UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

DESARROLLO DE UN MODELO DE PREDICCIÓN DE FUGA DE SALDOS DE CUENTAS CORRIENTES EN UN BANCO DEL PAÍS

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

CHRISTIAN ANDRÉS RAMÍREZ FLORES

PROFESOR GUÍA:

CARLOS PULGAR ARATA

PROFESOR CO-GUÍA:

CARLOS REYES RUBIO

COMISIÓN:

JUAN PABLO ROMERO GODOY

SANTIAGO DE CHILE

2023

RESUMEN DE LA MEMORIA
PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INDUSTRIAL
POR: CHRISTIAN ANDRÉS RAMÍREZ FLORES
AÑO: 2023
PROF. GUÍA: CARLOS PULGAR ARATA

DESARROLLO DE UN MODELO DE PREDICCIÓN DE FUGA DE SALDOS DE CUENTAS CORRIENTES EN UN BANCO DEL PAÍS

El objetivo del trabajo es realizar un modelo de predicción de fuga de saldos de cuentas corrientes en un banco del país, comprender qué variables afectan al comportamiento de estos y cuantificar la capacidad predictiva del modelo. Esto traerá como resultado la capacidad de proyectar los saldos de cuentas corrientes de los clientes del banco, ayudando a la gestión diaria de la mesa de dinero y al cumplimiento normativo asociado a la liquidez.

Existe todo un contexto de incertidumbre local y mundial, con una pandemia afectando a la economía, altos niveles de inflación y tasas históricamente altas, lo cual impacta en el comportamiento del saldo en cuentas corrientes y saldos vistas.

La realización de este trabajo se justifica en la necesidad de poder proyectar y anticiparse a fugas que puedan impactar a la liquidez del banco, lo cual se vuelve clave en periodos de alta inestabilidad financiera y escasa holgura en los límites de riesgo internos y normativos.

La metodología aplicada está relacionada con los modelos ARIMA y sus derivados de tal forma de identificar las variables intrínsecas y variables externas, para obtener una predicción lo más ajustada a la serie real de saldos posible. También se realiza un análisis estadístico a la data, con la finalidad de dar robustez a los resultados obtenidos, los cuales pasan por una serie de tests que validen el comportamiento resultante y justifiquen la elección de los parámetros utilizados. La validación de los resultados se realiza sobre la serie de datos que no se utiliza para modelar, de tal forma de contrastar empíricamente contra los mismos datos fuera de muestra.

Los resultados más relevantes de este trabajo son que no todos los segmentos de clientes responden de la misma manera a las variables externas, los cuales incluso algunos tienen peor performance al incluir variables exógenas. Además, el impacto que tienen algunas variables no fue el esperado a priori.

A modo de trabajo futuro, se vuelve muy relevante extender el horizonte de tiempo en el que se desarrolla esta memoria; incluir variables específicas a los clientes y evaluar distintos modelos de predicción al propuesto en este trabajo.

Tabla de contenido

Capítulo 1

Introducción

1.1. Motivación

La medición adecuada de la liquidez de un banco es de vital importancia, no tan solo para la institución en sí, si no que, para el sistema financiero completo, esto debido a que solo basta que un banco no sea capaz de cumplir sus obligaciones con sus clientes para que se genere una desconfianza generalizada a todo el sistema, la cual puede traer graves repercusiones financieras al país. Dicho lo anterior, es por lo cual la Comisión para el Mercado Financiero (en adelante "CMF") constantemente elabora, revisa y actualiza normativas que permiten tener un sistema financiero más robusto y confiable, entre las cuales se encuentran los distintos capítulos de la Recopilación Actualizada de Normas (en adelante "RAN"), manual de normas bajo la cual todos los participantes financieros del país deben cumplir.

Dentro de las instituciones bancarias existe un *trade-off* entre las áreas de riesgo y las distintas mesas de dineros, la cual consiste en que el área de riesgo tiene como objetivo principal la medición y limitación de potenciales riesgos a incumplir la normativa vigente y los distintos indicadores de riesgo internos. Por otro lado, el objetivo principal de la mesa de dinero es rentabilizar lo más posible el dinero destinado a inversión que posee el banco, el cual es fundamental en las utilidades que el banco genera año a año. Es por esto que ambas áreas trabajan en conjunto, armonizando sus respectivos objetivos a favor de una gestión adecuada del riesgo de liquidez.

Una forma de mejorar esta gestión de la liquidez es poder predecir movimientos futuros que puedan causar salidas de flujos de dinero no previstos por las áreas respectivas. Uno de los productos financieros que más se han visto afectados por las condiciones macroeconómicas del país y del mundo son las cuentas corrientes. Esto debido a que, a partir del brote de Coronavirus en el año 2020, el gobierno de Chile realizó importantes ayudas económicas con la finalidad de hacer frente a las restricciones de movilidad impuestas: retiros de fondos previsionales, transferencias directas al bolsillo de las personas, entre otras. Adicionalmente, el contexto laboral y macroeconómico han impactado el dinero de las personas: aumento del desempleo, aumento histórico de la inflación, aumento de la Tasa de Política Monetaria (en adelante "TPM") por parte del Banco Central de Chile, entre otras.

Actualmente la mesa de dinero gestiona a sus clientes más importantes, los cuales son clientes mayoristas que tienen ejecutivos dedicados a cada uno de ellos, por lo cual poseen el conocimiento de que movimientos realizarán durante los siguientes días. Sin embargo, no existe

una gestión de clientes más pequeños, debido a una restricción lógica de capacidad y tiempo, por lo cual la mesa de dinero se enfrenta a la caída (o aumento) de los saldos en las cuentas corrientes el mismo día que disponen para gestionar, esto una vez que el área contable deja disponible esta información en las bases de datos del banco y posteriormente que las áreas de riesgo validan esta información.

Sumado a esta ausencia de gestión de sus clientes más pequeños, también existe la nula predicción de movimientos de flujos futuros. Esta falta de proyección impide anticiparse a caídas bruscas en el saldo de cuentas corrientes, lo cual puede tener un grave impacto en los límites de liquidez si es que no se dispone la suficiente holgura al momento de esta caída, obligando a la mesa de dinero a financiarse rápidamente, incluso con instrumentos financieros que sean más costosos, de tal forma de cumplir los requerimientos normativos a cambio de una disminución en los márgenes. Adicionalmente, con una buena proyección no solo se podrá anticipar a caídas en los saldos, sino que también existirá la posibilidad de adelantarse a aumentos en los saldos, permitiendo gestionar de mejor manera el dinero que se dispone en exceso y no estaba contemplado, ayudando a cumplir de mejor forma el objetivo de la mesa de dinero, que es maximizar la rentabilidad.

1.2. Alcances y Objetivo General

El objetivo de este trabajo es desarrollar y proponer un modelo de predicción de saldos en cuentas corrientes y saldos vistas para un banco del país, para cada uno de los segmentos de clientes definidos internamente por la institución. En particular se evaluarán distintas metodologías basadas en el modelo ARIMA (y sus derivados tales como modelos SARIMA, ARIMAX y sus combinaciones), buscando aquellos que permitan explicar de mejor forma el comportamiento de los distintos segmentos de clientes que el banco posee.

El alcance de este trabajo se limita al producto de cuentas corrientes y saldos vistas de clientes minoristas, segmentados de acuerdo con definiciones internas de la institución, dentro de un horizonte temporal que abarca de enero de 2019 a diciembre de 2021.

1.3. Objetivos Específicos

Los objetivos específicos de este trabajo son los siguientes:

- Comprender el comportamiento histórico de los flujos de cuentas corrientes en el banco y las variables internas y externas que influyen en él

- Evaluar cuáles son las variables macroeconómicas que más afectan al comportamiento de los saldos de cuentas corrientes

- Modelar y predecir para cada segmento de clientes

- Medir la capacidad predictiva del modelo para cada segmento

1.4. Estructura de la Memoria

La estructura utilizada en este documento es la siguiente:

- Capítulo 1. Introducción: Corresponde a la descripción del tema, la motivación de éste y los alcances y objetivos del trabajo realizado.

- Capítulo 2. Antecedentes: Corresponde a la revisión bibliográfica o antecedentes. En este capítulo se explican los conceptos necesarios para la comprensión y contextualización del trabajo.

- Capítulo 3. Resultados: Corresponde a la exposición de los resultados de los modelos evaluados, con sus respectivos análisis de errores de predicción.

- Capítulo 4. Discusión y conclusiones: Se enumeran las conclusiones del trabajo realizado y se proponen trabajos a realizar en el futuro

Capítulo 2

Antecedentes

2.1. Marco Teórico

2.1.1. Riesgo de Liquidez

Como primer acercamiento el mundo de la liquidez es necesario comprender el por qué se exigen reportes normativos de indicadores de liquidez y qué consecuencias puede provocar un mal manejo de estos indicadores.

Un riesgo clave que enfrentan las entidades bancarias es el riesgo de liquidez. Este riesgo corresponde a la imposibilidad de:

- Cumplir oportunamente con las obligaciones contractuales.
- Liquidar posiciones sin pérdidas significativas ocasionadas por volúmenes anormales de operación.
- Evitar sanciones regulatorias por incumplimiento de índices normativos.
- Financiar de forma competitiva la actividad comercial y de tesorería.

Se distinguen dos fuentes de riesgo:

- Endógenas: situaciones de riesgo derivadas de decisiones corporativas controlables.
    - Alta liquidez alcanzada por una reducida base de activos líquidos o descalces de activos y pasivos significativos.
    - Baja diversificación o alta concentración de activos financieros y comerciales en término de emisores, plazos y factores de riesgo.

- Deficiente gestión de coberturas de valor, flujos o crédito en términos de la eficiencia de la cobertura, correlación de los cambios de valor, ratios de sensibilidad del elemento cubierto y el derivado, entre otras.

- Efectos reputacionales corporativos adversos que se traduzcan en acceso no competitivo a financiamiento o falta de éste.

- Exógenas: situaciones de riesgo producto de movimientos de los mercados financieros no controlables.

2.1.1.1. Reporte C46 o Descalces a Plazo

Este reporte busca medir los descalces a plazos de las distintas entidades bancarias para determinados plazos. Así, según el capítulo III.B.2 del Banco Central de Chile, "la posición de liquidez se medirá y controlará a través de la diferencia entre los flujos de efectivo de egresos y de ingresos, dentro y fuera del balance, para un determinado plazo o banda temporal". La fórmula que representa lo anterior es la siguiente:

$$Descalce_t = \sum_i Pasivos_{i,t} - \sum_i Activos_{i,t}$$

Donde: t indica el plazo al cual se determinará el descalce

i representa a todos los productos del banco que puedan ser catalogados como activos o pasivos respectivamente

Según el capítulo 12-20 de la RAN, los límites normativos de los descalces son[1]:

- La suma de todos los descalces de plazos hasta 30 días no podrá ser superior al capital básico
- El mismo requisito deberá cumplirse considerando solo los flujos en moneda extranjera
- La suma de los descalces de plazo hasta 90 días no podrá ser superior a dos veces el capital básico

Adicionalmente, la normativa define una serie de plazos de tal forma que el reporte y la lectura se este sea homogénea para todas las entidades. La siguiente tabla la relación entre las bandas normativas definidas por la CMF y su correspondiente símil en cuanto a días calendarios.

[1] Norma modificada a partir de 2022, la cual no afecta al periodo analizado en este trabajo

Código	Banda Temporal
101	Primer día
102	Segundo día
103	Tercer día
104	Cuarto día
105	Quinto día
106	Sexto día
107	Séptimo día
205	Desde 8 hasta 15 días.
310	Desde 16 hasta 30 días.
415	Desde 31 hasta 60 días.
520	Desde 61 hasta 90 días.
625	Desde 91 hasta 180 días.
730	Desde 181 días hasta un año.
831	Mayor a 1 año

Tabla 1. Tabla 82 del Manual de Sistemas de Información de Bancos referente a bandas temporales

Con respecto a los productos financieros que son considerados dentro del descalce a plazos, el Manual de Sistemas de Información también define categorías en las cuales se agrupan los distintos productos que ofrezcan los bancos, con el objetivo de estandarizar el reporte. Esta categorización se encuentra en **¡Error! No se encuentra el origen de la referencia.** para una r evisión más detallada.

2.1.2. Métodos de Proyección

2.1.2.1. ARIMA

El modelo ARIMA (Autoregressive Integrated Moving Average en inglés) permite pronosticar series de tiempo. Está compuesta por una componente autorregresiva (AR), otra integrada (I) y media móvil (MA), en donde los parámetros (p, d, q) indican el número de términos autorregresivos, número de diferencias no estacionarias para alcanzar la estacionariedad y el número de rezagos del pronóstico de errores, respectivamente. Su forma general es la siguiente:

$$Y_t = -(\Delta^d Y_t - Y_t) + \phi_o + \sum_{i=1}^{p} \phi_i \Delta^d Y_{t-i} - \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t$$

Se define un modelo como autorregresivo si la variable endógena de un período t es explicada por las observaciones de ella misma correspondientes a períodos anteriores añadiéndose, como en los modelos estructurales, un término de error. En el caso de procesos estacionarios con distribución normal, la teoría estadística de los procesos estocásticos dice que,

bajo determinadas condiciones previas, toda Y_t puede expresarse como una combinación lineal de sus valores pasados (parte sistemática) más un término de error (innovación).

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 sucesión de errores correspondientes a períodos precedentes, ponderados convenientemente.

2.1.2.2. SARIMA

El modelo SARIMA (Seasonal Autoregressive Integrated Moving Average en inglés) o ARIMA estacional, es una extensión del modelo ARIMA la cual añade información de serie de tiempo con una componente estacional. Este modelo agrega 3 nuevos parámetros para especificar la autorregresión (AR), diferenciación (I) y la media móvil (MA) para la componente estacional de la serie, así como se adiciona un parámetro de periodo de la estacionalidad.

$$(1-\phi_1 B)\ (1-\Phi_1 B^4)\ (1-B)\ (1-B^4)y_t \ = \ (1+\theta_1 B)\ (1+\Theta_1 B^4)e_t.$$

(Non-seasonal AR(1)) (Seasonal AR(1)) (Non-seasonal difference) (Seasonal difference) (Non-seasonal MA(1)) (Seasonal MA(1))

2.1.2.3. ARIMAX

El modelo ARIMAX es simplemente el modelo ARIMA incluyendo variables exógenas a las variables explicativas. Así, se incorporan nuevos parámetros z_{t-i} exógenos e independientes de la serie original y_t, los cuales poseen sus respectivos coeficientes θ_i que permiten cuantificar el impacto marginal de la respectiva variable exógena sobre la variable dependiente.

$$y_t = \beta x_t + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} - \theta_1 z_{t-1} - \cdots - \theta_q z_{t-q} + z_t$$

2.1.3. Test estadísticos

2.1.3.1. Test Dickey-Fuller y Dickey-Fuller aumentado

El test Dickey-Fuller evalúa la hipótesis nula de que existe presencia de una raíz unitaria en un modelo de series de tiempo autorregresivo. Respecto a la hipótesis alternativa, esta es que la serie es estacionaria.

2.1.3.2. Test Kwiatkowski–Phillips–Schmidt–Shin

El test Kwiatkowski–Phillips–Schmidt–Shin o también conocido como KPSS es un test utilizado para evaluar la hipótesis nula de que una serie de tiempo es estacionaria alrededor de una tendencia determinística, contra la alternativa de la existencia de una raíz unitaria.

A diferencia de la mayoría de los test de raíz unitaria, la presencia de una raíz unitaria no es la hipótesis nula, sino que es la hipótesis alternativa. Adicionalmente, en el test KPSS la ausencia de una raíz unitaria no es prueba de estacionariedad, si no que de una tendencia estacionaria[2].

2.1.4. Métodos de error

2.1.4.1. AIC y BIC

El criterio de información de Akaike (AIC) es una medida de la calidad relativa de un modelo estadístico, para un conjunto dado de datos. Como tal, el AIC proporciona un medio para la selección del modelo.

AIC maneja un *trade-off* entre la bondad de ajuste del modelo y la complejidad del modelo. Se basa en la entropía de información: se ofrece una estimación relativa de la información perdida cuando se utiliza un modelo determinado para representar el proceso que genera los datos.

AIC no proporciona una prueba de un modelo en el sentido de probar una hipótesis nula, es decir AIC no puede decir nada acerca de la calidad del modelo en un sentido absoluto. Si todos los modelos candidatos encajan mal, AIC no dará ningún aviso de ello.

La forma general de AIC es:

$$AIC = 2k - 2\ln(L)$$

El criterio de información Bayesiano (BIC) es un criterio para la selección de modelos entre un conjunto finito de modelos. Se basa, en parte, de la función de probabilidad y que está estrechamente relacionado con el AIC.

Cuando se ajustan modelos es posible aumentar la probabilidad mediante la adición de parámetros, pero si lo hace puede resultar en sobreajuste. Tanto el BIC y AIC resuelven este

[2] La tendencia estacionaria es un proceso estocástico en el cual puede ser removida la tendencia subyacente, resultando un proceso estacionario.

problema mediante la introducción de un término de penalización para el número de parámetros en el modelo, sin embargo, el término de penalización es mayor en el BIC que en el AIC.

La forma general de BIC es:

$$\mathrm{BIC} = -2 \cdot \ln \hat{L} + k \ln(n).$$

2.1.4.2. Error cuadrático medio

El error cuadrático medio (ECM) de un estimador mide el promedio de los errores al cuadrado, es decir, la diferencia entre el estimador y lo que se estima. El ECM es una función de riesgo, correspondiente al valor esperado de la pérdida del error al cuadrado o pérdida cuadrática. La diferencia se produce debido a la aleatoriedad o porque el estimador no tiene en cuenta la información que podría producir una estimación más precisa.

La forma general del ECM es:

$$\mathrm{ECM} = \frac{1}{n} \sum_{i-1}^{n} (\hat{Y}_i - Y_i)^2.$$

Donde $\hat{Y}_i$ es el vector de predicciones y Y_i es el vector de los valores reales.

2.1.6. Autocorrelación

La autocorrelación es un término estadístico que se utiliza para describir la presencia o ausencia de correlación en los datos de las series temporales, indicando, si las observaciones pasadas influyen en las actuales. Por tanto, se puede decir que la autocorrelación hace referencia cuando los valores que toman una variable en el tiempo no son independientes entre sí, sino que un valor determinado depende de los valores anteriores.

Para medir la autocorrelación se suele usar la función de autocorrelación y la función de autocorrelación parcial.

La función de autocorrelación (ACF) mide la correlación entre dos variables separadas por k periodos y mide el grado de asociación lineal que existe entre dos variables del mismo proceso estocástico.

La función de autocorrelación parcial (PACF) mide la correlación entre dos variables separadas por k periodos cuando no se considera la dependencia creada por los retardos intermedios existentes entre ambas y mide la autocorrelación que existe entre dos variables separadas k períodos descontando los posibles efectos debidos a variables intermedias.

2.1.7. Segmentación

La segmentación utilizada dentro de la institución bancaria en la cual se realiza este trabajo corresponde a una segmentación que agrupa a los clientes minoristas mediante una serie de características en común, las cuales son:

- Cuenta IFRS: corresponde a una clasificación dada por la CMF y representa el tipo de producto financiero.

