,], outlier433[1:765,], outlier637[1:765,],
outlier650[1:765,], outlier701[1:765,], outlier706p[1:765,])
reg12=cbind(feriados[1:766,],nochebuena[1:766,], efecto_feriado1[1:766,],
efecto_feriado2[1:766,], doble_efecto_feriado1[1:766,],
doble_efecto_feriado2[1:766,], faltantel[1:766,], faltante2[1:766,],
faltante3[1:766,], faltante4[1:766,], faltante5[1:766,], outlier7[1:766,],
outlier17[1:766,], outlier52[1:766,], outlier89[1:766,], outlier101[1:766,],
outlier150[1:766,], outlier433[1:766,], outlier637[1:766,],
outlier650[1:766,], outlier701[1:766,], outlier706p[1:766,])
reg13=cbind(feriados[1:767,],nochebuena[1:767,], efecto_feriado1[1:767,],
efecto_feriado2[1:767,], doble_efecto_feriado1[1:767,],
doble_efecto_feriado2[1:767,], faltantel[1:767,], faltante2[1:767,],
faltante3[1:767,], faltante4[1:767,], faltante5[1:767,], outlier7[1:767,],
outlier17[1:767,], outlier52[1:767,], outlier89[1:767,], outlier101[1:767,],
outlier150[1:767,], outlier433[1:767,], outlier637[1:767,],
outlier650[1:767,], outlier701[1:767,], outlier706p[1:767,])
reg14=cbind(feriados[1:768,],nochebuena[1:768,], efecto_feriado1[1:768,],
efecto_feriado2[1:768,], doble_efecto_feriado1[1:768,],

```

```

doble_efecto_feriado2[1:768,],      faltante1[1:768,],      faltante2[1:768,],
faltante3[1:768,], faltante4[1:768,], faltante5[1:768,], outlier7[1:768,],
outlier17[1:768,], outlier52[1:768,], outlier89[1:768,], outlier101[1:768,],
outlier150[1:768,], outlier433[1:768,], outlier637[1:768,],
outlier650[1:768,], outlier701[1:768,], outlier706p[1:768,])
reg15=cbind(feriados[1:769,],nochebuena[1:769,], efecto_feriado1[1:769,],
efecto_feriado2[1:769,], doble_efecto_feriado1[1:769,],
doble_efecto_feriado2[1:769,], faltante1[1:769,], faltante2[1:769,],
faltante3[1:769,], faltante4[1:769,], faltante5[1:769,], outlier7[1:769,],
outlier17[1:769,], outlier52[1:769,], outlier89[1:769,], outlier101[1:769,],
outlier150[1:769,], outlier433[1:769,], outlier637[1:769,],
outlier650[1:769,], outlier701[1:769,], outlier706p[1:769,])

```

```

seriepart11=serie[1:765]
seriepart11=ts(seriepart11)
seriepart12=serie[1:766]
seriepart12=ts(seriepart12)
seriepart13=serie[1:767]
seriepart13=ts(seriepart13)
seriepart14=serie[1:768]
seriepart14=ts(seriepart14)
seriepart15=serie[1:769]
seriepart15=ts(seriepart15)

```

```

m11=arima(seriepart11,
method="ML",order=c(1,0,0),seasonal=list(order=c(0,1,1),period=5),xreg=reg11)
m12=arima(seriepart12,
method="ML",order=c(1,0,0),seasonal=list(order=c(0,1,1),period=5),xreg=reg12)
m13=arima(seriepart13,
method="ML",order=c(1,0,0),seasonal=list(order=c(0,1,1),period=5),xreg=reg13)
m14=arima(seriepart14,
method="ML",order=c(1,0,0),seasonal=list(order=c(0,1,1),period=5),xreg=reg14)
m15=arima(seriepart15,
method="ML",order=c(1,0,0),seasonal=list(order=c(0,1,1),period=5),xreg=reg15)

```