- Moneda: indica si el producto está en moneda local (CLP) o moneda extranjera (USD). Esta división se realiza debido a que es una exigencia normativa separar por tipo de moneda al realizar modelos sobre los productos.

- Cuenta remunerada: indica si el producto financiero realiza pago de intereses (cuenta remunerada) o no (cuenta no remunerada). Intuitivamente esta división permite clasificar la importancia del cliente y su potencial comportamiento, ya que las cuentas remuneradas son productos más exclusivos de acuerdo con el nivel de ingreso del cliente.

- Clasificación: corresponde al tipo de clasificación interna que se le asigna a un cliente cuando este ingresa por primera vez al banco. Separa entre instituciones financieras, instituciones no financieras, PYME y personas naturales.

A continuación, se muestran los 6 segmentos resultantes de acuerdo con las características mencionadas previamente y sobre los cuales se realizará este trabajo:

	Nombre	**Moneda**	**Remunerada**	**Clasificación**
Segmento 1	Cuentas corrientes de otras personas jurídicas	CLP	No	Instituciones no financieras y PYME
Segmento 2	Cuentas corrientes de otras personas jurídicas	CLP	Sí	Instituciones no financieras y PYME
Segmento 3	Cuentas corrientes de otras personas jurídicas	USD	No	No financieras y PYME
Segmento 4	Cuentas corrientes de personas naturales	CLP	No y sí	Personas naturales y PYME

Segmento 5	Cuentas corrientes de personas naturales	USD	No y sí	Personas naturales y PYME
Segmento 6	Cuentas de depósitos a la vista	CLP	No y sí	Personas naturales, PYME e Instituciones no financieras

Tabla 2. Segmentos utilizados en el modelo

2.1.8. Contexto macroeconómico del país

2.1.8.1. Pandemia del Coronavirus y ayudas económicas

En diciembre de 2019 se detectan los primeros casos de Coronavirus en la ciudad de Wuhan, China. Unos meses después, el 3 de marzo de 2020 se confirma el primer caso en Chile, generando restricciones de movilidad a la población para detener el creciente contagio.

Para hacer frente a las restricciones de movilidad y su impacto en los puestos de trabajo, se propusieron una serie de ayudas económicas las cuales se listan a continuación:

- En mayo de 2020 el Congreso de Chile aprobó la ley que estipula el Ingreso Familiar de Emergencia (IFE). La ayuda se entrega mediante transferencia directa y el monto depende de las condiciones socioeconómicas de la persona, fijándose un máximo de $65.000 el primer mes, $55.000 el segundo mes y $45.000 el tercer mes. Posteriormente se extiende el IFE hasta octubre 2020.

- El 24 de julio de 2020, fue promulgada la ley que permite retirar anticipada y voluntariamente el 10% de fondos de pensión de las cuentas de ahorro individual.

- El 10 de diciembre de 2020 comenzó a regir la Ley N° 21.295 que permite a las y los afiliados del sistema realizar un segundo retiro excepcional de fondos.

- A partir de abril 2021 se entrega el IFE ampliado, destinado a los hogares pertenecientes al 80% más vulnerables según el Registro Social de Hogares.

- El 28 de abril de 2021 se publica en el diario oficial la Ley N° 21.330 que permite un tercer retiro excepcional de fondos de pensiones.

- Desde junio 2021 comenzó a regir el IFE Universal, beneficio entregado hasta noviembre del mismo año y el monto de transferencia directa depende del número de integrantes del hogar, yendo desde $88.000 para una familia con 10 integrantes hasta $177.000 si la familia está compuesta por una persona.

Los puntos mencionados corresponden a aquellas ayudas más relevantes, sin embargo, existen otras ayudas más focalizadas y de menor impacto económico como, por ejemplo: subsidios al empleo, bonos a PYME, subsidios de arriendo, préstamos solidarios, entre otros.

2.1.8.2. Impacto en la economía del país

La pandemia tuvo una serie de efectos negativos tanto en la salud como en la economía de Chile. Sumado a esto, las distintas ayudas mencionadas en la sección 2.1.8.1. Pandemia del Coronavirus y ayudas económicas, tuvieron un impacto no previsto en la economía, específicamente en la inflación.

El objetivo del Banco Central de Chile es mantener la inflación contralada entre 2% y 4%, sin embargo, como se observa en el siguiente gráfico, a mediados de 2021 ocurrió un aumento significativo en el IPC (índice que mide la inflación), superando incluso el 6%, las cifras más altas observadas desde 2015.

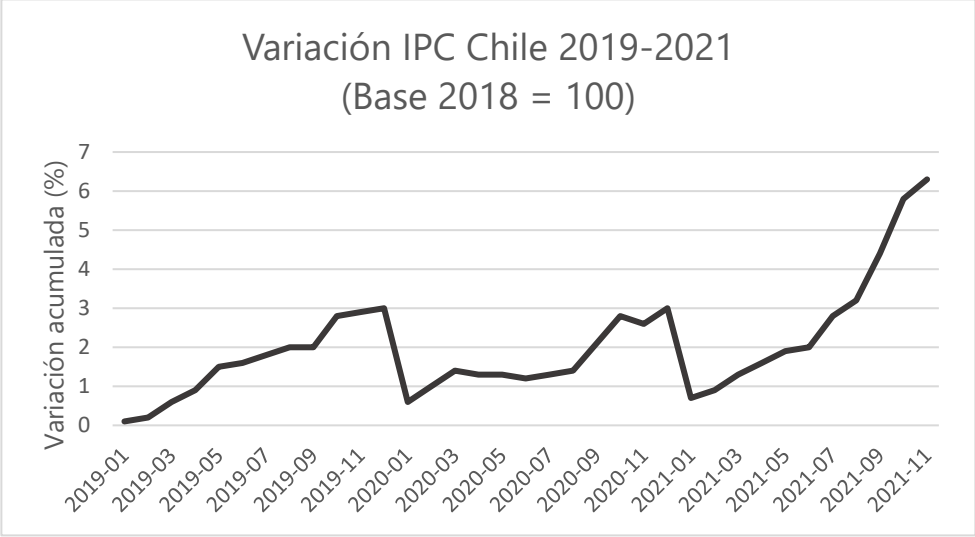

Ilustración 1. Gráfico de la variación del IPC entre 2019 y 2021

Esto tiene como impacto directo en los precios de los productos y disminuyendo el poder de compra de las personas. Para subsanar y evitar que la inflación siga aumentando descontroladamente, el Banco Central realizó un histórico aumento en la Tasa de Política Monetaria (TPM), subiendo 125 puntos básicos en la reunión llevada a cabo en octubre de 2021.

2.1.8.3. Variabilidad en saldos de cuentas corrientes

Los distintos fenómenos explicados anteriormente han tenido un fuerte efecto en la variación de los saldos de cuentas corrientes. En primera instancia, los retiros de los fondos

previsionales y las transferencias directas del Gobierno han llevado a números récords en los saldos en las cuentas de las personas.

Por otro lado, luego del fuerte aumento en la inflación y el aumento de las tasas por parte del Banco Central, no es conveniente tener el dinero en las cuentas corrientes debido a que no se ajustan por inflación, perdiendo capacidad de compra si no se invierte en otros instrumentos, llevando una fuga de los saldos que anteriormente se encontraban en números récords.

2.2. Descripción del modelo

La hipótesis que se propone en este trabajo considera que es posible proyectar de forma precisa los saldos de cuentas corrientes de los 6 segmentos utilizados en la institución financiera en la cual está inmerso este trabajo. Adicionalmente se evalúa el posible impacto que tiene el contexto macroeconómico sobre el comportamiento de los segmentos de clientes ya nombrados.

A modo de contexto, las series de saldos de cuentas corrientes se encuentran en el horizonte temporal iniciando en 2019-01-02 y finalizando en 2021-12-30. Para realizar y contrastar los modelos que se utilizan a continuación, esta data se separa en una data de entrenamiento y una data de validación. La data de entrenamiento utiliza el 80% inicial de la data, mientras que la data de validación corresponde al último 20% de la data, ordenada de forma temporal, así la separación de las fechas queda de la siguiente forma:

- Data de modelamiento: inicia en 2019-01-02 y finaliza en 2021-05-26
- Data de validación: inicia en 2021-05-27 y finaliza en 2021-12-30

El objetivo de esta separación es medir qué tan bien los modelos determinados utilizando la data de entramiento se ajustan a la data de validación, dando un indicio de la calidad de la proyección realizada.

La metodología utilizada para modelar es la siguiente:

- De forma previa, se realizan análisis generales del comportamiento de los segmentos, ya sea mediante estadística descriptiva y gráficos, sin la aplicación de ningún modelo ni alteración de la información.

- Luego, se evalúa el método ARIMA para cada uno de los segmentos, identificando los mejores parámetros mediante la exploración de las características de las series y de sus respectivas funciones de autocorrelación y de autocorrelación parcial.

- Posteriormente se evalúa la inclusión de variables estacionales, identificando, si aplica, modelos SARIMA para cada uno de los segmentos.

- Luego, se incluyen variables exógenas al modelo ARIMA, generando modelos ARIMAX. Estas variables exógenas, a priori, corresponden a:
    - Tasa de los depósitos a plazo a 30 días en moneda CLP
    - Tasa de los depósitos a plazo a 30 días en moneda USD
    - Serie de la TPM
    - Serie de la tasa de desempleo
    - Serie del IMACEC
    - Variables *dummies* que indican el día calendario

Estas proyecciones se evaluarán mediante distintos indicadores de desempeño, comparando la serie proyectada contra la serie real de saldos (el 20% final del set de datos). Los métodos para calcular el desempeño de la predicción son el AIC, BIC y error cuadrático medio, en donde un menor valor de estos indicadores indica que la proyección es más ajustada a la serie real, indicando que el modelo es mejor que el resto.

2.2.1. Herramientas utilizadas

Este trabajo se realiza utilizando el software R y su interfaz RStudio, el cual es un software de acceso libre (*open source*), debido a que es el programa que se utiliza en la institución donde se realiza este trabajo.

Para realizar los distintos gráficos y modelos se utiliza una serie de paquetes estadísticos y de modelamiento que están disponibles gratuitamente para instalar en la interfaz RStudio. Los paquetes más relevantes utilizados son:

- Tseries
- Forecast
- Ggplot2
- Psych
- Tidyr
- Urca
- Lmtest
- MASS
- Glmnet
- caret

2.2.2 Información utilizada

Para realizar este trabajo de memoria, se utilizaron datos de distintas fuentes. A continuación, se detallan las características de estos datos y sus fuentes:

- Variables internas: corresponden a datos que tienen como fuente de información la misma institución bancaria.
    - Series de saldos de cuentas corrientes: los datos se esta serie fueron obtenidos de las bases de datos de la institución bancaria en la cual se realizó este trabajo de memoria. Los datos están correctamente validados, ya que un área del banco se encarga de confirmar que la información este cargada correctamente, aplicando test para validar la calidad de la data. Estos datos corresponden a 6 segmentos de clientes distintos, los cuales se definieron previamente para la realización de modelos de estabilidad de saldos, por lo cual este trabajo de memoria se basa en estos segmentos ya definidos para trabajar en la proyección de estos saldos. El horizonte temporal utilizado va desde 2019-01-02 hasta 2021-12-30, serie de datos diario y sin considerar los fines de semana y festivos, totalizando una cantidad de 737 datos por cada segmento. Debido a la gran cantidad de saldos en los segmentos analizados, durante esta memoria se trabaja siempre en miles de millones de pesos (MMMCLP), de tal forma de simplificar la lectura de los resultados y gráficos. Cabe mencionar que los segmentos que se indican en moneda USD de igual forma se representan en MMMCLP, esto es debido a que la base de datos ya viene aplicada la conversión de USD a CLP, luego de haber aplicado el tipo de cambio a la fecha correspondiente.

- Variables externas: corresponden a datos que fueron obtenidas desde bases de datos externas a la institución bancaria.
    - Tasas de depósitos a plazo: con la finalidad de incluir variables financieras externas, se trabaja con las series de tasas de depósitos a plazo de 30 días en moneda CLP y USD. Estos datos fueron descargados de la página web del Banco Central de Chile, específicamente de la base de datos estadísticos. El horizonte temporal utilizado va desde 2019-01-02 hasta 2021-12-30, serie de datos diario y sin considerar los fines de semana y festivos, totalizando una cantidad de 737 datos por cada segmento.

    - TPM / desempleo / IPC / IMACEC: estas variables macroeconómicas se utilizan dentro de este trabajo de memoria. Estos datos fueron descargados de la página web del Banco Central de Chile, específicamente de la base de datos estadísticos. El horizonte temporal utilizado va desde enero de 2019 hasta

diciembre de 2021, serie de datos mensual, llevada a cada día del mes para hacer coincidir la forma de las series.

- Variables creadas: corresponden a variables creadas específicamente para el trabajo de esta memoria, es decir, no provienen de bases de datos internas ni externas a la institución bancaria.

    - Dummies de días: se crean variables dummies para cada día del mes, así esta variable tiene el valor de 1 si el día corresponde a su respectiva variable dummy y 0 si no corresponde.

Resultados

A continuación, se presentan los resultados obtenidos al desarrollar este trabajo. Primero se realiza una exploración preliminar de todos los segmentos, para así observar a modo general y simple el comportamiento de estos. Posteriormente se detallan los resultados segmento a segmento, lo cual es la parte central de este trabajo y posee el nivel de detalle necesario para obtener conclusiones bien respaldadas. Por último, se recopilan todos los resultados anteriores para observarlos y analizarlos de manera más ordenada y clara.

3.1. Exploración preliminar

Antes de aplicar cualquier tipo de herramienta estadística sobre la data disponible, se realiza una breve exploración de carácter preliminar, de forma que permita observar el comportamiento de la data sin ningún tipo de alteración.

En primera instancia, se realiza una inspección gráfica del comportamiento de las series dentro del periodo estudiado:

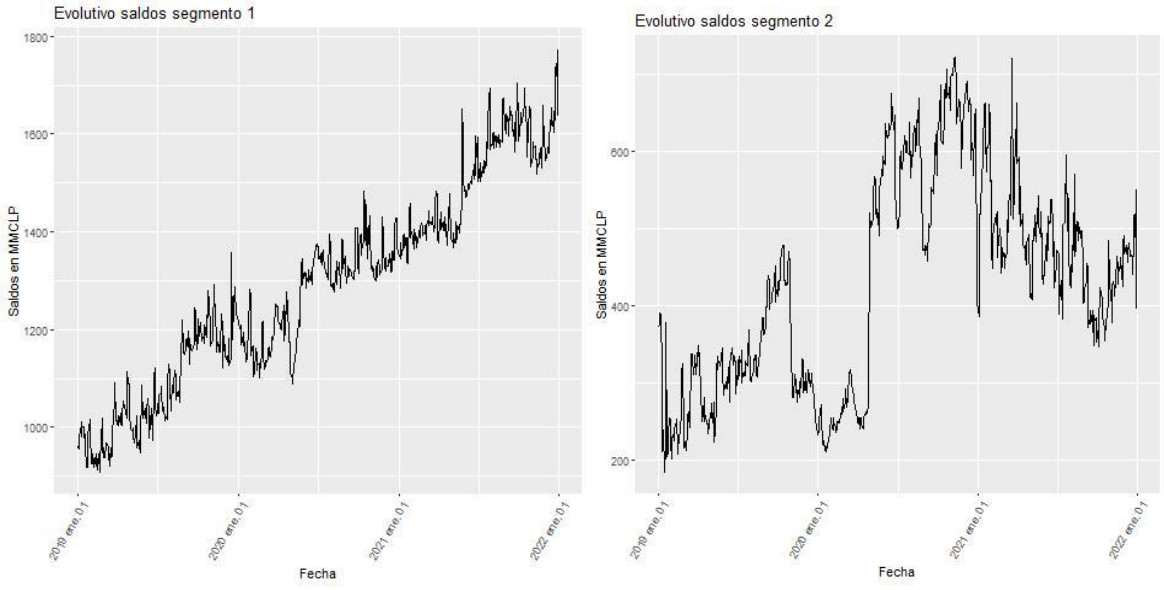

Ilustración 2. Evolutivo de los segmentos 1 y 2

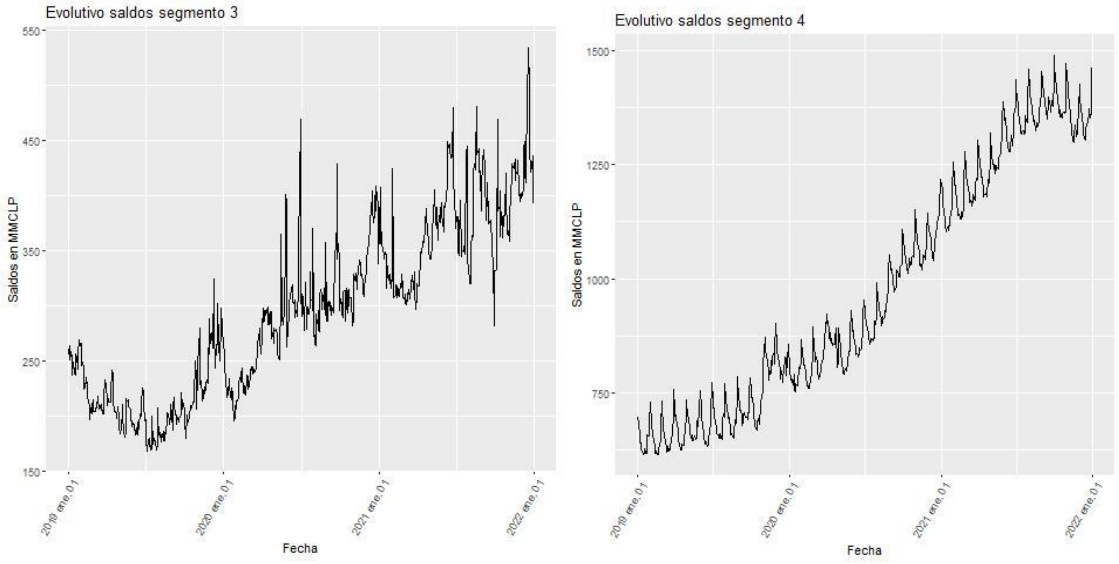

Ilustración 3. Evolutivo de los segmentos 3 y 4

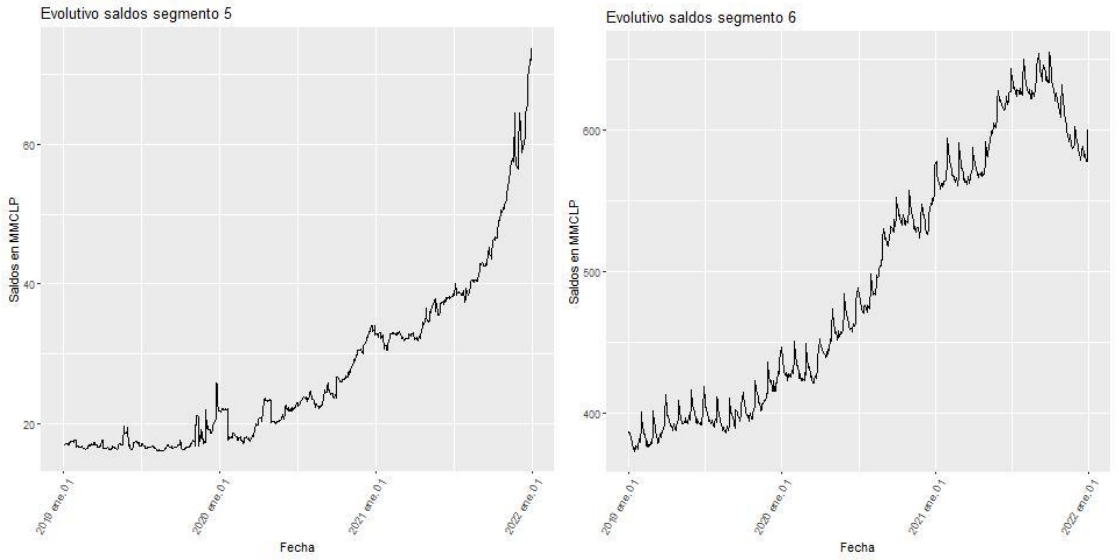

Ilustración 4. Evolutivo de los segmentos 5 y 6

De los gráficos anteriormente mostrados se observa que, si bien todos los segmentos tienen un comportamiento al alza, se observa que los segmentos 2, 4 y 6 experimentan una desaceleración y caída de sus saldos en la última parte del periodo en estudio. Lo contrario ocurre con los segmentos 1, 3 y 5, los cuales mantienen su tendencia al alza, especialmente el segmento 5 el cual aumenta su velocidad de crecimiento en el último tramo del horizonte temporal.