```

pred11=predict(m11,n.ahead=1,newxreg=newreg)
pred12=predict(m12,n.ahead=1,newxreg=newreg)
pred13=predict(m13,n.ahead=1,newxreg=newreg)
pred14=predict(m14,n.ahead=1,newxreg=newreg)
pred15=predict(m15,n.ahead=1,newxreg=newreg)

```

```

datos11=serie[766:766]
datos12=serie[767:767]
datos13=serie[768:768]
datos14=serie[769:769]
datos15=serie[770:770]

```

```

pred11
$pred
Time Series:
Start = 766
End = 766
Frequency = 1
row1
1026.823

```

```

$se
Time Series:
Start = 766

```

```
End = 766
Frequency = 1
[1] 100.412
```

```
datos11
968
```

```
pred12
$pred
Time Series:
Start = 767
End = 767
Frequency = 1
  row1
772.5499
```

```
$se
Time Series:
Start = 767
End = 767
Frequency = 1
[1] 100.5003
```

```
datos12
799
```

```
pred13
$pred
Time Series:
Start = 768
End = 768
Frequency = 1
  row1
692.7446
```

```
$se
Time Series:
Start = 768
End = 768
Frequency = 1
[1] 100.4663
```

```
datos13
716
```

```
pred14
$pred
Time Series:
Start = 769
End = 769
Frequency = 1
  row1
682.4306
```

```
$se
Time Series:
Start = 769
End = 769
```

```
Frequency = 1
[1] 100.4381
datos14
738
```

```
pred15
$pred
Time Series:
Start = 770
End = 770
Frequency = 1
row1
745.1517
```

```
$se
Time Series:
Start = 770
End = 770
Frequency = 1
[1] 100.3838
```

```
datos15
788
```

```
##### ANTEPENULTIMA SEMANA #####
newreg <- matrix(c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0),
nrow = 1, ncol=22, byrow=TRUE, dimnames = list(c("row1"), c("C.1",
"C.2", "C.3", "C.4", "C.5", "C.6", "C.7", "C.8", "C.9", "C.10", "C.11",
"C.12", "C.13", "C.14", "C.15", "C.16", "C.17", "C.18", "C.19", "C.20",
"C.21", "C.22")))
reg21=cbind(feriados[1:760,],nochebuena[1:760,], efecto_feriadol[1:760,],
efecto_feriado2[1:760,], doble_efecto_feriadol[1:760,],
doble_efecto_feriado2[1:760,], faltantel[1:760,], faltante2[1:760,],
faltante3[1:760,], faltante4[1:760,], faltante5[1:760,], outlier7[1:760,],
outlier17[1:760,], outlier52[1:760,], outlier89[1:760,], outlier101[1:760,],
outlier150[1:760,], outlier433[1:760,], outlier637[1:760,],
outlier650[1:760,], outlier701[1:760,], outlier706p[1:760,])
reg22=cbind(feriados[1:761,],nochebuena[1:761,], efecto_feriadol[1:761,],
efecto_feriado2[1:761,], doble_efecto_feriadol[1:761,],
doble_efecto_feriado2[1:761,], faltantel[1:761,], faltante2[1:761,],
faltante3[1:761,], faltante4[1:761,], faltante5[1:761,], outlier7[1:761,],
outlier17[1:761,], outlier52[1:761,], outlier89[1:761,], outlier101[1:761,],
outlier150[1:761,], outlier433[1:761,], outlier637[1:761,],
outlier650[1:761,], outlier701[1:761,], outlier706p[1:761,])
reg23=cbind(feriados[1:762,],nochebuena[1:762,], efecto_feriadol[1:762,],
efecto_feriado2[1:762,], doble_efecto_feriadol[1:762,],
doble_efecto_feriado2[1:762,], faltantel[1:762,], faltante2[1:762,],
faltante3[1:762,], faltante4[1:762,], faltante5[1:762,], outlier7[1:762,],
outlier17[1:762,], outlier52[1:762,], outlier89[1:762,], outlier101[1:762,],
outlier150[1:762,], outlier433[1:762,], outlier637[1:762,],
outlier650[1:762,], outlier701[1:762,], outlier706p[1:762,])
reg24=cbind(feriados[1:763,],nochebuena[1:763,], efecto_feriadol[1:763,],
efecto_feriado2[1:763,], doble_efecto_feriadol[1:763,],
doble_efecto_feriado2[1:763,], faltantel[1:763,], faltante2[1:763,],
faltante3[1:763,], faltante4[1:763,], faltante5[1:763,], outlier7[1:763,],
outlier17[1:763,], outlier52[1:763,], outlier89[1:763,], outlier101[1:763,],
outlier150[1:763,], outlier433[1:763,], outlier637[1:763,],
outlier650[1:763,], outlier701[1:763,], outlier706p[1:763,])
```