Otro aspecto que resaltar es la forma marcada de “serrucho” de los segmentos 4 y 6, los cuales pueden dar ciertas pistas de la existencia de una potencial estacionalidad en sus saldos, potenciada por la clara tendencia al alza que se observa de sus gráficas.

Siguiendo con el análisis preliminar, a continuación, se presenta una tabla de estadística descriptiva de los segmentos, con las herramientas más ampliamente conocidos y utilizados de la estadística.

	Mínimo	Máximo	Media	Desv. estándar	Curtosis	Asimetría
Segmento 1	907,39	1772,47	1289,04	204,95	-0,89	0,07
Segmento 2	183,18	721,62	426,80	137,96	-1,08	0,18
Segmento 3	168,17	533,98	295,13	77,12	-0,83	0,31
Segmento 4	613,74	1490,89	973,26	262,42	-1,33	0,30
Segmento 5	16,01	73,79	27,39	12,06	1,56	1,35
Segmento 6	372,60	654,79	493,70	88,92	-1,47	0,23

Tabla 3. Estadística descriptiva de los 6 segmentos

De la tabla anterior se pueden comentar las siguientes observaciones:

- Para todos los segmentos el valor de la media se encuentra relativamente centrada respecto al mínimo y máximo, a excepción del segmento 5, el cual tiene su media más cercana al valor mínimo que el resto de los segmentos. Este comportamiento se relaciona con su gráfica, en donde la mayor parte del primer tramo de la serie los saldos permanecen relativamente estables cercano al valor mínimo, lo cual cambia abruptamente en los últimos meses en donde el crecimiento pareciese tener componentes exponenciales.

- El comportamiento anterior también se ve reflejado en la curtosis, en donde el único segmento con un valor positivo es el segmento 5, lo cual indica que los valores de la serie están muy concentrados hacia la media (distribución leptocúrtica), en línea con el punto anterior debido al comportamiento inicial de la serie, la cual permaneció por varios meses cercano al valor medio.

- Con respecto al coeficiente de asimetría, para todos los segmentos el valor es positivo, lo cual indica que la distribución está sesgada hacia la derecha, lo cual se observa gráficamente con el crecimiento casi constante en todos los segmentos. El valor que se destaca es nuevamente el segmento 5, el cual tiene un coeficiente de asimetría mayor que el resto de los segmentos, explicado por el crecimiento explosivo en la última parte de la curva, concentrando aún más los datos hacia la derecha.

- Por último, en pertinente comentar que los datos reflejan el típico comportamiento de las series de tiempo: alta desviación y asimetría positiva.

De la misma forma que se analizó gráficamente las distintas series de saldos de cuentas corrientes, se realizará lo mismo, pero ahora sobre las variables exógenas que serán utilizadas más adelante en el modelamiento. Dicho lo anterior, los gráficos son los siguientes:

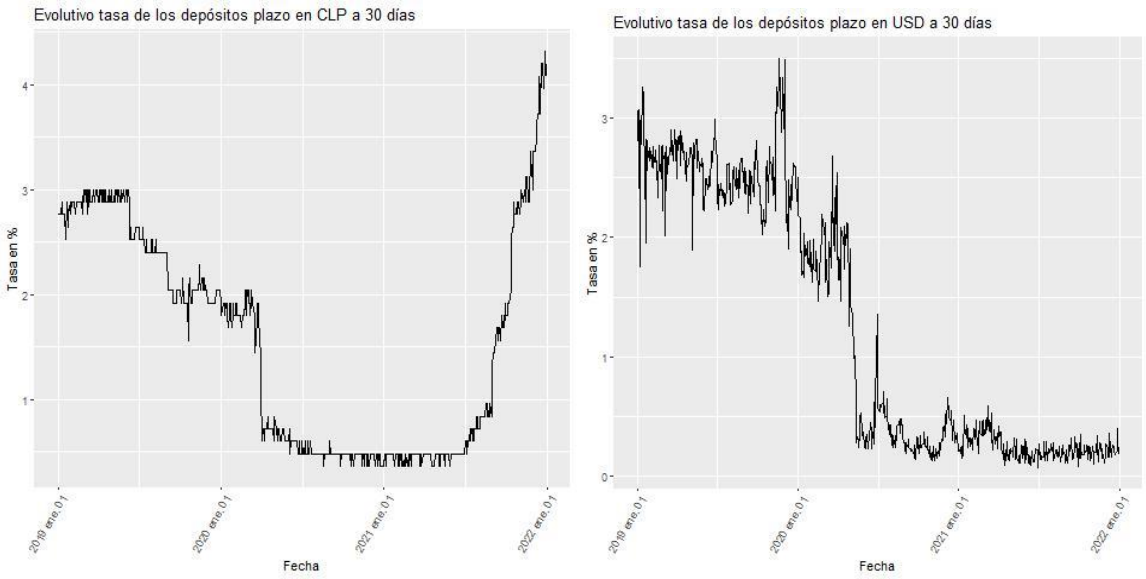

Ilustración 5. Evolutivo de la serie de tasas de depósitos a plazo a 30 días en CLP y USD

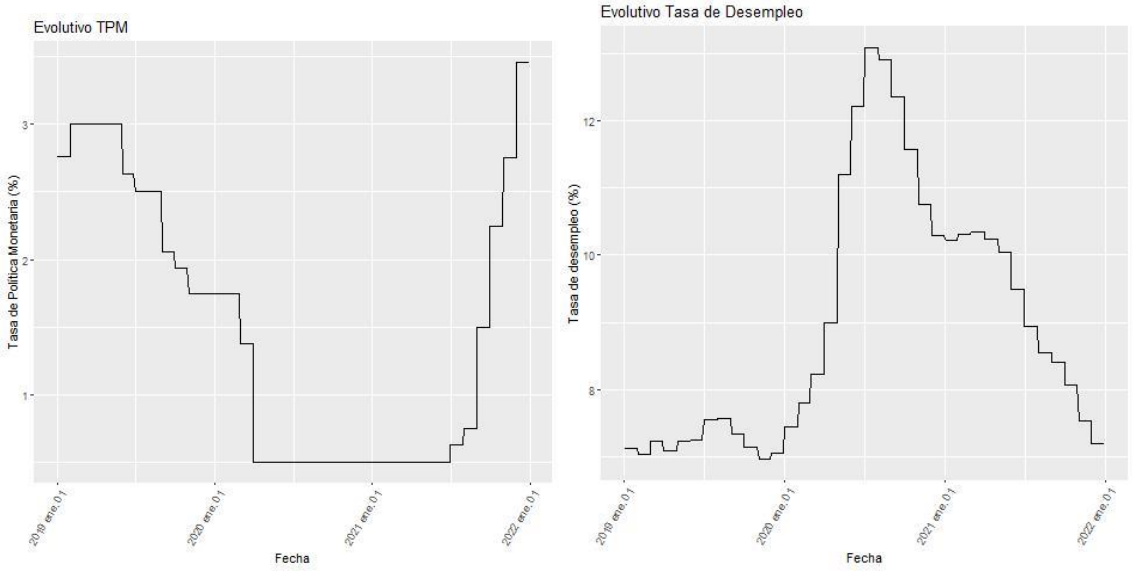

Ilustración 6. Evolutivo de las series de la TPM y tasa de desempleo

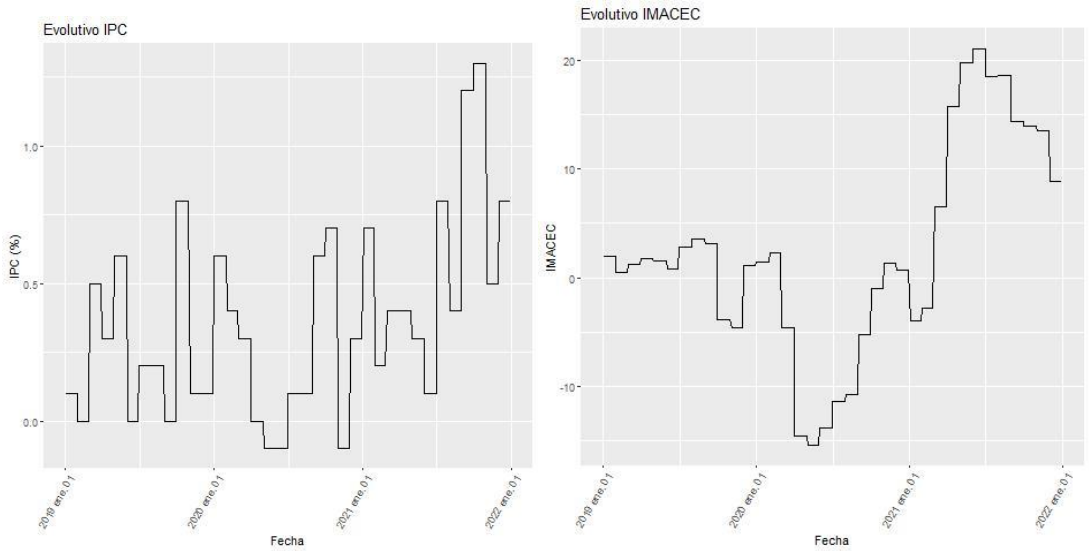

Ilustración 7. Evolutivos de las series del IPC e IMACEC

De los gráficos anteriores se puede comentar lo siguiente:

- Las series de depósitos a plazo en CLP y USD son variables con información diaria, mientras que el resto de las variables contienen información mensual, de ahí el comportamiento menos detallado que se observa en los gráficos de TPM, desempleo, IPC e IMACEC.

- Se observa un comportamiento casi calcado entre la serie de la TPM y la tasa de depósitos a plazo en CLP, esto debido a que la TPM es una tasa que guía todas las demás tasas de corto plazo del mercado.

- Respecto a los gráficos de desempleo e IMACEC, se observa un comportamiento contrario en ambas series, cuando una de ellas va a la baja, la otra va al alza y viceversa. Esto se debe a que, a un mayor índice de actividad económica, mayor empleo en el país y, por ende, provoca una disminución en la tasa de desempleo.

- La tasa de los depósitos a plazo en USD pareciera no estar relacionada con el comportamiento de ninguna de las otras variables. Una explicación de esto es que la tasa al estar en USD, no se verá afectado por variables internas del país, si no que debería verse influenciada por variables externas que no son consideradas dentro de este trabajo.

3.2. Selección de variables

Como se mencionó en la sección 2.2.2 Información utilizada, para este trabajo se utiliza una serie de variables macroeconómicas con la finalidad de mejorar la predicción de los modelos ARIMA y de evaluar el impacto y significancia de estas variables sobre los modelos. Aunque antes de utilizar las variables, es pertinente comentar el porqué de la elección de estas variables en el modelamiento.

Respecto a las tasas de los depósitos a plazo, tanto en CLP como en USD, se considera una variable relevante para el análisis ya que el costo de oportunidad de tener el dinero en las cuentas corrientes (las cuales en su mayoría no generan rentabilidad o en algunos casos específicos sí rentabilizan, pero a tasas muy bajas comparativamente) es alto dependiendo de la tasa de los depósitos, lo cual podría explicar una fuga de saldos desde las cuentas corrientes hacia los depósitos a plazo a medida que esta tasa sea lo suficientemente atractiva.

Respecto a las variables macroeconómicas seleccionadas, la justificación es la siguiente:

- TPM: Es una tasa que regula o sirve de comparación para las tasas de otros productos financieros de corto plazo, ya sea tasas de depósitos, tasas de interés de créditos de consumo y tarjetas de crédito, entre otras. Lo anterior podría impactar de forma indirecta en el comportamiento de las personas y empresas al momento de decidir mantener su dinero en las cuentas corrientes o llevarlo a otros instrumentos financieros.

- Desempleo: La tasa de desempleo puede impactar directamente en el nivel de vida de las personas y afectar fuertemente a PYMES, por lo cual esta variable podría impactar directamente en el saldo que las personas tengan en sus cuentas corrientes.

- IPC: El IPC mide el nivel de inflación que hay en el país, en donde un IPC alto significa una alta inflación y, por ende, una disminución del poder adquisitivo real que tienen las personas. Esta variable podría empujar a retirar el dinero de las cuentas corrientes, las cuales se desvalorizan al no reajustarse mediante este indicador, obligando a las personas a buscar nuevos instrumentos financieros para tratar de cubrirse frente a una potencial alta inflación.

- IMACEC: El índice IMACEC refleja el nivel de actividad económica del país, por lo cual un IMACEC alto implicaría que el país posee un alto nivel de actividad y, por ende, podría significar un aumento en las cuentas corrientes, sobre todo de empresas financieras y PYMES.

Una vez justificada la inclusión de estas variables en el modelamiento posterior que se realiza, es necesario evitar una alta correlación entre estas variables, ya que reducen el

desempeño de los modelos. Dicho lo anterior, se procede a eliminar variables que puedan ser redundantes en el análisis pues alguna otra variable pueda capturar el mismo efecto. Para realizar esto, en primera instancia se realiza una matriz de correlación de variables para observar que tan grande o no es este nivel de correlación y si es significativo. Los resultados se muestran a continuación:

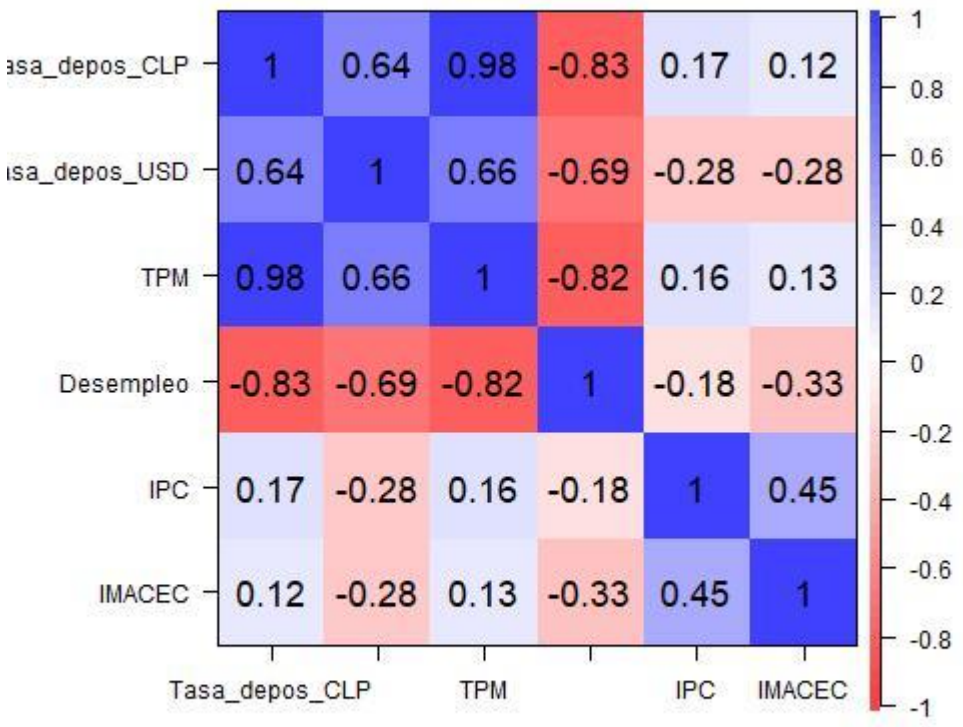

Ilustración 8. Matriz de correlación de las variables exógenas

Del gráfico de correlación anterior se puede desprender lo siguiente:

- Las tasas de depósitos (tanto en CLP como en USD, pero esta última en menor medida) están fuertemente correlacionadas con la serie de la TPM y la correlación es positiva. Esta dependencia era esperable, pues la TPM es la tasa que regula a las demás tasas del mercado, entonces un aumento en la TPM implicará un aumento en las demás tasas y, específicamente, implicará un aumento en las tasas de los depósitos en CLP (fuertemente) y en USD (en menor medida, pero aun así de forma relevante)

- Respecto a la serie de desempleo, se observa que está fuertemente correlacionada de forma negativa con la TPM (y, por ende, con las tasas de depósitos). Este

resultado podría ser sorpresivo debido a la distinta naturaleza de estas variables. La interpretación anterior nos dice que al aumentar la TPM, empujaría a la baja al desempleo (o viceversa), lo cual parece no tener una relación directa. Una posible respuesta es que exista una tercera variable que no se está incluyendo en el análisis, provocando multicausalidad o correlación espuria.

- Respecto a la correlación entre las demás variables, se observa que estas no están fuertemente correlacionadas, lo cual no afectaría en la inclusión de todas estas variables en los modelos propuestos en las secciones posteriores.

Para estar seguros de qué variables eliminar o no, se emplea la metodología LASSO, la cual servirá como complemento al análisis realizado anteriormente. Utilizando el *software* RStudio y añadiendo todas las variables, sin realizar ninguna eliminación el resultado es el siguiente:

```
                          s0
(Intercept)    1661.569156
Tasa_depos_CLP    .
Tasa_depos_USD -157.226920
TPM             -15.017269
Desempleo       -21.460789
IPC              84.041008
IMACEC            2.403075
```

Ilustración 9. Resultados test LASSO, primera iteración

Del resultado se observa que la variable Tasa_depos_CLP queda con parámetro cero, indicando que es una variable que hay que eliminar del análisis. De acuerdo con el análisis de correlaciones tiene sentido el resultado, pues esta variable está altamente correlacionada con la TPM, sin embargo, la serie de los depósitos es más relevante para este trabajo que la TPM, pues es una variable con información diaria y, además, a priori, está más estrechamente relacionada con el producto de cuentas corrientes, por lo cual se decide eliminar la TPM del análisis. Dicho esto, se ejecuta nuevamente el método LASSO, resultando lo siguiente:

```
                          s0
(Intercept)    1591.952286
Tasa_depos_CLP    .
Tasa_depos_USD -159.595742
Desempleo       -15.681519
IPC              78.273087
IMACEC            2.587278
```

Ilustración 10.Resultados test LASSO, segunda iteración

Nuevamente asigna el valor cero a la variable Tasa_depos_CLP, esto debido a que la tasa de los depósitos en CLP está correlacionada con la tasa en USD, sin embargo, en los análisis posteriores no se utilizarán ambas variables simultáneamente, si no que se utilizarán de forma separada dependiendo del tipo de moneda que posea el segmento (es decir, para los segmentos en CLP se utilizará la tasa de depósitos en pesos y para los segmentos en USD se utilizará la tasa

de depósitos en USD). Luego de esta observación, se elimina del análisis (pero no del set de variables) la variable Tasa_depos_USD, obteniéndose el siguiente resultado:

	s0
(Intercept)	798.71986
Tasa_depos_CLP	-25.05134
Desempleo	48.07442
IPC	203.96802
IMACEC	10.00253

Ilustración 11.Resultados test LASSO, tercera iteración y final

Del resultado anterior se observa que la metodología LASSO no obliga a eliminar más variables, por lo cual se decide mantener la variable de desempleo, eliminando del modelamiento solo la serie de la TPM. Estas variables serán utilizadas en secciones posteriores de este trabajo.