```

reg25=cbind(feriados[1:764,],nochebuena[1:764,],          efecto_feriado1[1:764,],
efecto_feriado2[1:764,],                                doble_efecto_feriado1[1:764,],
doble_efecto_feriado2[1:764,],          faltante1[1:764,],          faltante2[1:764,],
faltante3[1:764,], faltante4[1:764,], faltante5[1:764,], outlier7[1:764,],
outlier17[1:764,], outlier52[1:764,], outlier89[1:764,], outlier101[1:764,],
outlier150[1:764,],          outlier433[1:764,],          outlier637[1:764,],
outlier650[1:764,], outlier701[1:764,], outlier706p[1:764,])

```

```

seriepart21=serie[1:760]
seriepart21=ts(seriepart21)
seriepart22=serie[1:761]
seriepart22=ts(seriepart22)
seriepart23=serie[1:762]
seriepart23=ts(seriepart23)
seriepart24=serie[1:763]
seriepart24=ts(seriepart24)
seriepart25=serie[1:764]
seriepart25=ts(seriepart25)

```

```

m21=arima(seriepart21,
method="ML",order=c(1,0,0),seasonal=list(order=c(0,1,1),period=5),xreg=reg21)
m22=arima(seriepart22,
method="ML",order=c(1,0,0),seasonal=list(order=c(0,1,1),period=5),xreg=reg22)
m23=arima(seriepart23,
method="ML",order=c(1,0,0),seasonal=list(order=c(0,1,1),period=5),xreg=reg23)
m24=arima(seriepart24,
method="ML",order=c(1,0,0),seasonal=list(order=c(0,1,1),period=5),xreg=reg24)
m25=arima(seriepart25,
method="ML",order=c(1,0,0),seasonal=list(order=c(0,1,1),period=5),xreg=reg25)

```

```

pred21=predict(m21,n.ahead=1,newxreg=newreg)
pred22=predict(m22,n.ahead=1,newxreg=newreg)
pred23=predict(m23,n.ahead=1,newxreg=newreg)
pred24=predict(m24,n.ahead=1,newxreg=newreg)
pred25=predict(m25,n.ahead=1,newxreg=newreg)

```

```

datos21=serie[761:761]
datos22=serie[762:762]
datos23=serie[763:763]
datos24=serie[764:764]
datos25=serie[765:765]

```

```

pred21
$pred
Time Series:
Start = 761
End = 761
Frequency = 1
  row1
1060.666

```

```

$se
Time Series:
Start = 761
End = 761
Frequency = 1
[1] 100.6027

```

```

datos21

```

1132

```
pred22
$pred
Time Series:
Start = 762
End = 762
Frequency = 1
  row1
876.535
```

```
$se
Time Series:
Start = 762
End = 762
Frequency = 1
[1] 100.5397
```

```
datos22
859
```

```
pred23
$pred
Time Series:
Start = 763
End = 763
Frequency = 1
  row1
732.1344
```

```
$se
Time Series:
Start = 763
End = 763
Frequency = 1
[1] 100.5563
```

```
datos23
786
```

```
pred24
$pred
Time Series:
Start = 764
End = 764
Frequency = 1
  row1
723.248
```

```
$se
Time Series:
Start = 764
End = 764
Frequency = 1
[1] 100.5027
```

```
datos24
842
```

```

pred25
$pred
Time Series:
Start = 765
End = 765
Frequency = 1
  row1
812.1119

$se
Time Series:
Start = 765
End = 765
Frequency = 1
[1] 100.4395

datos25
833

#####
##### PREDICCIÓN DENTRO DE LA MUESTRA A 5 PASOS #####

##### ULTIMA SEMANA #####
newreg5 <- matrix(c(0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0,
0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0,
0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0,
0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0), nrow = 5, ncol=22, byrow=TRUE, dimnames =
list(c("row1", "row2", "row3", "row4", "row5"), c("C.1", "C.2", "C.3", "C.4",
"C.5", "C.6", "C.7", "C.8", "C.9", "C.10", "C.11", "C.12", "C.13", "C.14",
"C.15", "C.16", "C.17", "C.18", "C.19", "C.20", "C.21", "C.22")))
reg055=cbind(feriados[1:770,],nochebuena[1:770,], efecto_feriado1[1:770,],
efecto_feriado2[1:770,], doble_efecto_feriado1[1:770,],
doble_efecto_feriado2[1:770,], faltante1[1:770,], faltante2[1:770,],
faltante3[1:770,], faltante4[1:770,], faltante5[1:770,], outlier7[1:770,],
outlier17[1:770,], outlier52[1:770,], outlier89[1:770,], outlier101[1:770,],
outlier150[1:770,], outlier433[1:770,], outlier637[1:770,],
outlier650[1:770,], outlier701[1:770,], outlier706p[1:770,])

seriepart055=serie[1:770]
seriepart055=ts(seriepart055)

m055=arima(seriepart055,
method="ML",order=c(1,0,0),seasonal=list(order=c(0,1,1),period=5),xreg=reg055)

pred055=predict(m055,n.ahead=5,newxreg=newreg5)
datos055=serie[771:775]

pred055
$pred
Time Series:
Start = 771
End = 775
Frequency = 1
  row1    row2    row3    row4    row5
987.5448 834.7932 764.6487 764.3591 813.4595

$se
Time Series:
Start = 771

```

```

End = 775
Frequency = 1
[1] 100.3385 115.9114 120.6727 122.2240 122.7385

datos055
1041    870    987    933    1192

##### PENULTIMA SEMANA #####
newreg5 <- matrix(c(0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0,
0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0,
0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0,
0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0,
0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0), nrow = 5, ncol=22, byrow=TRUE, dimnames =
list(c("row1", "row2", "row3", "row4", "row5"), c("C.1", "C.2", "C.3", "C.4",
"C.5", "C.6", "C.7", "C.8", "C.9", "C.10", "C.11", "C.12", "C.13", "C.14",
"C.15", "C.16", "C.17", "C.18", "C.19", "C.20", "C.21", "C.22")))
reg155=cbind(feriados[1:765,],nochebuena[1:765,], efecto_feriado1[1:765,],
efecto_feriado2[1:765,], doble_efecto_feriado1[1:765,],
doble_efecto_feriado2[1:765,], faltante1[1:765,], faltante2[1:765,],
faltante3[1:765,], faltante4[1:765,], faltante5[1:765,], outlier7[1:765,],
outlier17[1:765,], outlier52[1:765,], outlier89[1:765,], outlier101[1:765,],
outlier150[1:765,], outlier433[1:765,], outlier637[1:765,],
outlier650[1:765,], outlier701[1:765,], outlier706p[1:765,])

seriepart155=serie[1:765]
seriepart155=ts(seriepart155)

m155=arima(seriepart155,
method="ML",order=c(1,0,0),seasonal=list(order=c(0,1,1),period=5),xreg=reg155)

pred155=predict(m155,n.ahead=5,newxreg=newreg5)
datos155=serie[766:770]

pred155
$pred
Time Series:
Start = 766
End = 770
Frequency = 1
      row1      row2      row3      row4      row5
1026.8226  863.4432  787.0335  780.9772  827.4018