3.3. Resultados Segmento 1

Como ya se mencionó en 2.1.7. Segmentación, el segmento 1 corresponde a la agrupación de todos los clientes dentro de la cuenta IFRS 2100109 (Cuentas corrientes de otras personas jurídicas), que tienen sus saldos en moneda CLP, que sus cuentas no son remuneradas y pertenecientes a la clasificación interna de Instituciones no financieras y PYME.

De aquí en adelante, todo el análisis y modelamiento se realiza sobre la data de modelamiento del segmento 1, de forma de no influir sobre la data de validación cuando se realice el análisis de validación correspondiente.

En la sección 3.1. Exploración preliminar se puede observar que el comportamiento de la serie posee una clara tendencia hacia el crecimiento, lo cual implica que la serie no es estacionaria. La condición de estacionariedad es necesaria para la utilización de los modelos ARIMA y sus derivados, por lo cual se evaluará de forma estadística esta condición. Utilizando los test ADF[3] y KPSS[4] se evalúa si la serie es estacionaria o hay existencia de una raíz unitaria. Los resultados son los siguientes:

[3] Función ur.df, librería urca en Rstudio
[4] Función ur.kpss, librería urca en Rstudio

```
###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression none


Call:
lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)

Residuals:
     Min       1Q   Median       3Q      Max
-159.268  -15.460    0.918   15.486  170.063

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
z.lag.1     0.0005071  0.0009188   0.552  0.58120
z.diff.lag -0.1133093  0.0367233  -3.085  0.00211 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32.48 on 733 degrees of freedom
Multiple R-squared:  0.01306,   Adjusted R-squared:  0.01037
F-statistic: 4.851 on 2 and 733 DF,  p-value: 0.00807


Value of test-statistic is: 0.5519

Critical values for test statistics:
      1pct  5pct 10pct
tau1 -2.58 -1.95 -1.62
```

Ilustración 12. Test ADF Segmento 1, serie original

```
######################
# KPSS Unit Root Test #
######################

Test is of type: mu with 6 lags.

Value of test-statistic is: 9.9635

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.347 0.463  0.574 0.739
```

Ilustración 13. Test KPSS Segmento 1, serie original

Los test anteriores entregan una gran cantidad de información respecto a residuos, valor de los coeficientes y distintas métricas de error. Para evaluar la estacionariedad de la serie, el valor de los estadísticos (test-statistic en las imágenes anteriores) y sus valores críticos son los relevantes para este objetivo. En este caso, el estadístico del test ADF es 0.5519 y el estadístico del test KPSS es 9.9635. Es necesario mencionar que la comparación de los estadísticos ya mencionados versus los valores críticos es distinta en ambos tipos de test, debido al signo contrario que estos poseen. En el caso de los valores críticos del test KPSS, estos aumentan a medida que el nivel de significancia disminuye, y el signo de estos es positivo. Por el contrario, en el test ADF los valores críticos son negativos y a medida que el nivel de significancia disminuye el

valor crítico se hace más negativo. Dicho lo anterior, la forma de evaluar el test en contraria en ambos casos. Así para el caso del test KPSS, el estadístico (9.9635) es mayor que los valores críticos, para todos los niveles de significancia, por lo cual se rechaza la hipótesis nula, en favor de la alternativa la cual indica que hay existencia de una raíz unitaria. Así mismo, para el caso del test ADF, el estadístico (0.5519) es mayor que los valores críticos (los cuales son negativos), para todos los niveles de significancia, por lo cual no es posible rechazar la hipótesis nula, concluyendo de igual manera la existencia de una raíz unitaria.

Como se mencionó anteriormente, que la serie no sea estacionaria significa un problema para la aplicación del modelo ARIMA, por lo cual es necesario modificar la serie original para que cumpla con la condición de estacionariedad. La forma de realizar esta modificación es diferenciando la serie, es decir, creando una nueva serie la cual es la resta del saldo en el periodo *t* menos el saldo en el periodo *t-1.* Esta nueva serie tiene la siguiente forma:

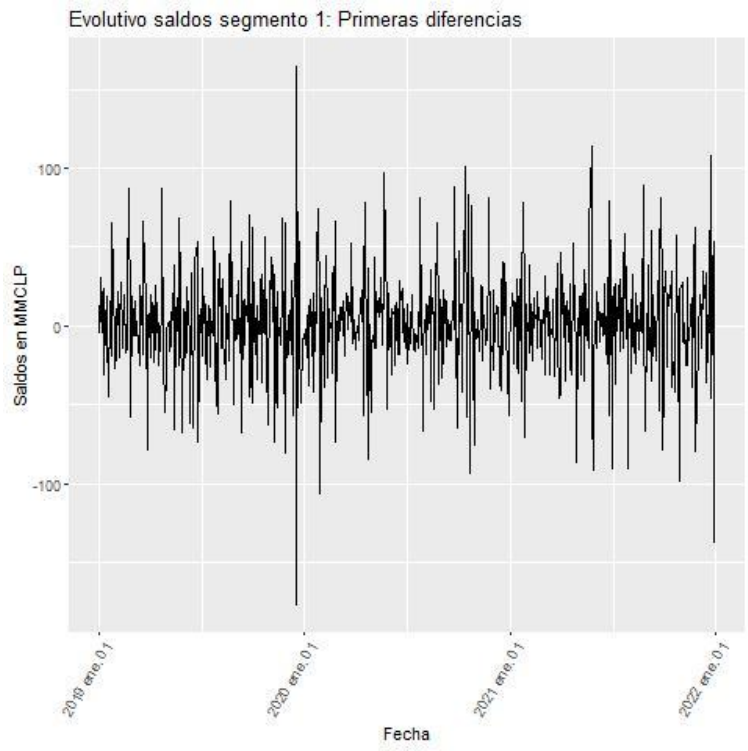

Ilustración 14. Evolutivo del segmento 1, en sus primeras diferencias

Del gráfico anterior se observa inmediatamente una diferencia en el comportamiento, mientras la serie original posee un marcado crecimiento a medida que se avanza en el tiempo, la serie de sus diferencias está centrada en cero, con distintas amplitudes que implican mayores o menores diferencias de saldos entre un día y otro, pero siempre variando respecto al eje. Dicho lo anterior, esta nueva serie pareciera cumplir con las condiciones de la estacionariedad.

Para confirmar el comportamiento analizado gráficamente, se vuelven a realizar los test ADF y KPSS, pero ahora sobre la serie de diferencias. Los resultados son los siguientes:

```
###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression none


Call:
lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)

Residuals:
     Min       1Q   Median       3Q      Max
-153.657  -15.218    1.231   15.886  171.502

Coefficients:
           Estimate Std. Error t value Pr(>|t|)
z.lag.1    -1.19436    0.05533 -21.587   <2e-16 ***
z.diff.lag  0.07378    0.03729   1.978   0.0483 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32.41 on 732 degrees of freedom
Multiple R-squared:  0.5588,    Adjusted R-squared:  0.5576
F-statistic: 463.6 on 2 and 732 DF,  p-value: < 2.2e-16


Value of test-statistic is: -21.587

Critical values for test statistics:
      1pct  5pct 10pct
tau1 -2.58 -1.95 -1.62
```

Ilustración 15.Test ADF Segmento 1, serie diferenciada

```
######################
# KPSS Unit Root Test #
######################

Test is of type: mu with 6 lags.

Value of test-statistic is: 0.0093

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.347 0.463  0.574 0.739
```

Ilustración 16.Test KPSS Segmento 1, serie diferenciada

Como se intuía del gráfico de la serie de las diferencias, ahora ambos test cambian opuestamente sus estadísticos. En el caso del test ADF, su valor es de -21.587, menor que todos los valores críticos (o, dicho de otra forma, es más negativo que los valores críticos) por lo cual se rechaza la hipótesis nula, en favor de la alternativa la cual indica que la serie es estacionaria. Respecto al test KPSS, el valor del estadístico es de 0.0093, menor que los valores críticos para todos los niveles de significancia, lo cual implica que no es posible rechazar la hipótesis nula y de este modo la serie es estacionaria o posee tendencia estacionaria. Complementando ambos test

se puede concluir que la serie de diferencias de la serie original de saldos del segmento 1 es estacionaria, lo cual da el paso para aplicar los modelos ARIMA al cumplir la condición necesaria de estacionariedad.

3.3.1. ARIMA sin estacionalidad

El siguiente paso es modelar mediante la serie ARIMA y sus parámetros clásicos (parámetros *p*, *d* y *q* no estacionales). Del análisis anterior, podemos concluir rápidamente que el parámetro *d*, asociado a la parte integrada del ARIMA es igual a 1, ya que esta componente corresponde al número de diferencias no estacionarias para alcanzar la estacionariedad. Dicho lo anterior, ahora resta conocer los parámetros *p* y *q*, asociados a las componentes de media móvil y autorregresiva del modelo. Para lograr este objetivo se utilizan las funciones de autocorrelación y autocorrelación parcial, resultando los siguientes gráficos:

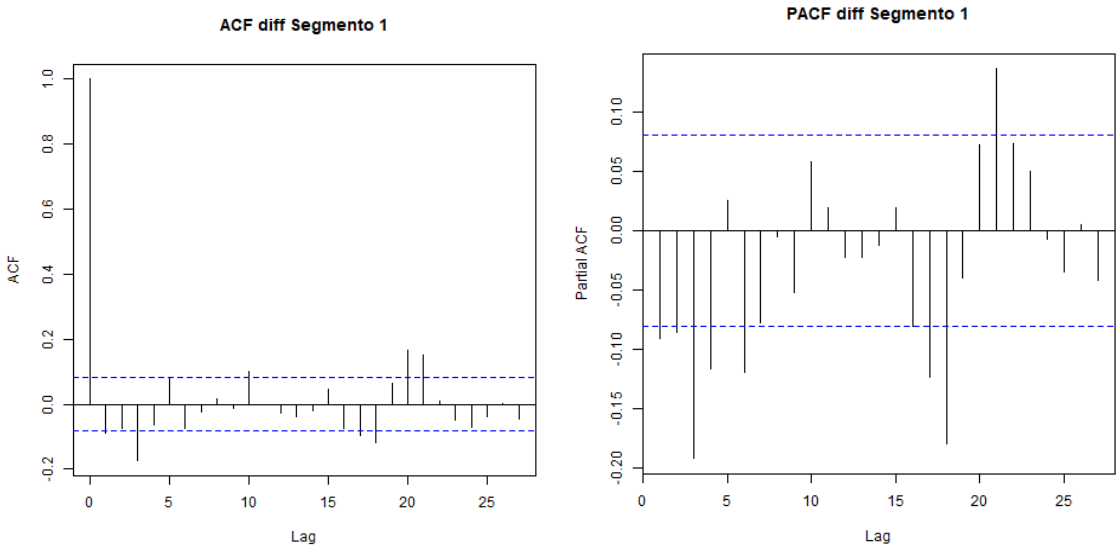

Ilustración 17. Gráficos ACF y PACF Segmento 1, primeras diferencias

De los gráficos anteriores se pueden comentar las siguientes observaciones:

- Respecto al gráfico de ACF, existen *lags* significativos en el *lag* 1 (al borde de la significancia), *lag* 3 y *lag* 5 y sus múltiplos.
- Respecto al gráfico de PACF, hasta el *lag* 6 son todos significativos a excepción del *lag* 5. Además, la significancia reaparece en *lags* más tardíos, superiores al *lag* 15.

De acuerdo con las significancias identificadas, se evalúan las posibles combinaciones para AR(1), AR(2), AR(3), AR(4) y AR(6) respecto a su componente autorregresiva, y MA(1), MA(3) y MA(5) respecto a su componente de media móvil. Así, los parámetros ARIMA a evaluar son los siguientes:

- ARIMA(1,1,1)
- ARIMA(1,1,3)
- ARIMA(1,1,5)
- ARIMA(2,1,1)
- ARIMA(2,1,3)
- ARIMA(2,1,5)
- ARIMA(3,1,1)
- ARIMA(3,1,3)
- ARIMA(3,1,5)
- ARIMA(4,1,1)
- ARIMA(4,1,3)
- ARIMA(4,1,5)
- ARIMA(6,1,1)
- ARIMA(6,1,3)
- ARIMA(6,1,5)

A continuación, se muestra una tabla en donde se muestran los criterios AIC,BIC y Error Cuadrático Medio para cada uno de los parámetros ARIMA indicados anteriormente.

ARIMA	AIC	BIC	RMSE
(1,1,1)	5.740	5.753	31,52
(2,1,1)	5.738	5.755	31,40
(3,1,1)	5.716	5.738	30,76
(4,1,1)	5.708	5.735	30,50
(6,1,1)	5.702	5.737	30,23
(1,1,3)	5.704	5.725	30,39
(2,1,3)	5.707	5.733	30,40
(3,1,3)	5.705	5.735	30,32
(4,1,3)	5.706	5.741	30,32
(6,1,3)	5.701	5.745	30,09
(1,1,5)	5.696	5.727	30,10
(2,1,5)	5.698	5.733	30,09
(3,1,5)	5.697	5.736	30,01
(4,1,5)	5.699	5.742	30,00

(6,1,5)	5.700	5.753	29,92

Tabla 4. Criterios AIC, BIC y RMSE. Segmento 1

De la tabla anterior se observa que no hay un ARIMA que sea el mínimo en los 3 criterios evaluados. Así, los posibles candidatos son:

- Menor AIC: ARIMA(1,1,5)
- Menor BIC: ARIMA(1,1,3)
- Menor RMSE: ARIMA(6,1,5)

Para seleccionar los parámetros ARIMA, el criterio BIC es el criterio principal que se considera, debido a que tiene la ventaja de penalizar más fuertemente sobre la adición de parámetros en comparación al AIC. Bajo esta lógica, el ARIMA a considerar sería el ARIMA(1,1,3), sin embargo, el evaluar conjuntamente tanto AIC y BIC, el mejor ARIMA pareciese ser el ARIMA(1,1,5), debido a que se gana más en AIC que lo se que pierde en BIC. El criterio de RMSE se considera como un tercer criterio, debido a que siempre disminuye al incorporar más parámetros al modelo, lo cual hace disminuir el error de ajuste, pero aumenta el sobreajuste, así que se considera un criterio meramente informativo. Así, dicho lo anterior, el ARIMA seleccionado para este segmento es el ARIMA(1,1,5).

A modo comparativo y de incluir una segunda opinión dentro de la selección de parámetros, existe una función[5] en RStudio que realiza la selección de parámetros de forma automática, solo considerando los menores valores de los criterios AIC y BIC, es decir, sin preocuparse de la validez estadística desarrollada en este trabajo. Dicho lo anterior, la verificación de esta metodología es simplemente para considerarla como un cable a tierra, sin perjuicio de que estas puedan diferir.

Ejecutando el comando indicado anteriormente, el resultado es el siguiente:

```
Series: segmento_ts_model[[1]]
ARIMA(1,1,4) with drift

Coefficients:
          ar1     ma1      ma2      ma3      ma4   drift
      -0.5470  0.3852  -0.2305  -0.3364  -0.2413  0.9213
s.e.   0.2031  0.2006   0.0548   0.0561   0.0579  0.4665

sigma^2 estimated as 915.5:  log likelihood=-2841.31
AIC=5696.61   AICc=5696.81   BIC=5727.26

Training set error measures:
                    ME     RMSE      MAE          MPE     MAPE      MASE          ACF1
Training set 0.0228467 30.07696 21.62908 -0.05222967 1.803349 0.9647366 -0.003042654
```

[5] Función auto.arima, del paquete forecast

Ilustración 18. Resultados auto.arima, Segmento 1

Se observa que los parámetros son muy similares (ARIMA(1,1,5) propuesto versus ARIMA(1,1,4) entregado por el *software*) a los escogidos mediante la metodología propuesta en este trabajo, los cuales fueron avalados por todo el desarrollo estadístico presentado. Adicionalmente, los criterios AIC, BIC y RMSE son prácticamente idénticos, debido a la naturaleza de elección del programa utilizado.

Un paso final para evaluar la validez y consistencia del modelo es realizar un análisis de los residuos. El objetivo, es comprobar que los residuos se comporten como ruido blanco, es decir que la serie de residuos sea independiente e idénticamente distribuida, con media cero e igual varianza. Más específicamente, se evaluará si los residuos están correlacionados o no, ya que de existir autocorrelación implicaría que existe información en los residuos que debiese estar capturada por el modelo. Para comprobar esto, se evalúa tanto la serie de los residuos, su distribución y la función de autocorrelación ACF:

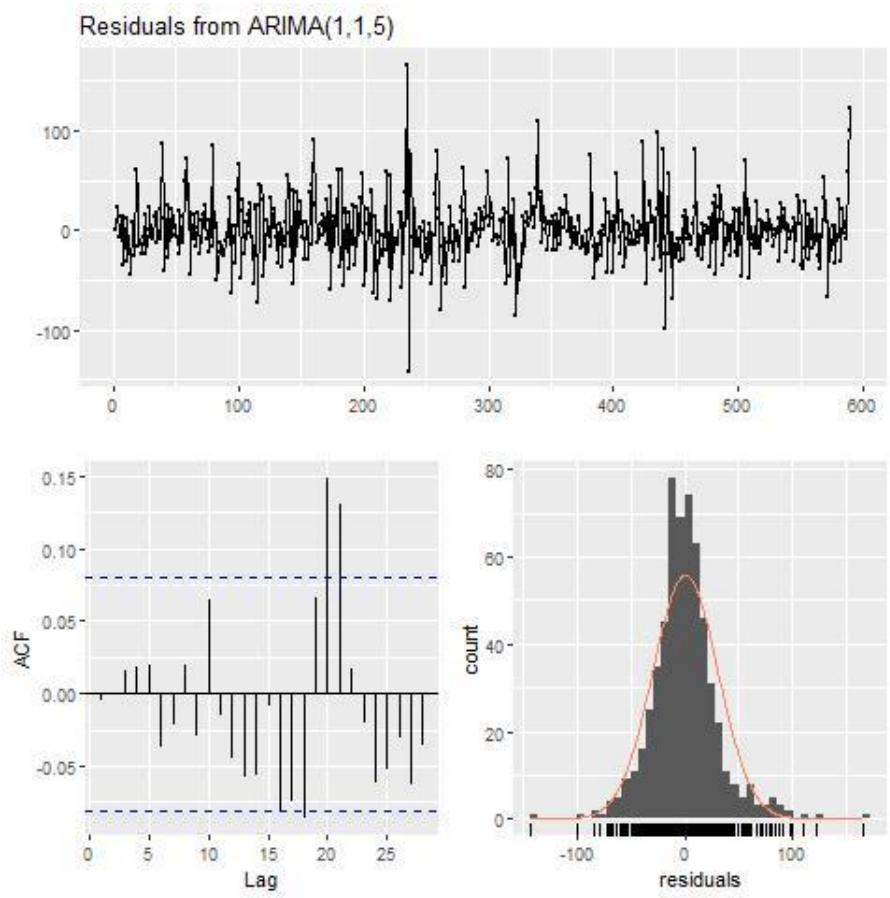

Ilustración 19. Análisis de residuos. Segmento 1

Del gráfico anterior se puede observar lo siguiente:

- Los residuos están centrados en cero, con una varianza relativamente constante a lo largo del tiempo, excepto en fechas muy puntuales.
- Los residuos siguen una distribución muy similar a la distribución normal.
- Del gráfico ACF se observa que existen algunos *lags* que son significativos, en torno al *lag* 20

Dicho todo lo anterior, el comportamiento de los residuos tiene características muy similares a las de ruido blanco, a excepción de la existencia de un par de *lags* significativos, lo cual puede indicar que exista información relevante que no se está utilizando para incluir en el modelo. Esta información, dado el número del *lag* que es significativo, puede estar capturado por alguna variable estacional (los meses poseen 21 días hábiles) que no está siendo incorporada en el modelo ARIMA. Esto se evaluará más adelante en este trabajo.