$se
Time Series:
Start = 766
End = 770
Frequency = 1
[1] 100.4120 115.6827 120.2601 121.7209 122.1952

datos155
968    799    716    738    788

##### ANTEPENULTIMA SEMANA #####
newreg5 <- matrix(c(0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0,
0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0,
0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0,
0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0), nrow = 5, ncol=22, byrow=TRUE, dimnames =
list(c("row1", "row2", "row3", "row4", "row5"), c("C.1", "C.2", "C.3", "C.4",
"C.5", "C.6", "C.7", "C.8", "C.9", "C.10", "C.11", "C.12", "C.13", "C.14",
"C.15", "C.16", "C.17", "C.18", "C.19", "C.20", "C.21", "C.22")))

```

```

reg255=cbind(feriados[1:760,],nochebuena[1:760,],      efecto_feriado1[1:760,],
efecto_feriado2[1:760,],                                doble_efecto_feriado1[1:760,],
doble_efecto_feriado2[1:760,],      faltante1[1:760,],      faltante2[1:760,],
faltante3[1:760,], faltante4[1:760,], faltante5[1:760,], outlier7[1:760,],
outlier17[1:760,], outlier52[1:760,], outlier89[1:760,], outlier101[1:760,],
outlier150[1:760,],      outlier433[1:760,],      outlier637[1:760,],
outlier650[1:760,], outlier701[1:760,], outlier706p[1:760,])

seriepart255=serie[1:760]
seriepart255=ts(seriepart255)

m255=arima(seriepart255,
method="ML",order=c(1,0,0),seasonal=list(order=c(0,1,1),period=5),xreg=reg255)

pred255=predict(m255,n.ahead=5,newxreg=newreg5)
datos255=serie[761:765]

pred255
$pred
Time Series:
Start = 761
End = 765
Frequency = 1
      row1      row2      row3      row4      row5
1060.6665  891.2164  806.1370  790.5005  838.4137

$se
Time Series:
Start = 761
End = 765
Frequency = 1
[1] 100.6027 115.8203 120.3581 121.7983 122.2633

datos255
1132      859      786      842      833

#####
##### PREDICCION DENTRO DE LA MUESTRA A 1 PASO #####
##### EFECTO CALENDARIO #####

## EF1 ##
newregEF1 <- matrix(c(0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0),
nrow = 1, ncol=18, byrow=TRUE, dimnames = list(c("row1"), c("C.1", "C.2",
"C.3", "C.4", "C.5", "C.6", "C.7", "C.8", "C.9", "C.10", "C.11", "C.12",
"C.13", "C.14", "C.15", "C.16", "C.17", "C.18")))
newregEF1
      C.1 C.2 C.3 C.4 C.5 C.6 C.7 C.8 C.9 C.10 C.11 C.12 C.13 C.14 C.15 C.16 C.17 C.18
row1  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0

regEF1=cbind(feriados[1:616,],nochebuena[1:616,],      efecto_feriado1[1:616,],
efecto_feriado2[1:616,],                                doble_efecto_feriado1[1:616,],
doble_efecto_feriado2[1:616,],      faltante1[1:616,],      faltante2[1:616,],
faltante3[1:616,], faltante4[1:616,], faltante5[1:616,], outlier7[1:616,],
outlier17[1:616,], outlier52[1:616,], outlier89[1:616,], outlier101[1:616,],
outlier150[1:616,], outlier433[1:616,])

seriepartEF1=serie[1:616]
seriepartEF1=ts(seriepartEF1)

```

```

mEF1=arima(seriepartEF1,
method="ML",order=c(1,0,0),seasonal=list(order=c(0,1,1),period=5),xreg=regEF1)

predEF1=predict(mEF1,n.ahead=1,newxreg=newregEF1)
datosEF1=serie[617:617]
predEF1
$pred
Time Series:
Start = 617
End = 617
Frequency = 1
  row1
1033.101

$se
Time Series:
Start = 617
End = 617
Frequency = 1
[1] 96.56609

datosEF1
1036

## EF2 ##
newregEF2 <- matrix(c(0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0),
nrow = 1, ncol=18, byrow=TRUE, dimnames = list(c("row1"), c("C.1", "C.2",
"C.3", "C.4", "C.5", "C.6", "C.7", "C.8", "C.9", "C.10", "C.11", "C.12",
"C.13", "C.14", "C.15", "C.16", "C.17", "C.18")))
newregEF2
  C.1 C.2 C.3 C.4 C.5 C.6 C.7 C.8 C.9 C.10 C.11 C.12 C.13 C.14 C.15 C.16 C.17 C.18
row1  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0

regEF2=cbind(feriados[1:617,],nochebuena[1:617,], efecto_feriado1[1:617,],
efecto_feriado2[1:617,], doble_efecto_feriado1[1:617,],
doble_efecto_feriado2[1:617,], faltante1[1:617,], faltante2[1:617,],
faltante3[1:617,], faltante4[1:617,], faltante5[1:617,], outlier7[1:617,],
outlier17[1:617,], outlier52[1:617,], outlier89[1:617,], outlier101[1:617,],
outlier150[1:617,], outlier433[1:617,])

seriepartEF2=serie[1:617]
seriepartEF2=ts(seriepartEF2)

mEF2=arima(seriepartEF2,
method="ML",order=c(1,0,0),seasonal=list(order=c(0,1,1),period=5),xreg=regEF2)

predEF2=predict(mEF2,n.ahead=1,newxreg=newregEF2)
datosEF2=serie[618:618]

predEF2
$pred
Time Series:
Start = 618
End = 618
Frequency = 1
  row1
848.7293

$se
Time Series:

```

```

Start = 618
End = 618
Frequency = 1
[1] 96.48691

datosEF2
820

## DEF1 ##
newregDEF1 <- matrix(c(0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0),
nrow = 1, ncol=17, byrow=TRUE, dimnames = list(c("row1"), c("C.1", "C.2",
"C.3", "C.4", "C.5", "C.6", "C.7", "C.8", "C.9", "C.10", "C.11", "C.12",
"C.13", "C.14", "C.15", "C.16", "C.17")))
newregDEF1
  C.1 C.2 C.3 C.4 C.5 C.6 C.7 C.8 C.9 C.10 C.11 C.12 C.13 C.14 C.15 C.16 C.17
row1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0

regDEF1=cbind(feriados[1:472,],nochebuena[1:472,], efecto_feriado1[1:472,],
efecto_feriado2[1:472,], doble_efecto_feriado1[1:472,],
doble_efecto_feriado2[1:472,], faltante1[1:472,], faltante2[1:472,],
faltante3[1:472,], faltante4[1:472,], outlier7[1:472,], outlier17[1:472,],
outlier52[1:472,], outlier89[1:472,], outlier101[1:472,], outlier150[1:472,],
outlier433[1:472,])

seriepartDEF1=serie[1:472]
seriepartDEF1=ts(seriepartDEF1)

mDEF1=arima(seriepartDEF1,
method="ML",order=c(1,0,0),seasonal=list(order=c(0,1,1),period=5),xreg=regDEF1)

predDEF1=predict(mDEF1,n.ahead=1,newxreg=newregDEF1)
datosDEF1=serie[473:473]