Un último análisis sobre los residuos corresponde al test Ljung-Box, el cual evalúa que las autocorrelaciones de una serie de tiempo son distintas de cero. Así la hipótesis nula indica que los datos se distribuyen de forma independiente, mientras que la hipótesis alternativa indica que los datos no se distribuyen de forma independiente. Los resultados son los siguientes:

```
        Ljung-Box test

data:  Residuals from ARIMA(1,1,5)
Q* = 4.9992, df = 4, p-value = 0.2874

Model df: 6.   Total lags used: 10
```

Ilustración 20. Test Ljung-Box. Segmento 1

Se observa que el p-valor es de 0.2874, un número mayor que el nivel de significancia del 0.05, lo cual indica que la hipótesis nula no puede ser rechazada, implicando que los residuos se distribuyen independientemente, y por ende son ruido blanco.

3.3.2. ARIMA estacional o SARIMA

El siguiente paso es evaluar la inclusión de variables estacionales al modelo ARIMA. La intuición detrás de este análisis es que puede existir una estacionalidad en las series de saldos debido a la naturaleza del producto bancario de cuentas corrientes, debido a que existen fechas definidas dentro del mes en las cuales se realizan depósitos a estas cuentas por el pago de sueldos, así como también pueden existir fechas definidas de salida de flujos, como lo son las fechas de pagos de cuotas de algún crédito o tarjetas y líneas de créditos.

La forma de incluir esta variable estacional es la siguiente: al modelo ARIMA escogido en la sección anterior se le añaden como variable externa distintas variables dummies, asociadas a días financieros claves dentro del mes, los cuales son ampliamente utilizados en el sector financiero

como fechas de abono o de pago de las cuentas corrientes. La forma de estas variables dummies es la siguiente:

$$\delta_t = \begin{cases} 1 \; si \; el \; día \; calendario \; es \; t, t \in [1; 31] \\ 0 \; si \; no \end{cases}$$

Las variables dummies escogidas para evaluar su impacto son: δ_1, δ_5, δ_{15}, δ_{25}, δ_{30} y δ_{31}. La razón de considerar estos días calendario es la siguiente: en general, las fechas de pago de deudas (cuotas de crédito de consumo y/o hipotecario o cuotas de la tarjeta de crédito) son los días 1 y 5 del mes. Por otro lado, generalmente las fechas de pago de sueldo es a fin de mes (días 30 o 31 dependiendo del mes) y poco menos común pero también relevante evaluar es el pago de avances de sueldo (días 15 del mes). Adicionalmente, se considera el día 25 del mes debido a que es el día en el cual se le pagan los sueldos a los trabajadores de la institución bancaria en la cual se realiza este trabajo y en su mayoría se realiza en cuentas corrientes del mismo banco, por lo cual toma relevancia este día en particular.

Con las variables dummies ya definidas, se procede a evaluar el impacto y la significancia de la inclusión de estas variables en el modelo ARIMA previamente determinado. Así, el resultado es el siguiente:

```
Call:
arima(x = segmento_ts_model[[1]], order = c(1, 1, 5), xreg = base_model[, c(8,
    12, 22, 32, 37, 38)], method = "ML")

Coefficients:
          ar1     ma1      ma2      ma3      ma4     ma5  Dummy_dia_1  Dummy_dia_5  Dummy_dia_15
      -0.4197  0.2530  -0.1936  -0.2649  -0.1978  0.0266       2.2099       7.9368        3.7852
s.e.   0.2724  0.2724   0.0591   0.0579   0.0716  0.0556       6.1703       4.8888        4.9992
      Dummy_dia_25  Dummy_dia_30  Dummy_dia_31
           25.1706       -8.6986        3.1816
s.e.        5.2378        5.6204        8.8368

sigma^2 estimated as 867:  log likelihood = -2828.23,  aic = 5682.45

Training set error measures:
                    ME    RMSE      MAE       MPE     MAPE      MASE          ACF1
Training set 2.170634 29.4191 21.36823 0.1291616 1.776364 0.9531018 -0.006074144
```

```
z test of coefficients:

              Estimate Std. Error z value  Pr(>|z|)
ar1          -0.419697   0.272397 -1.5408  0.123376
ma1           0.252956   0.272431  0.9285  0.353141
ma2          -0.193579   0.059074 -3.2769  0.001050 **
ma3          -0.264903   0.057862 -4.5782 4.690e-06 ***
ma4          -0.197822   0.071626 -2.7619  0.005747 **
ma5           0.026601   0.055559  0.4788  0.632086
Dummy_dia_1   2.209909   6.170274  0.3582  0.720228
Dummy_dia_5   7.936808   4.888829  1.6235  0.104492
Dummy_dia_15  3.785231   4.999193  0.7572  0.448949
Dummy_dia_25 25.170645   5.237765  4.8056 1.543e-06 ***
Dummy_dia_30 -8.698554   5.620356 -1.5477  0.121698
Dummy_dia_31  3.181569   8.836816  0.3600  0.718820
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ilustración 21. Estadísticos y p-valor variables dummy estacionales. Segmento 1

De los resultados anteriores, se observa que solo la dummy asociada al día 25 es significativa al 99,9%, y su coeficiente es positivo, lo cual indica que en los días 25 del mes, se explica un incremento de 25.000 MMCLP en los saldos de este segmento, solo por el día calendario.

3.3.3. ARIMA con variables exógenas o ARIMAX

El paso siguiente es evaluar la inclusión de variables exógenas al modelo ARIMA escogido anteriormente. Como se discutió en la sección 3.2. Selección de variables, las variables externas que se evaluarán, tanto en significancia estadística como en el valor de sus coeficientes son:

- Tasa de depósitos a plazo de 30 días (ya sea en CLP o USD dependiendo del tipo de moneda del segmento que se esté modelando)
- Desempleo
- IPC
- IMACEC

De la misma forma que los análisis anteriores, se incluye este conjunto de nuevas variables sobre el modelo ARIMA escogido previamente, de tal forma de analizar una potencial mejora en el modelamiento. Cabe mencionar que las variables exógenas incluidas en el modelo están diferenciadas, de tal forma de capturar la variación diaria de estas. Los resultados son los siguientes:

```
Call:
arima(x = diff(segmento_ts_model[[1]]), order = c(1, 1, 5), xreg = cbind(depos_CLP,
    desempleo, IPC, IMACEC), method = "ML")

Coefficients:
         ar1      ma1      ma2      ma3     ma4     ma5  depos_CLP  desempleo      IPC   IMACEC
      -0.565  -0.5915  -0.6378  -0.1023  0.0877  0.2438     3.6355    -8.1096  10.3600  -1.3872
s.e.   0.190   0.1874   0.2185   0.0565  0.0452  0.0545    14.2883     8.8558  13.7462   1.2204

sigma^2 estimated as 904.2:  log likelihood = -2839.97,  aic = 5701.93

Training set error measures:
                    ME     RMSE      MAE      MPE     MAPE      MASE         ACF1
Training set 0.6635347 30.04444 21.52892 82.73116 153.9348 0.6254522 -0.005996433
```

```
z test of coefficients:

            Estimate Std. Error z value  Pr(>|z|)
ar1       -0.564966   0.189956 -2.9742  0.002938 **
ma1       -0.591471   0.187435 -3.1556  0.001602 **
ma2       -0.637765   0.218484 -2.9190  0.003511 **
ma3       -0.102273   0.056528 -1.8092  0.070414 .
ma4        0.087694   0.045185  1.9408  0.052285 .
ma5        0.243829   0.054546  4.4702 7.815e-06 ***
depos_CLP  3.635501  14.288320  0.2544  0.799157
desempleo -8.109582   8.855785 -0.9157  0.359804
IPC       10.359986  13.746213  0.7537  0.451053
IMACEC    -1.387189   1.220411 -1.1367  0.255681
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ilustración 22.Estadísticos y p-valor variables exógenas. Segmento 1

De los resultados anteriores se observa que ninguna variable exógena es significativa. Por lo cual no hay suficiente evidencia como para incorporarlas al modelo.

3.3.4. Proyección de los modelos

Luego de todo el desarrollo estadístico anterior, que tuvo como objetivo seleccionar los mejores modelos en base a criterios de pérdida de información, el siguiente paso medir la capacidad predictiva de estos modelos. Esta sección es la más relevante dentro del análisis, pues ataca directamente al objetivo general de este trabajo.

La metodología que se seguirá en esta sección es la siguiente: considerando los modelos escogidos anteriormente, se realizará una proyección de los saldos, abarcando 147 registros hacia adelante. La razón de esto es que la data de validación (que contiene el 20% final de la data completa) posee esa cantidad de registros. Una vez realizada la proyección para cada modelo anterior, los resultados que se obtengan de esto serán comparados contra la data de validación, así utilizando el criterio de RMSE, se podrá determinar cual modelo predice de mejor forma el comportamiento de los saldos en cuenta corriente hacia el futuro.

En primera instancia, se muestran los gráficos proyectados utilizando la función *forecast:*

- ARIMA(1,1,5) sin estacionalidad ni variables exógenas (con drift).

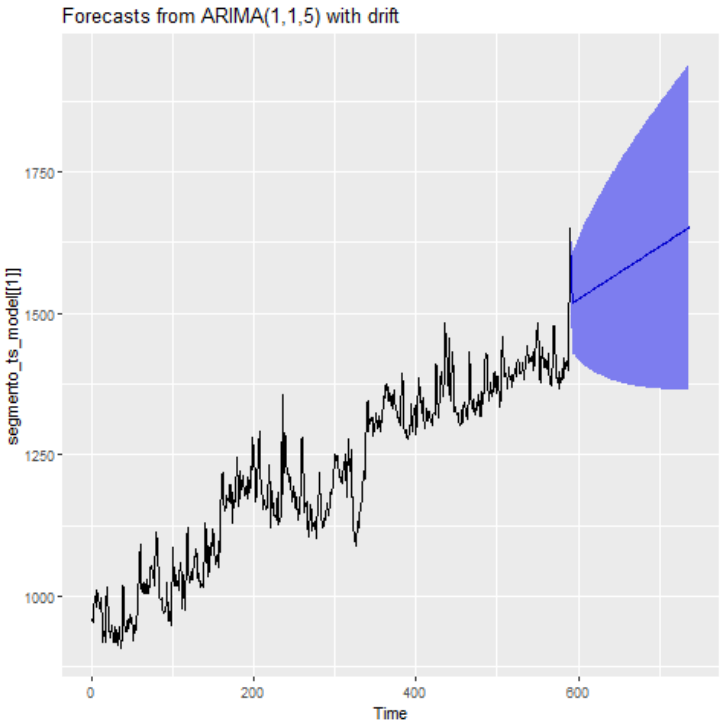

Ilustración 23. Proyección ARIMA no estacional. Segmento 1

- ARIMA(1,1,5) con δ_{25} y sin variable exógena (con drift).

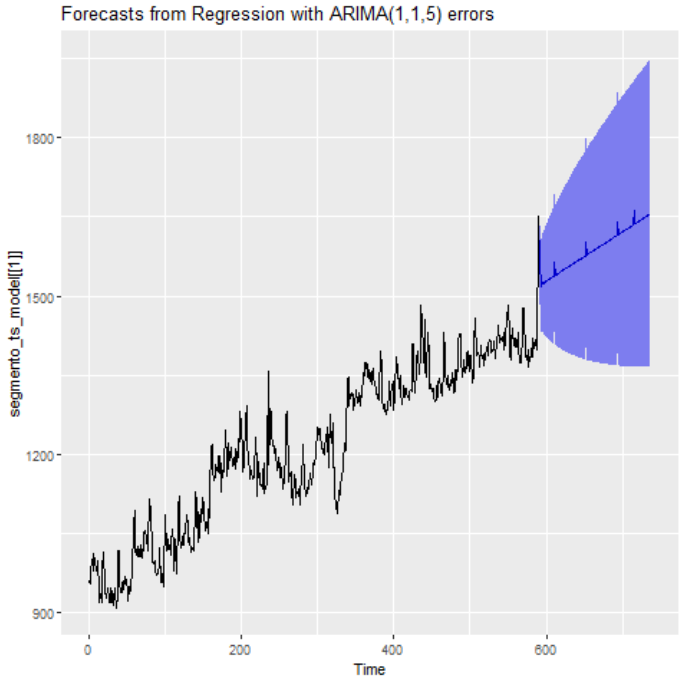

Ilustración 24. Proyección ARIMA con variable estacional. Segmento 1

Luego de esta observación gráfica de las proyecciones, se procede a comparar estos valores predichos versus la data de validación. Los resultados se muestran a continuación:

ARIMA	RMSE
(1,1,5)	50,85
(1,1,5) con δ_{25}	50,47

Tabla 5. Resultados RSME sobre data de validación

3.4. Resultados Segmento 2

Como ya se mencionó anteriormente, el segmento 2 corresponde a la agrupación de todos los clientes dentro de la cuenta IFRS 2100109 (Cuentas corrientes de otras personas jurídicas), que tienen sus saldos en moneda CLP, que sus cuentas son remuneradas y pertenecientes a la clasificación interna de Instituciones no financieras y PYME.

De aquí en adelante, todo el análisis y modelamiento se realiza sobre la data de modelamiento del segmento 2.

En la sección 3.1. Exploración preliminar se puede observar que el comportamiento de la serie posee una clara tendencia hacia el crecimiento, pero con periodos con caídas abruptas. Independiente de estas caídas específicas en los saldos, la serie de todas formas es potencialmente no estacionaria. De forma análoga a la realizado en la sección 3.3. Resultados

Segmento 1, se utilizan los test KPSS y DFA para evaluar estadísticamente la presencia o no de estacionariedad. Los resultados son los siguientes:

```
###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression none


Call:
lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)

Residuals:
     Min       1Q   Median       3Q      Max
-241.807  -10.287    2.018   13.778  242.833

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
z.lag.1    -0.001938   0.002570  -0.754    0.451
z.diff.lag -0.208412   0.036128  -5.769 1.18e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31.24 on 733 degrees of freedom
Multiple R-squared:  0.04459,   Adjusted R-squared:  0.04199
F-statistic: 17.11 on 2 and 733 DF,  p-value: 5.483e-08


Value of test-statistic is: -0.7539

Critical values for test statistics:
      1pct  5pct 10pct
tau1 -2.58 -1.95 -1.62
```

Ilustración 25. Test ADF Segmento 2, serie original

```
######################
# KPSS Unit Root Test #
######################

Test is of type: mu with 6 lags.

Value of test-statistic is: 4.6832

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.347 0.463  0.574 0.739
```

Ilustración 26. Test KPSS Segmento 2, serie original

Para este segmento, el estadístico del test ADF es -0.7539 y el estadístico del test KPSS es 4.6832. Comparándolos contra los valores críticos se tiene que el estadístico del test ADF es mayor al de los valores críticos, por lo cual no es posible rechazar la hipótesis nula. Respecto al

estadístico del test KPSS, este es mayor que los valores críticos, por ende, se rechaza la hipótesis nula en favor de la hipótesis alternativa. Así, se prueba que la serie no es estacionaria pues se comprueba la existencia de una raíz unitaria.

Dado el resultado anterior, se hace necesario diferenciar la serie de saldos de este segmento para comprobar si de esta forma la serie se vuelve estacionaria. La nueva serie de saldos del segmento 2 diferenciada posee la siguiente forma:

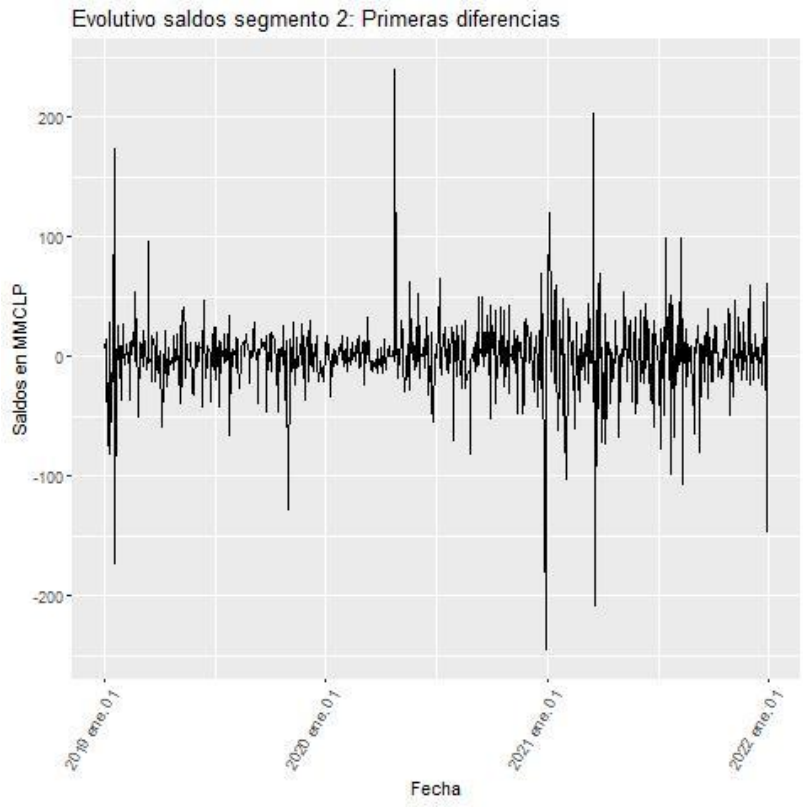

Ilustración 27. Evolutivo del segmento 2 en sus primeras diferencias

Del gráfico anterior, se observa que la serie de las diferencias está centrada en cero, con algunas variaciones más amplias que la del resto de la serie, pero sin influir en el comportamiento general. Dicho lo anterior, la serie de diferencias pareciera cumplir con las condiciones de la estacionariedad.