predDEF1
$pred
Time Series:
Start = 473
End = 473
Frequency = 1
  row1
1336.205

$se
Time Series:
Start = 473
End = 473
Frequency = 1
[1] 97.55443

datosDEF1
1178

## DEF2 ##
newregDEF2 <- matrix(c(0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0),
nrow = 1, ncol=17, byrow=TRUE, dimnames = list(c("row1"), c("C.1", "C.2",
"C.3", "C.4", "C.5", "C.6", "C.7", "C.8", "C.9", "C.10", "C.11", "C.12",
"C.13", "C.14", "C.15", "C.16", "C.17")))
newregDEF2
  C.1 C.2 C.3 C.4 C.5 C.6 C.7 C.8 C.9 C.10 C.11 C.12 C.13 C.14 C.15 C.16 C.17
row1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0

```

```

regDEF2=cbind(feriados[1:473,],nochebuena[1:473,],      efecto_feriado1[1:473,],
efecto_feriado2[1:473,],                                doble_efecto_feriado1[1:473,],
doble_efecto_feriado2[1:473,],      faltante1[1:473,],      faltante2[1:473,],
faltante3[1:473,], faltante4[1:473,], outlier7[1:473,], outlier17[1:473,],
outlier52[1:473,], outlier89[1:473,], outlier101[1:473,], outlier150[1:473,],
outlier433[1:473,])

seriepartDEF2=serie[1:473]
seriepartDEF2=ts(seriepartDEF2)

mDEF2=arima(seriepartDEF2,
method="ML",order=c(1,0,0),seasonal=list(order=c(0,1,1),period=5),xreg=regDEF)

predDEF2=predict(mDEF2,n.ahead=1,newxreg=newregDEF2)
datosDEF2=serie[474:474]

predDEF2
$pred
Time Series:
Start = 474
End = 474
Frequency = 1
  row1
847.3722

$se
Time Series:
Start = 474
End = 474
Frequency = 1
[1] 97.62792

datosDEF2
970

```

```
#####
##### PREDICCIÓN DENTRO DE LA MUESTRA A 5 PASO #####
##### EFECTO CALENDARIO #####

## EF ##
newregEF <- matrix(c(1,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 1,0,0,0,0,
0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,1, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0,
0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0), nrow = 5,
ncol=18, byrow=TRUE, dimnames = list(c("row1", "row2", "row3", "row4",
"row5"), c("C.1", "C.2", "C.3", "C.4", "C.5", "C.6", "C.7", "C.8", "C.9",
"C.10", "C.11", "C.12", "C.13", "C.14", "C.15", "C.16", "C.17", "C.18")))
newregEF
      C.1 C.2 C.3 C.4 C.5 C.6 C.7 C.8 C.9 C.10 C.11 C.12 C.13 C.14 C.15 C.16 C.17 C.18
row1  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
row2  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
row3  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0
row4  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
row5  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0

regEF=cbind(feriados[1:615,],nochebuena[1:615,], efecto_feriado1[1:615,],
efecto_feriado2[1:615,], doble_efecto_feriado1[1:615,],
doble_efecto_feriado2[1:615,], faltante1[1:615,], faltante2[1:615,],
faltante3[1:615,], faltante4[1:615,], faltante5[1:615,], outlier7[1:615,],
outlier17[1:615,], outlier52[1:615,], outlier89[1:615,], outlier101[1:615,],
outlier150[1:615,], outlier433[1:615,])

seriepartEF=serie[1:615]
seriepartEF=ts(seriepartEF)

mEF=arima(seriepartEF,
method="ML",order=c(1,0,0),seasonal=list(order=c(0,1,1),period=5),xreg=regEF)

predEF=predict(mEF,n.ahead=5,newxreg=newregEF)
datosEF=serie[616:620]

predEF
$pred
Time Series:
Start = 616
End = 620
Frequency = 1
      row1      row2      row3      row4      row5
221.3353 1149.6408  907.8489  864.1930  903.6262

$se
Time Series:
Start = 616
End = 620
Frequency = 1
[1] 96.24882 108.92106 112.22075 113.12954 113.38329

datosEF
0      1036      820      778      962

## DEF ##
newregDEF <- matrix(c(1,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,1,0,0, 0,0,0,0,0,
0,0,0,0,0, 0,0,0,0,0, 0,0,0,1,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,1,0,0,0,
0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0, 0,0,0,0,0), nrow = 5, ncol=17,
byrow=TRUE, dimnames = list(c("row1", "row2", "row3", "row4", "row5"),
```

```

c("C.1", "C.2", "C.3", "C.4", "C.5", "C.6", "C.7", "C.8", "C.9", "C.10",
"C.11", "C.12", "C.13", "C.14", "C.15", "C.16", "C.17"))
newregDEF
  C.1 C.2 C.3 C.4 C.5 C.6 C.7 C.8 C.9 C.10 C.11 C.12 C.13 C.14 C.15 C.16 C.17
row1  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
row2  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
row3  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0
row4  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0
row5  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0

regDEF=cbind(feriados[1:470,],nochebuena[1:470,],          efecto_feriado1[1:470,],
efecto_feriado2[1:470,],                                doble_efecto_feriado1[1:470,],
doble_efecto_feriado2[1:470,],                          faltante1[1:470,],          faltante2[1:470,],
faltante3[1:470,],  faltante4[1:470,],  outlier7[1:470,],  outlier17[1:470,],
outlier52[1:470,],  outlier89[1:470,],  outlier101[1:470,], outlier150[1:470,],
outlier433[1:470,])

seriepartDEF=serie[1:470]
seriepartDEF=ts(seriepartDEF)

mDEF=arima(seriepartDEF,
method="ML",order=c(1,0,0),seasonal=list(order=c(0,1,1),period=5),xreg=regDEF)

predDEF=predict(mDEF,n.ahead=5,newxreg=newregDEF)
datosDEF=serie[471:475]

predDEF
$pred
Time Series:
Start = 471
End = 475
Frequency = 1
      row1      row2      row3      row4      row5
222.60643  33.29819 1353.04483  929.48453  905.59412

$se
Time Series:
Start = 471
End = 475
Frequency = 1
[1]  97.14044 110.80189 114.59616 115.71406 116.04850

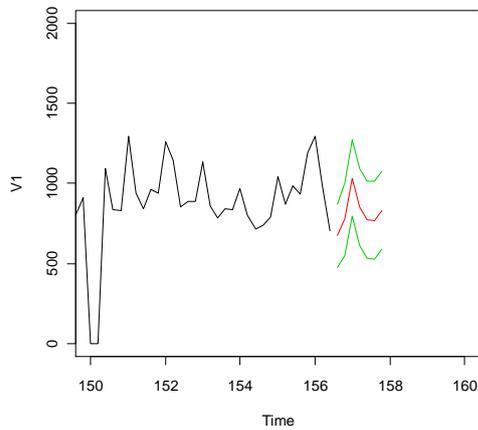
datosDEF
0      0      1178      970      980

```



```
pred=predict(m22p,n.ahead=7,newxreg=newreg2)
```

```
plot(serie,xlim=c(150,160),ylim=c(0,2000))  
lines(pred$pred,col=2)  
lines(pred$pred + 1.96*pred$se,col=3)  
lines(pred$pred - 1.96*pred$se,col=3)
```



```
pred
```

```
$pred
```

```
Time Series:
```

```
Start = c(156, 4)
```

```
End = c(157, 5)
```

```
Frequency = 5
```

row1	row2	row3	row4	row5	row6	row7
672.4341	776.4620	1032.7389	853.1576	772.9267	768.6951	832.1350

```
$se
```

```
Time Series:
```

```
Start = c(156, 4)
```

```
End = c(157, 5)
```

```
Frequency = 5
```

```
[1] 100.6499 116.2711 121.0472 122.6033 123.1194 123.8396 124.0796
```