Para confirmar el comportamiento analizado gráficamente, se vuelven a realizar los test ADF y KPSS, pero ahora sobre la serie de diferencias. Los resultados son los siguientes:

```
###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression none


Call:
lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)

Residuals:
     Min       1Q   Median       3Q      Max
-243.201  -11.101    1.281   13.090  242.353

Coefficients:
           Estimate Std. Error t value Pr(>|t|)
z.lag.1    -1.18328    0.05805 -20.384   <2e-16 ***
z.diff.lag -0.02167    0.03752  -0.578    0.564
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31.27 on 732 degrees of freedom
Multiple R-squared:  0.605,     Adjusted R-squared:  0.6039
F-statistic: 560.6 on 2 and 732 DF,  p-value: < 2.2e-16


Value of test-statistic is: -20.3837

Critical values for test statistics:
      1pct  5pct 10pct
tau1 -2.58 -1.95 -1.62
```

Ilustración 28.Test ADF Segmento 2, serie diferenciada

```
#######################
# KPSS Unit Root Test #
#######################

Test is of type: mu with 6 lags.

Value of test-statistic is: 0.0457

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.347 0.463  0.574 0.739
```

Ilustración 29.Test KPSS Segmento 2, serie diferenciada

Con la diferenciación de la serie, se genera una modificación importante en los estadísticos KPSS y ADF. En el caso del test ADF, su valor es de -20.3837, menor que todos los valores críticos (o, dicho de otra forma, es más negativo que los valores críticos) por lo cual se rechaza la hipótesis nula, en favor de la alternativa la cual indica que la serie es estacionaria. Respecto al test KPSS, el valor del estadístico es de 0.0457, menor que los valores críticos para todos los niveles de significancia, lo cual implica que no es posible rechazar la hipótesis nula y de este modo la serie

es estacionaria o posee tendencia estacionaria. Complementando ambos test se puede concluir que la serie de diferencias de la serie original de saldos del segmento 2 es estacionaria, lo cual da el paso para aplicar los modelos ARIMA al cumplir la condición necesaria de estacionariedad.

3.4.1. ARIMA sin estacionalidad

De forma análoga a lo realizado en la sección 3.3.1. ARIMA sin estacionalidad, el siguiente paso es encontrar los parámetros *p* y *q*, pues gracias a la diferenciación el parámetro *d* es igual a 1. Para lograr lo anterior se utilizan las funciones ACF y PACF, las cuales se muestran a continuación:

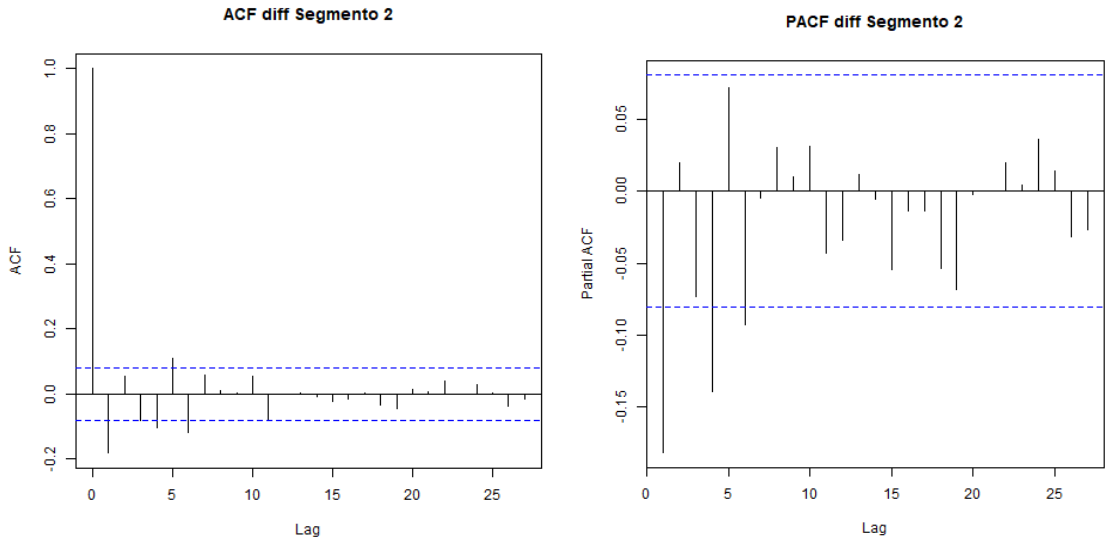

Ilustración 30. Gráficos ACF y PACF Segmento 2, primeras diferencias

De los gráficos anteriores se pueden comentar las siguientes observaciones:

- Respecto al gráfico de ACF, existen *lags* significativos en el *lag* 1, *lag* 4, *lag* 5 y *lag* 6.
- Respecto al gráfico de PACF, los únicos *lags* significativos son el *lag* 1, el *lag* 4 y el *lag* 6.

De acuerdo con las significancias identificadas, se evalúan las posibles combinaciones para AR(1), AR(4) y AR(6) respecto a su componente autorregresiva, y MA(1), MA(4), MA(5) y MA(6) respecto a su componente de media móvil. Así, los parámetros ARIMA a evaluar son los siguientes:

- ARIMA(1,1,1)
- ARIMA(1,1,4)

- ARIMA(1,1,5)
- ARIMA(1,1,6)
- ARIMA(4,1,1)
- ARIMA(4,1,4)
- ARIMA(4,1,5)
- ARIMA(4,1,6)
- ARIMA(6,1,1)
- ARIMA(6,1,4)
- ARIMA(6,1,5)
- ARIMA(6,1,6)

A continuación, se muestra una tabla en donde se muestran los criterios AIC,BIC y Error Cuadrático Medio para cada uno de los parámetros ARIMA indicados anteriormente.

ARIMA	AIC	BIC	RMSE
(1,1,1)	5.727	5.740	31,17
(4,1,1)	5.719	5.745	30,77
(6,1,1)	5.714	5.749	30,55
(1,1,4)	5.722	5.749	30,87
(4,1,4)	5.716	5.756	30,55
(6,1,4)	5.720	5.768	30,55
(1,1,5)	5.717	5.748	30,68
(4,1,5)	5.718	5.762	30,55
(6,1,5)	5.720	5.772	30,49
(1,1,6)	5.718	5.753	30,65
(4,1,6)	5.720	5.768	30,55
(6,1,6)	5.722	5.779	30,48

Tabla 6. Criterios AIC, BIC y RMSE. Segmento 2

De la tabla anterior se observa que no hay un ARIMA que sea el mínimo en los 3 criterios evaluados. Así, los posibles candidatos son:

- Menor AIC: ARIMA(6,1,1)
- Menor BIC: ARIMA(1,1,1)
- Menor RMSE: ARIMA(6,1,6)

Considerando al BIC como el indicador principal, el ARIMA a escoger sería el ARIMA(1,1,1), sin embargo, el evaluar conjuntamente tanto AIC y BIC, el mejor ARIMA pareciese ser el

ARIMA(6,1,1), debido a que se gana más en AIC que lo que pierde en BIC. El criterio de RMSE se considera como un tercer criterio, debido a que siempre disminuye al incorporar más parámetros al modelo, lo cual hace disminuir el error de ajuste, pero aumenta el sobreajuste, así que se considera un criterio meramente informativo. Así, dicho lo anterior, el ARIMA seleccionado para este segmento es el ARIMA(6,1,1).

A modo comparativo y de incluir una segunda opinión dentro de la selección de parámetros, se ejecuta la función auto.arima. El resultado es el siguiente:

```
Series: segmento_ts_model[[2]]
ARIMA(4,1,1)

Coefficients:
          ar1      ar2      ar3      ar4     ma1
      -0.6953  -0.0812  -0.0951  -0.1830  0.5234
s.e.   0.1427   0.0552   0.0495   0.0409  0.1419

sigma^2 estimated as 946:  log likelihood=-2851.33
AIC=5714.67   AICc=5714.81   BIC=5740.94

Training set error measures:
                    ME     RMSE      MAE        MPE     MAPE     MASE         ACF1
Training set 0.3407792 30.60108 18.18672 -0.2988553 4.609314 1.007045 0.001608393
```

Ilustración 31. Resultados auto.arima, Segmento 2

Se observa que los parámetros difieren levemente en su componente autorregresiva, sin embargo, era una de las opciones intermedias entre las evaluadas anteriormente, por lo cual la diferencia no es tan relevante. Adicionalmente, los criterios AIC y BIC son levemente inferiores (menor pérdida de información) que los encontrados en la exploración de los distintos parámetros, sin embargo, no afectan a la conclusión del modelo seleccionado.

Por último, se realiza el análisis sobre los residuos, de tal forma de corroborar si cumplen con las características que lo hacen ser ruido blanco. Este análisis se muestra a continuación:

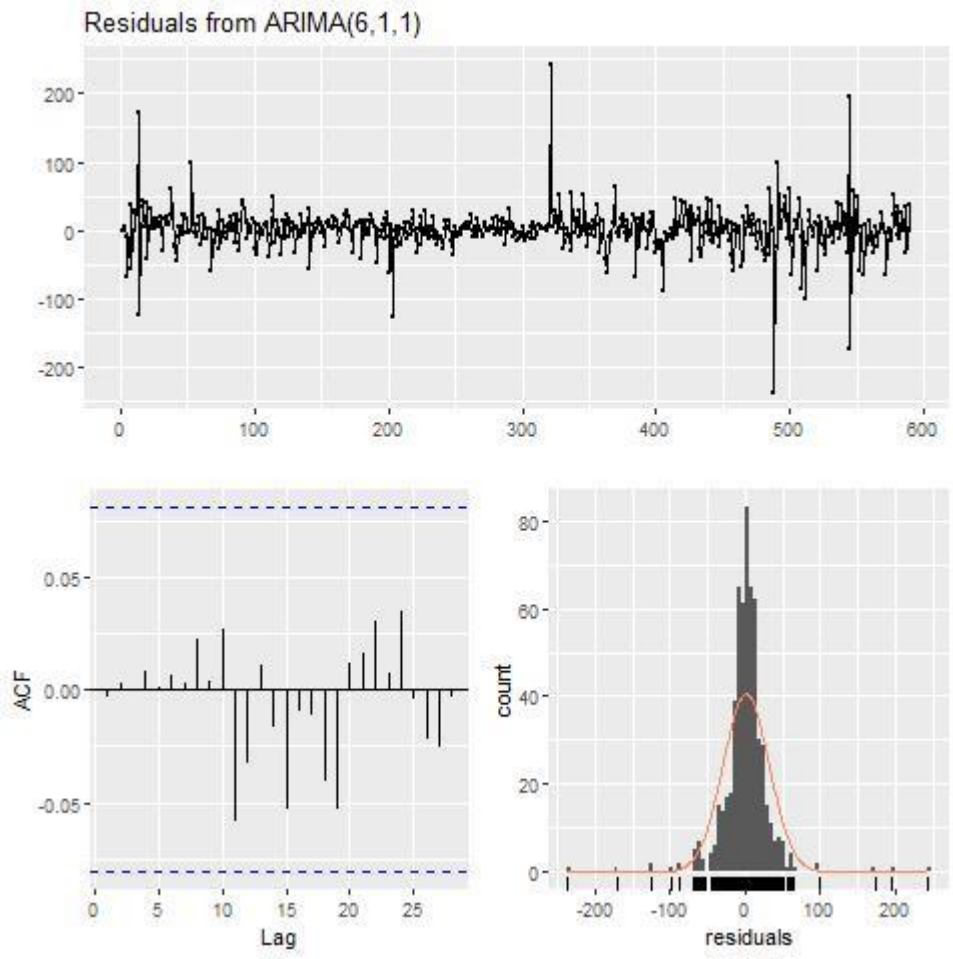

Ilustración 32. Análisis de residuos. Segmento 2

Del gráfico anterior se puede observar lo siguiente:

- Los residuos están centrados en cero, con una varianza relativamente constante a lo largo del tiempo, excepto en un par de fechas muy puntuales.
- Los residuos siguen una distribución muy similar a la distribución normal, donde la gran mayoría de los datos está muy concentrada en cero.
- Del gráfico ACF se observa que no existen *lags* significativos, todos están dentro de la banda de significancia, implicando que no hay pérdida de información en el modelo.

Dicho lo anterior, los residuos parecen cumplir a cabalidad las condiciones para ser considerado ruido blanco. A modo complementario, se realiza el test Ljung-Box, para confirmar lo anterior. Los resultados son los siguientes:

```
        Ljung-Box test

data:  Residuals from ARIMA(6,1,1)
Q* = 0.8133, df = 3, p-value = 0.8463

Model df: 7.   Total lags used: 10
```

Ilustración 33. Test Ljung-Box. Segmento 2

Se observa que el p-valor es de 0.8463, un número bastante mayor que el nivel de significancia del 0.05, lo cual indica que la hipótesis nula no puede ser rechazada, implicando que los residuos se distribuyen independientemente, y por ende son ruido blanco.

3.4.2. ARIMA estacional o SARIMA

De la misma forma que en la sección 3.3.2. ARIMA estacional o SARIMA, se evalúa la significancia de agregar al modelo las variables dummies asociadas a los días calendario previamente definidos. Los resultados son los siguientes:

```
Call:
arima(x = segmento_ts_model[[2]], order = c(6, 1, 1), xreg = base_model[, c(8,
    12, 22, 32, 37, 38)], method = "ML")

Coefficients:
          ar1     ar2      ar3      ar4     ar5      ar6      ma1  Dummy_dia_1  Dummy_dia_5
      -0.1678  0.0135  -0.1010  -0.1389  0.0536  -0.0925  -0.0014       7.1741       3.3843
s.e.   0.3705  0.0772   0.0427   0.0541  0.0672   0.0491   0.3710       6.1235       5.0325
      Dummy_dia_15  Dummy_dia_25  Dummy_dia_30  Dummy_dia_31
            7.2316        3.7260       -1.5829       -2.4019
s.e.        5.1112        5.1888        5.6127        8.5208

sigma^2 estimated as 926.4:  log likelihood = -2847.68,  aic = 5723.37

Training set error measures:
                    ME     RMSE      MAE       MPE     MAPE     MASE          ACF1
Training set 0.3617832 30.41091 18.09474 -0.303148 4.582673 1.001952 -0.0005053379
```

```
z test of coefficients:

              Estimate Std. Error z value Pr(>|z|)
ar1          -0.167827   0.370452 -0.4530  0.65053
ar2           0.013495   0.077199  0.1748  0.86123
ar3          -0.100956   0.042692 -2.3647  0.01804 *
ar4          -0.138860   0.054052 -2.5690  0.01020 *
ar5           0.053617   0.067190  0.7980  0.42488
ar6          -0.092479   0.049135 -1.8821  0.05982 .
ma1          -0.001422   0.371023 -0.0038  0.99694
Dummy_dia_1   7.174126   6.123520  1.1716  0.24137
Dummy_dia_5   3.384289   5.032470  0.6725  0.50127
Dummy_dia_15  7.231617   5.111193  1.4149  0.15711
Dummy_dia_25  3.725997   5.188802  0.7181  0.47271
Dummy_dia_30 -1.582936   5.612717 -0.2820  0.77792
Dummy_dia_31 -2.401949   8.520774 -0.2819  0.77803
```

Ilustración 34. Estadísticos y p-valor variables dummy estacionales. Segmento 2

De los resultados anteriores se observa que ninguna variable dummy es significativa para el modelo, por lo cual no es relevante incluirlas en el modelamiento de este segmento.

3.4.3. ARIMA con variables exógenas o ARIMAX

De la misma forma que en la sección 3.3.3. ARIMA con variables exógenas o ARIMAX, se procede a incorporar las variables macroeconómicas ya mencionadas, de tal de forma de evaluar la significancia de agregarlas al modelo ya seleccionado. Los resultados son los siguientes:

```
Call:
arima(x = diff(segmento_ts_model[[2]]), order = c(6, 1, 1), xreg = cbind(depos_CLP,
    desempleo, IPC, IMACEC), method = "ML")

Coefficients:
          ar1     ar2      ar3      ar4     ar5      ar6      ma1  depos_CLP  desempleo      IPC  IMACEC
      -0.1698  0.0067  -0.1009  -0.1209  0.0547  -0.0906  -1.0000   -19.5534     0.1355  26.4939  0.2475
s.e.   0.0412  0.0418   0.0415   0.0428  0.0424   0.0417   0.0052    14.7125     9.9716  14.9527  1.3338

sigma^2 estimated as 927.4:  log likelihood = -2846.69,  aic = 5717.38

Training set error measures:
                    ME     RMSE      MAE      MPE     MAPE      MASE         ACF1
Training set 1.053315 30.42709 18.30871 143.6259 219.2669 0.6262476 -0.003394957
```

```
z test of coefficients:

             Estimate  Std. Error   z value  Pr(>|z|)
ar1        -0.1697510   0.0412004   -4.1201 3.787e-05 ***
ar2         0.0066591   0.0417935    0.1593  0.873406
ar3        -0.1009346   0.0415102   -2.4316  0.015034 *
ar4        -0.1209486   0.0428319   -2.8238  0.004746 **
ar5         0.0546512   0.0423592    1.2902  0.196987
ar6        -0.0905839   0.0417188   -2.1713  0.029909 *
ma1        -0.9999975   0.0051703 -193.4137 < 2.2e-16 ***
depos_CLP -19.5534431  14.7125404   -1.3290  0.183837
desempleo   0.1355137   9.9716489    0.0136  0.989157
IPC        26.4938825  14.9526659    1.7719  0.076419 .
IMACEC      0.2475416   1.3338180    0.1856  0.852767
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ilustración 35.Estadísticos y p-valor variables exógenas. Segmento 2

De los resultados anteriores se observa que solo la variable IPC es significativa al 90% de confianza. Si bien no es altamente significativa, se incorporará de igual manera al modelo para evaluar una posible mejora de este.

3.4.4. Proyección de los modelos

Siguiendo la misma estructura y metodología que en la sección 3.3.4. Proyección de los modelos, se muestran los gráficos proyectados a continuación:

- ARIMA(6,1,1) sin estacionalidad ni variables exógenas (sin drift)

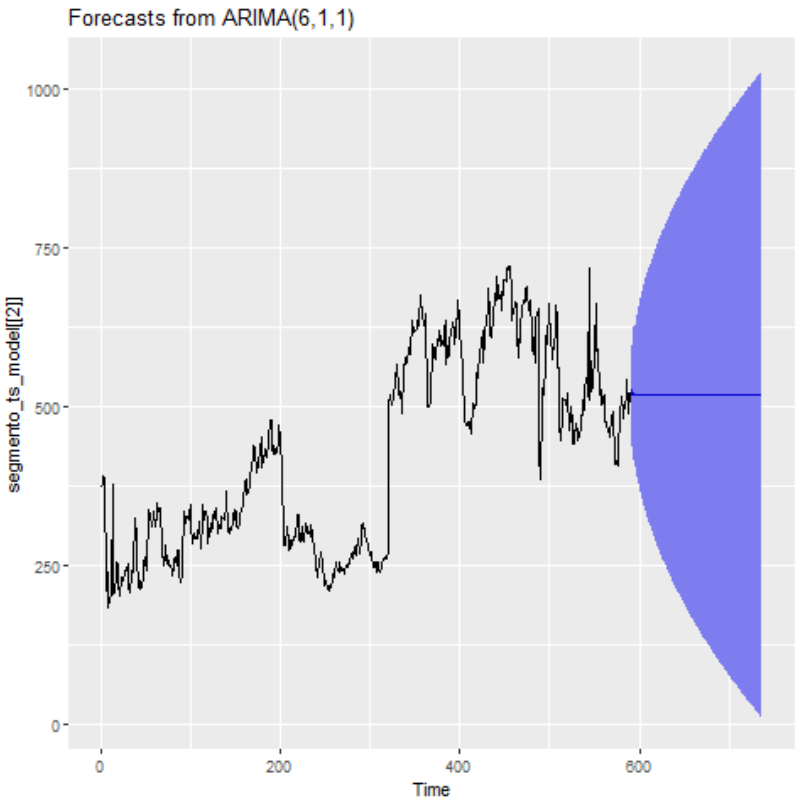

Ilustración 36. Proyección ARIMA no estacional. Segmento 2

- ARIMA(6,1,1) sin estacionalidad e IPC como variable exógena (sin drift)

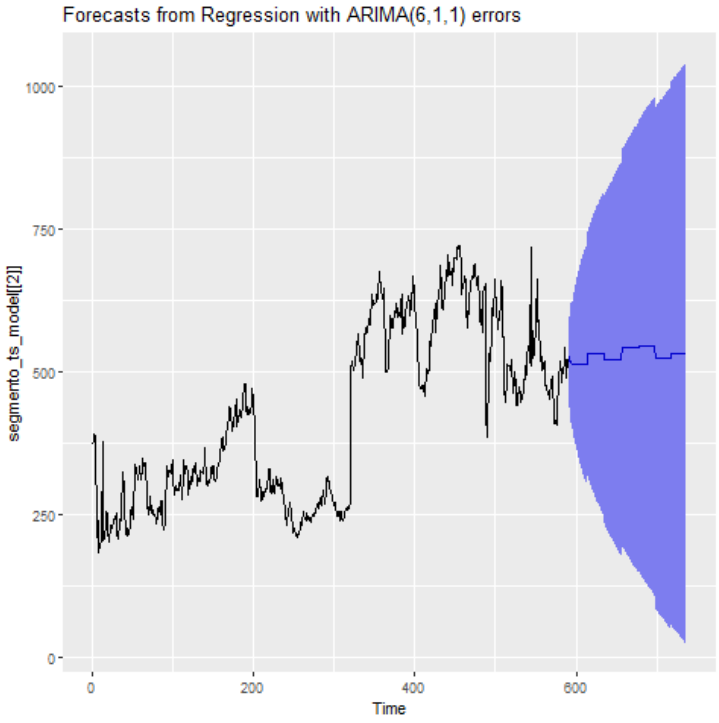

Ilustración 37. Proyección ARIMA con variable exógena. Segmento 2

Luego de esta observación gráfica de las proyecciones, se procede a comparar estos valores predichos versus la data de validación. Los resultados se muestran a continuación:

ARIMA	RMSE
(6,1,1)	87,79
(6,1,1) + IPC	100,50

Tabla 7. Resultados RSME sobre data de validación

3.5. Resultados Segmento 3

Como ya se mencionó anteriormente, el segmento 3 corresponde a la agrupación de todos los clientes dentro de la cuenta IFRS 2100109 (Cuentas corrientes de otras personas jurídicas), que tienen sus saldos en moneda USD, que sus cuentas no son remuneradas y pertenecientes a la clasificación interna de Instituciones no Financieras y PYME.

De aquí en adelante, todo el análisis y modelamiento se realiza sobre la data de modelamiento del segmento 3.

En la sección 3.1. Exploración preliminar se puede observar que el comportamiento de la serie posee una clara tendencia hacia el crecimiento, con leves caídas puntuales en sus saldos, lo cual da a inferir que la serie no es estacionaria. De forma análoga a la realizado en la sección 3.3. Resultados Segmento 1, se utilizan los test KPSS y DFA para evaluar estadísticamente la presencia o no de estacionariedad. Los resultados son los siguientes:

```
###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression none


Call:
lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)

Residuals:
     Min       1Q   Median       3Q      Max
-135.591   -7.027    0.030    7.334  125.757

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
z.lag.1    -0.0008883  0.0024491  -0.363    0.717
z.diff.lag -0.2573403  0.0357404  -7.200 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.23 on 733 degrees of freedom
Multiple R-squared:  0.06662,   Adjusted R-squared:  0.06407
F-statistic: 26.16 on 2 and 733 DF,  p-value: 1.062e-11


Value of test-statistic is: -0.3627

Critical values for test statistics:
      1pct  5pct 10pct
tau1 -2.58 -1.95 -1.62
```

Ilustración 38. Test ADF Segmento 3, serie original

```
#######################
# KPSS Unit Root Test #
#######################

Test is of type: mu with 6 lags.

Value of test-statistic is: 9.2341

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.347 0.463  0.574 0.739
```

Ilustración 39. Test KPSS Segmento 3, serie original

Para este segmento, el estadístico del test ADF es -0.3628 y el estadístico del test KPSS es 9.2341. Comparándolos contra los valores críticos se tiene que el estadístico del test ADF es mayor al de los valores críticos, por lo cual no es posible rechazar la hipótesis nula. Respecto al estadístico del test KPSS, este es mayor que los valores críticos, por ende, se rechaza la hipótesis nula en favor de la hipótesis alternativa. Así, se prueba que la serie no es estacionaria pues se comprueba la existencia de una raíz unitaria.

Dado el resultado anterior, se hace necesario diferenciar la serie de saldos de este segmento para comprobar si de esta forma la serie se vuelve estacionaria. La nueva serie de saldos del segmento 3 diferenciada posee la siguiente forma:

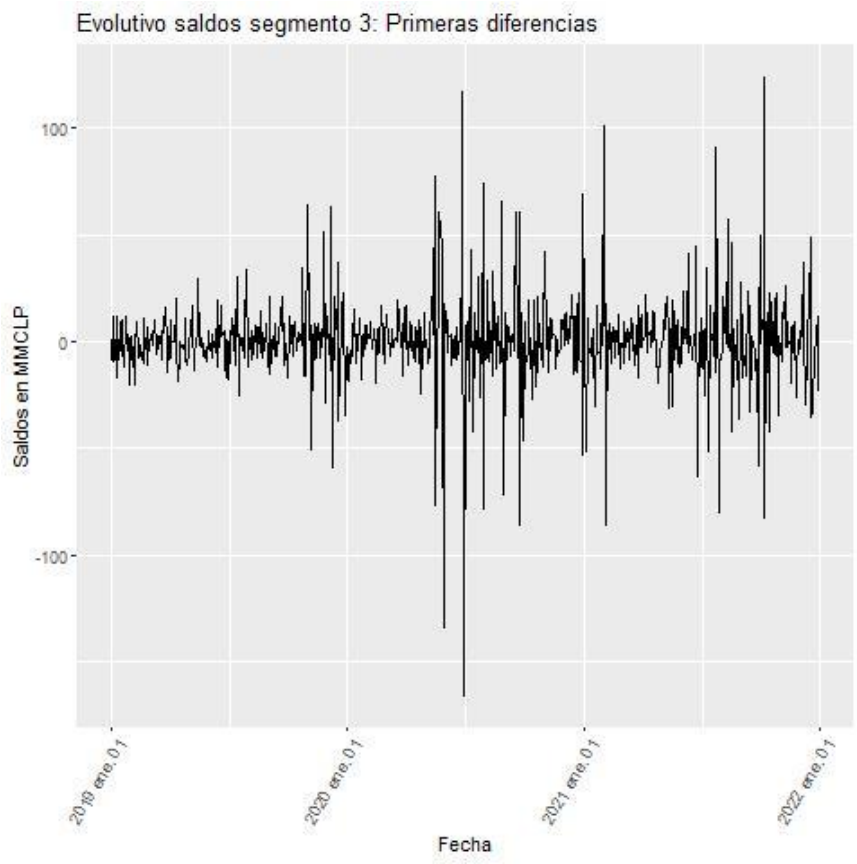

Ilustración 40. Evolutivo del segmento 3 en sus primeras diferencias

Del gráfico anterior, se observa que la serie de las diferencias está centrada en cero, con algunas variaciones más amplias que la del resto de la serie, pero sin influir en el comportamiento general. Dicho lo anterior, la serie de diferencias pareciera cumplir con las condiciones de la estacionariedad.

Para confirmar el comportamiento analizado gráficamente, se vuelven a realizar los test ADF y KPSS, pero ahora sobre la serie de diferencias. Los resultados son los siguientes:

```
###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression none


Call:
lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)

Residuals:
     Min       1Q   Median       3Q      Max
-134.636   -7.540   -0.284    6.789  126.538

Coefficients:
           Estimate Std. Error t value Pr(>|t|)
z.lag.1    -1.34730    0.05860 -22.992   <2e-16 ***
z.diff.lag  0.07118    0.03693   1.927   0.0543 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.19 on 732 degrees of freedom
Multiple R-squared:  0.6305,    Adjusted R-squared:  0.6295
F-statistic: 624.5 on 2 and 732 DF,  p-value: < 2.2e-16


Value of test-statistic is: -22.9915

Critical values for test statistics:
      1pct  5pct 10pct
tau1 -2.58 -1.95 -1.62
```

Ilustración 41.Test ADF Segmento 3, serie diferenciada

```
#######################
# KPSS Unit Root Test #
#######################

Test is of type: mu with 6 lags.

Value of test-statistic is: 0.0238

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.347 0.463  0.574 0.739
```

Ilustración 42.Test KPSS Segmento 3, serie diferenciada

Con la diferenciación de la serie, se genera una modificación importante en los estadísticos KPSS y ADF. En el caso del test ADF, su valor es de -22.9915, menor que todos los valores críticos (o, dicho de otra forma, es más negativo que los valores críticos) por lo cual se rechaza la hipótesis nula, en favor de la alternativa la cual indica que la serie es estacionaria. Respecto al test KPSS, el valor del estadístico es de 0.0238, menor que los valores críticos para todos los niveles de significancia, lo cual implica que no es posible rechazar la hipótesis nula y de este modo la serie es estacionaria o posee tendencia estacionaria. Complementando ambos test se puede concluir que la serie de diferencias de la serie original de saldos del segmento 3 es estacionaria, lo cual da el paso para aplicar los modelos ARIMA al cumplir la condición necesaria de estacionariedad.

3.5.1. ARIMA sin estacionalidad

De forma análoga a lo realizado en la sección 3.3.1. ARIMA sin estacionalidad, el siguiente paso es encontrar los parámetros *p* y *q*, pues gracias a la diferenciación el parámetro *d* es igual a 1. Para lograr lo anterior se utilizan las funciones ACF y PACF, las cuales se muestran a continuación:

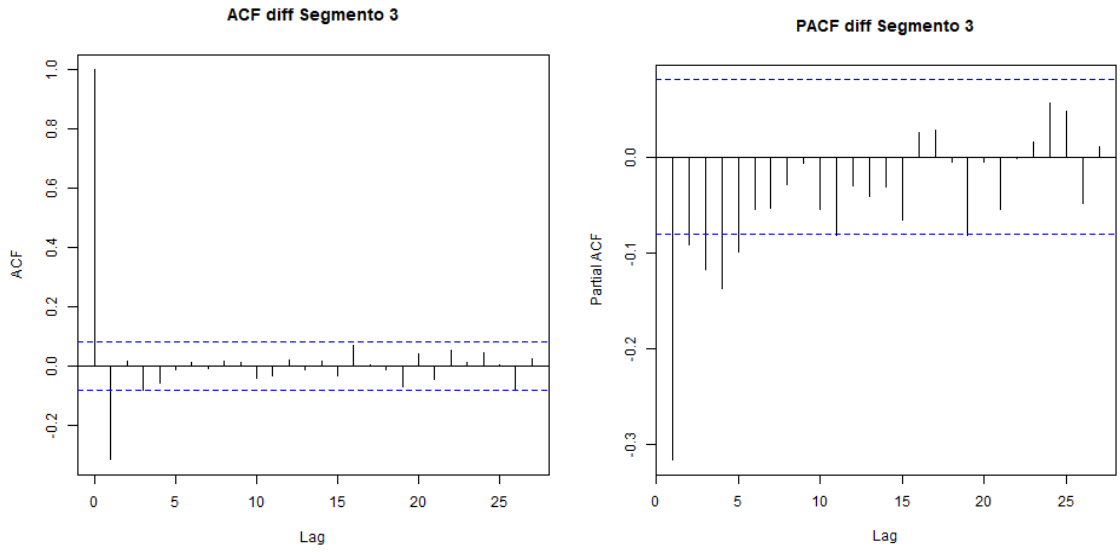

Ilustración 43. Gráficos ACF y PACF Segmento 3, primeras diferencias

De los gráficos anteriores se pueden comentar las siguientes observaciones:

- Respecto al gráfico de ACF, solo el *lag* 1 es significativo.
- Respecto al gráfico de PACF, todos los *lags* son significativos hasta el *lag* 5.

De acuerdo con las significancias identificadas, se evalúan las posibles combinaciones para AR(1), AR(2), AR(3), AR(4) y AR(5) respecto a su componente autorregresiva, y MA(1) respecto a su componente de media móvil. Así, los parámetros ARIMA a evaluar son los siguientes:

- ARIMA(1,1,1)
- ARIMA(2,1,1)
- ARIMA(3,1,1)
- ARIMA(4,1,1)
- ARIMA(5,1,1)

A continuación, se muestra una tabla en donde se muestran los criterios AIC,BIC y Error Cuadrático Medio para cada uno de los parámetros ARIMA indicados anteriormente.

ARIMA	AIC	BIC	RMSE
(1,1,1)	5.123	5.136	18,65
(2,1,1)	5.121	5.138	18,57
(3,1,1)	5.115	5.137	18,44
(4,1,1)	5.106	5.132	18,26
(5,1,1)	5.102	5.133	18,18

Tabla 8. Criterios AIC, BIC y RMSE. Segmento 3

De la tabla anterior se observa que no hay un ARIMA que sea el mínimo en los 3 criterios evaluados. Así, los posibles candidatos son:

- Menor AIC: ARIMA(5,1,1)
- Menor BIC: ARIMA(4,1,1)
- Menor RMSE: ARIMA(5,1,1)

Considerando al BIC como el indicador principal, el ARIMA a escoger sería el ARIMA(4,1,1), sin embargo, el evaluar conjuntamente tanto AIC y BIC, el mejor ARIMA pareciese ser el ARIMA(5,1,1), debido a que se gana más en AIC que lo que pierde en BIC. El criterio de RMSE se considera como un tercer criterio y en este caso coincide con el criterio de AIC, por lo cual aporta a la elección de estos parámetros. Así, dicho lo anterior, el ARIMA seleccionado para este segmento es el ARIMA(5,1,1).

A modo comparativo y de incluir una segunda opinión dentro de la selección de parámetros, se ejecuta la función auto.arima. El resultado es el siguiente:

```
Series: segmento_ts_model[[3]]
ARIMA(2,1,1)

Coefficients:
         ar1     ar2      ma1
      0.4789  0.1386  -0.8880
s.e.  0.0578  0.0488   0.0387

sigma^2 estimated as 328.4:  log likelihood=-2540.85
AIC=5089.69   AICc=5089.76   BIC=5107.2

Training set error measures:
                     ME     RMSE      MAE          MPE     MAPE     MASE        ACF1
Training set 0.6879808 18.05956 10.98249 -0.08968533 3.943972 1.012982 0.00372885
```

Ilustración 44. Resultados auto.arima, Segmento 3

Se observa que los parámetros difieren en su componente autorregresiva, a un grado mayor a lo observado en los segmentos anteriores. Adicionalmente, los criterios AIC y BIC son menores a todos los posibles encontrados con la metodología propuesta. La disparidad de estos resultados se puede deber a que al existir tan pocos *lags* significativos y, por ende, menores combinaciones para evaluar, se resalta la diferencia entre lo analizado en este trabajo y lo entregado por el *software*. Sin perjuicio de lo anterior, los parámetros escogidos siguen siendo el del ARIMA(5,1,1), ya que como se indicó anteriormente, este análisis es solo un complemento y no es vinculante en la decisión de los parámetros.

Por último, se realiza el análisis sobre los residuos, de tal forma de corroborar si cumplen con las características que lo hacen ser ruido blanco. Este análisis se muestra a continuación:

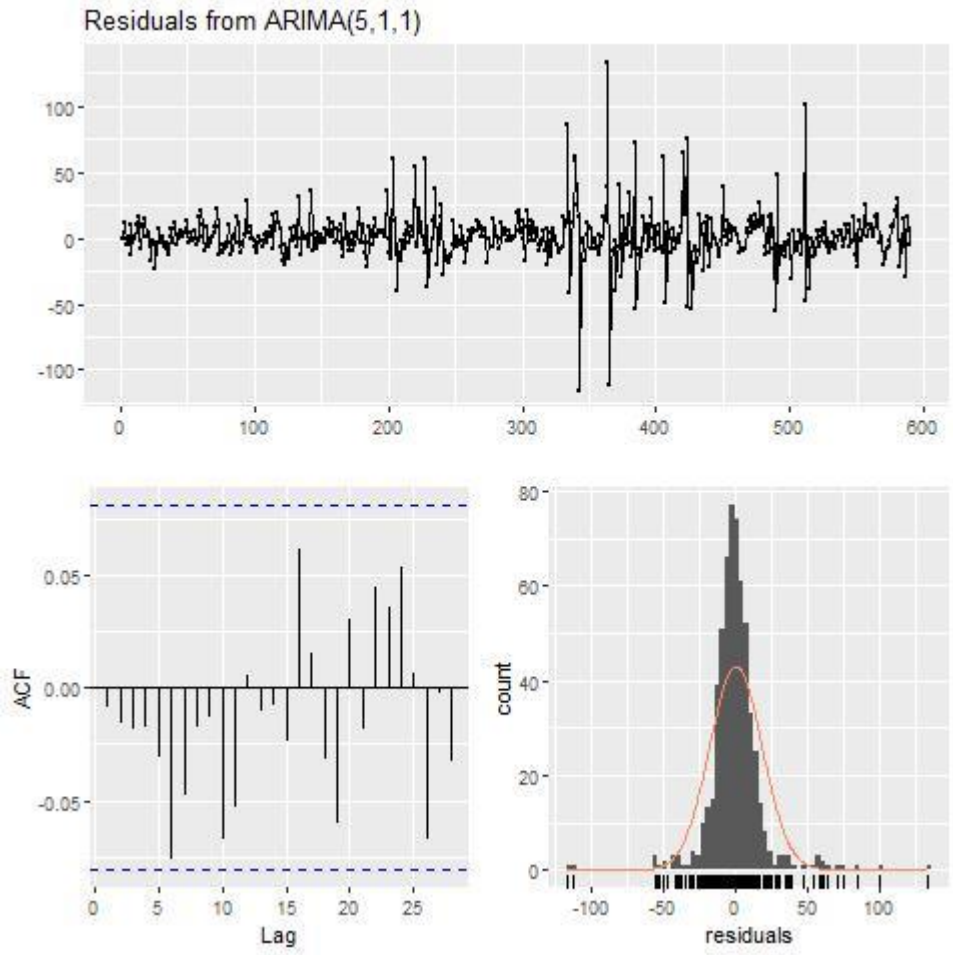

Ilustración 45. Análisis de residuos. Segmento 3

Del gráfico anterior se puede observar lo siguiente:

- Los residuos están centrados en cero, con una varianza relativamente constante a lo largo del tiempo, excepto en un par de fechas muy puntuales.
- Los residuos siguen una distribución muy similar a la distribución normal, donde la gran mayoría de los datos está muy concentrada en cero.
- Del gráfico ACF se observa que no existen *lags* significativos, todos están dentro de la banda de significancia, implicando que no hay pérdida de información en el modelo.

Dicho lo anterior, los residuos parecen cumplir a cabalidad las condiciones para ser considerado ruido blanco. A modo complementario, se realiza el test Ljung-Box, para confirmar lo anterior. Los resultados son los siguientes:

```
        Ljung-Box test

data:  Residuals from ARIMA(5,1,1)
Q* = 9.0056, df = 4, p-value = 0.06096

Model df: 6.   Total lags used: 10
```

Ilustración 46. Test Ljung-Box. Segmento 3

Se observa que el p-valor es de 0.06096, levemente mayor que el nivel de significancia del 0.05, un resultado un poco contraintuitivo debido al buen comportamiento de los residuos observados en el análisis anterior. Independiente de lo anterior, de todas formas, el p-valor es mayor al nivel de significancia, lo cual indica que la hipótesis nula no puede ser rechazada, implicando que los residuos se distribuyen independientemente, y por ende son ruido blanco.

3.5.2. ARIMA estacional o SARIMA

De la misma forma que en la sección 3.3.2. ARIMA estacional o SARIMA, se evalúa la significancia de agregar al modelo las variables dummies asociadas a los días calendario previamente definidos. Los resultados son los siguientes:

```
Call:
arima(x = segmento_ts_model[[3]], order = c(5, 1, 1), xreg = base_model[, c(8,
    12, 22, 32, 37, 38)], method = "ML")

Coefficients:
         ar1     ar2      ar3      ar4     ar5     ma1  Dummy_dia_1  Dummy_dia_5  Dummy_dia_15
      0.4952  0.1675  -0.0613  -0.0224  0.0533  -0.891     -11.1905      -2.3229        0.3520
s.e.  0.0752  0.0528   0.0481   0.0488  0.0488   0.062       4.0070       3.2921        3.3807
      Dummy_dia_25  Dummy_dia_30  Dummy_dia_31
           -2.1152       -3.0028       -8.5279
s.e.        3.3956        3.5750        5.5301

sigma^2 estimated as 320.3:  log likelihood = -2535.03,  aic = 5096.07

Training set error measures:
                    ME     RMSE      MAE         MPE     MAPE     MASE         ACF1
Training set 0.6789604 17.88199 11.03235 -0.08873259 3.979726 1.01758 -0.003040469
```

```
z test of coefficients:

               Estimate Std. Error  z value  Pr(>|z|)
ar1            0.495178   0.075237   6.5816 4.654e-11 ***
ar2            0.167452   0.052801   3.1714  0.001517 **
ar3           -0.061307   0.048055  -1.2758  0.202038
ar4           -0.022386   0.048760  -0.4591  0.646155
ar5            0.053310   0.048803   1.0924  0.274678
ma1           -0.890962   0.061958 -14.3800 < 2.2e-16 ***
Dummy_dia_1  -11.190492   4.007014  -2.7927  0.005227 **
Dummy_dia_5   -2.322850   3.292112  -0.7056  0.480449
Dummy_dia_15   0.352015   3.380693   0.1041  0.917070
Dummy_dia_25  -2.115221   3.395642  -0.6229  0.533336
Dummy_dia_30  -3.002827   3.575044  -0.8399  0.400941
Dummy_dia_31  -8.527934   5.530068  -1.5421  0.123049
```

Ilustración 47. Estadísticos y p-valor variables dummy estacionales. Segmento 3

De los resultados anteriores se observa que ninguna variable dummy es significativa para el modelo, por lo cual no es relevante incluirlas en el modelamiento de este segmento.

3.5.3. ARIMA con variables exógenas o ARIMAX

De la misma forma que en la sección 3.3.3. ARIMA con variables exógenas o ARIMAX, se procede a incorporar las variables macroeconómicas ya mencionadas, de tal de forma de evaluar la significancia de agregarlas al modelo ya seleccionado. Los resultados son los siguientes:

```
Call:
arima(x = diff(segmento_ts_model[[3]]), order = c(5, 1, 1), xreg = cbind(depos_USD,
    desempleo, IPC, IMACEC), method = "ML")

Coefficients:
          ar1      ar2      ar3      ar4      ar5      ma1  depos_USD  desempleo       IPC    IMACEC
      -0.4046  -0.1662  -0.1710  -0.1688  -0.0939  -1.0000     6.8890   -11.7741  -24.8833  -1.5775
s.e.   0.0415   0.0439   0.0441   0.0440   0.0413   0.0071     4.0746     5.2738    8.1834   0.7221

sigma^2 estimated as 319.4:  log likelihood = -2533.68,  aic = 5089.36

Training set error measures:
                    ME     RMSE      MAE      MPE    MAPE      MASE         ACF1
Training set 0.8743035 17.85616 11.12214 66.10057 262.474 0.6242515 -0.008513342
```

```
z test of coefficients:

             Estimate  Std. Error   z value  Pr(>|z|)
ar1        -0.4045982   0.0414795   -9.7542 < 2.2e-16 ***
ar2        -0.1661791   0.0439089   -3.7846 0.0001539 ***
ar3        -0.1710133   0.0440777   -3.8798 0.0001045 ***
ar4        -0.1688065   0.0440347   -3.8335 0.0001263 ***
ar5        -0.0938724   0.0413279   -2.2714 0.0231223 *
ma1        -0.9999989   0.0070621 -141.6001 < 2.2e-16 ***
depos_USD   6.8889569   4.0745928    1.6907 0.0908921 .
desempleo -11.7740974   5.2737554   -2.2326 0.0255764 *
IPC       -24.8832523   8.1833664   -3.0407 0.0023602 **
IMACEC     -1.5774596   0.7221008   -2.1845 0.0289224 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ilustración 48.Estadísticos y p-valor variables exógenas. Segmento 3

De los resultados anteriores se observa que todas las variables son significativas: el IPC es significativo al 99% de confianza y con su estimador negativo; la tasa de desempleo y el IMACEC son significativos al 95% y con sus estimadores negativos; la tasa de los depósitos a plazo en USD es significativa al 90% y con su estimador positivo. Analizando el signo de los estimadores, se puede decir que los estimadores con signo negativo tienen sentido financiero (a excepción de IMACEC), ya que, si aumenta el desempleo, entonces los clientes del banco tendrán menos dinero en sus cuentas, porque probablemente haya una mayor cantidad de estos que esté desempleado. Además, un aumento en el IPC significa que el costo de la vida es mayor y el dinero pierde cada vez más valor, por lo cual mantenerlo en las cuentas corrientes es muy costoso ya que se devalúa rápidamente. Por último, el signo de la variable IMACEC pareciera no ser consiste con la lógica financiera, en donde un aumento de la actividad económica implicaría una mejora en la condición económica de las personas. Una posible respuesta a esta potencial inconsistencia es que el producto asociado al segmento 3 son cuentas en dólares, por ende, no podría verse afectado de una manera directa ante variaciones del IMACEC. Otra posible interpretación a este resultado es que, si la actividad económica del país aumenta, será más atractivo invertir en proyectos, generando una potencial salida de estos saldos. Independiente de todo lo anterior, el estimador es pequeño (en valor absoluto) por ende no aporta de forma sustancial a la explicación de la cantidad de saldos en las cuentas.

3.5.4. Proyección de los modelos

Siguiendo la misma estructura y metodología que en la sección 3.3.4. Proyección de los modelos, se muestran los gráficos proyectados a continuación:

- ARIMA(5,1,1) sin estacionalidad ni variables exógenas (sin drift)

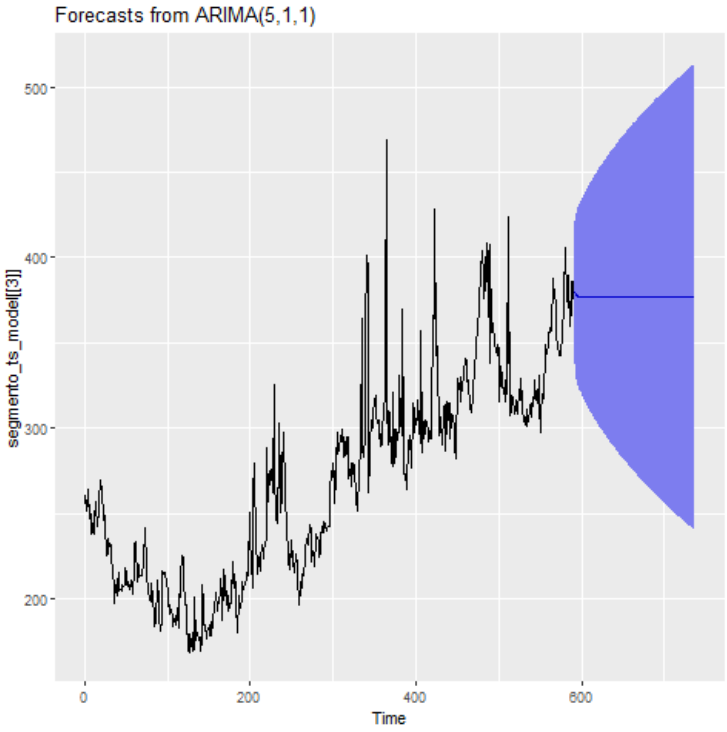

Ilustración 49. Proyección ARIMA no estacional. Segmento 3

- ARIMA(5,1,1) sin estacionalidad y variables: Tasa de depósitos en USD, IPC, IMACEC, desempleo (sin drift)

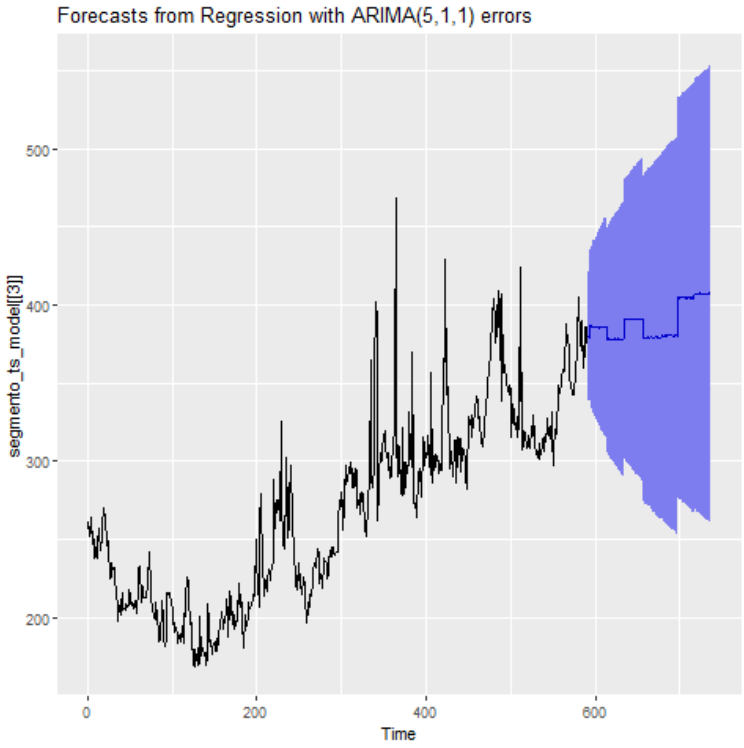

Ilustración 50. Proyección ARIMA con variable exógena. Segmento 3

Luego de esta observación gráfica de las proyecciones, se procede a comparar estos valores predichos versus la data de validación. Los resultados se muestran a continuación:

ARIMA	RMSE
(5,1,1)	46,54
(5,1,1) + IPC, Desempleo, IMACEC, depos_USD	38,65

Tabla 9. Resultados RSME sobre data de validación

3.6. Resultados Segmento 4

Como ya se mencionó anteriormente, el segmento 4 corresponde a la agrupación de todos los clientes dentro de la cuenta IFRS 2100110 (Cuentas corrientes de personas naturales), que tienen sus saldos en moneda CLP, sin diferenciación por remuneración de cuenta y pertenecientes a la clasificación interna de Personas naturales y PYME.

De aquí en adelante, todo el análisis y modelamiento se realiza sobre la data de modelamiento del segmento 4.

En la sección 3.1. Exploración preliminar se puede observar que el comportamiento de la serie posee una clara tendencia hacia el crecimiento, sin embargo, a diferencia de los segmentos estudiados anteriormente, se identifica una potencial estacionalidad en la serie, lo cual se observa en la forma de "serrucho" que posee el gráfico, es decir *peaks* que se van alcanzando de forma casi equiespaciada y luego una caída que lo lleva al mismo nivel que antes del *peak*. Este comportamiento puede generar alguna diferencia relevante al momento de modelar este segmento. Eso sí, independiente de este nuevo comportamiento la serie sigue pareciendo ser no estacionaria. De forma análoga a la realizado en la sección 3.3. Resultados Segmento 1, se utilizan los test KPSS y DFA para evaluar estadísticamente la presencia o no de estacionariedad. Los resultados son los siguientes:

```
###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression none


Call:
lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)

Residuals:
    Min      1Q  Median      3Q     Max
-89.828  -9.558  -1.393   8.600  77.148

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
z.lag.1    0.0008035  0.0007362   1.091  0.27546
z.diff.lag 0.0983198  0.0370820   2.651  0.00819 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.08 on 733 degrees of freedom
Multiple R-squared:  0.01159,   Adjusted R-squared:  0.00889
F-statistic: 4.296 on 2 and 733 DF,  p-value: 0.01396


Value of test-statistic is: 1.0914

Critical values for test statistics:
      1pct  5pct 10pct
tau1 -2.58 -1.95 -1.62
```

Ilustración 51. Test ADF Segmento 4, serie original

```
######################
# KPSS Unit Root Test #
######################

Test is of type: mu with 6 lags.

Value of test-statistic is: 10.372

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.347 0.463  0.574 0.739
```

Ilustración 52. Test KPSS Segmento 4, serie original

Para este segmento, el estadístico del test ADF es 1.0914 y el estadístico del test KPSS es 10.372. Comparándolos contra los valores críticos se tiene que el estadístico del test ADF es mayor al de los valores críticos, por lo cual no es posible rechazar la hipótesis nula. Respecto al estadístico del test KPSS, este es mayor que los valores críticos, por ende, se rechaza la hipótesis nula en favor de la hipótesis alternativa. Así, se prueba que la serie no es estacionaria pues se comprueba la existencia de una raíz unitaria.

Dado el resultado anterior, se hace necesario diferenciar la serie de saldos de este segmento para comprobar si de esta forma la serie se vuelve estacionaria. La nueva serie de saldos del segmento 4 diferenciada posee la siguiente forma:

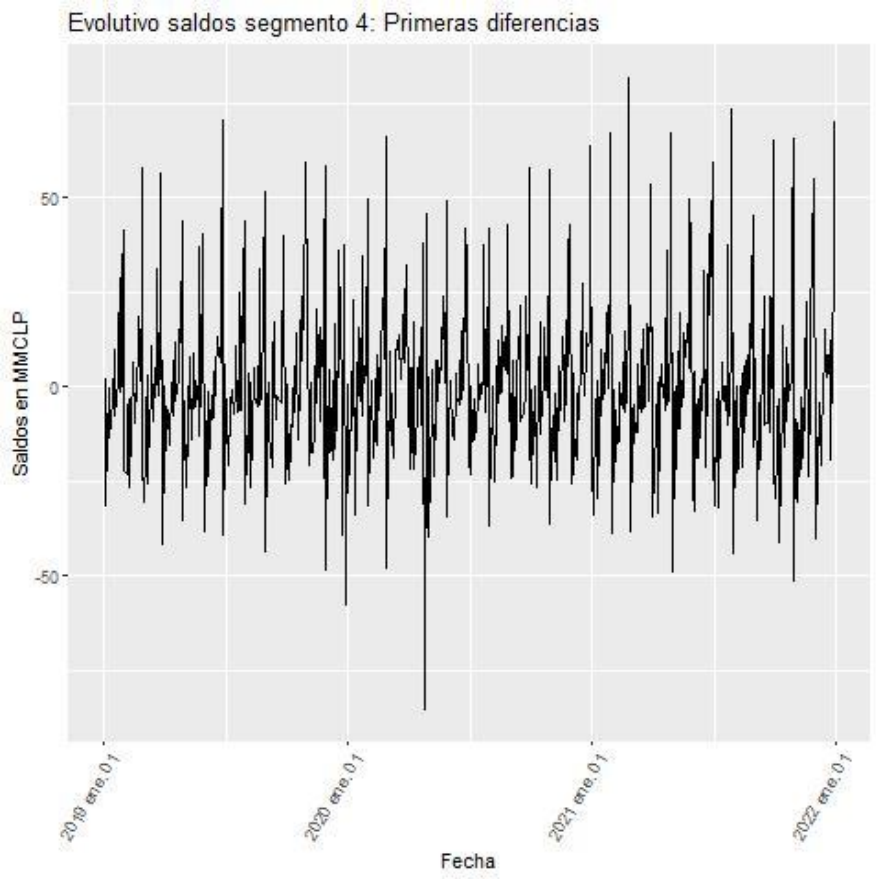

Ilustración 53. Evolutivo del segmento 4 en sus primeras diferencias

Del gráfico anterior, se observa que la serie de las diferencias está centrada en cero, con variaciones que poseen alguna forma estacional, pero lo cual no influye en el comportamiento general. Dicho lo anterior, la serie de diferencias pareciera cumplir con las condiciones de la estacionariedad.

Para confirmar el comportamiento analizado gráficamente, se vuelven a realizar los test ADF y KPSS, pero ahora sobre la serie de diferencias. Los resultados son los siguientes:

```
###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression none


Call:
lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)

Residuals:
    Min      1Q  Median      3Q     Max
-90.341  -8.766  -0.670   9.438  78.602

Coefficients:
           Estimate Std. Error t value Pr(>|t|)
z.lag.1    -0.92883    0.05014 -18.525   <2e-16 ***
z.diff.lag  0.03274    0.03727   0.878     0.38
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.09 on 732 degrees of freedom
Multiple R-squared:  0.446,     Adjusted R-squared:  0.4445
F-statistic: 294.7 on 2 and 732 DF,  p-value: < 2.2e-16


Value of test-statistic is: -18.5252

Critical values for test statistics:
      1pct  5pct 10pct
tau1 -2.58 -1.95 -1.62
```

Ilustración 54.Test ADF Segmento 4, serie diferenciada

```
#######################
# KPSS Unit Root Test #
#######################

Test is of type: mu with 6 lags.

Value of test-statistic is: 0.0456

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.347 0.463  0.574 0.739
```

Ilustración 55.Test KPSS Segmento 4, serie diferenciada

Con la diferenciación de la serie, se genera una modificación importante en los estadísticos KPSS y ADF. En el caso del test ADF, su valor es de -18.525, menor que todos los valores críticos (o, dicho de otra forma, es más negativo que los valores críticos) por lo cual se rechaza la hipótesis nula, en favor de la alternativa la cual indica que la serie es estacionaria. Respecto al test KPSS, el

valor del estadístico es de 0.0456, menor que los valores críticos para todos los niveles de significancia, lo cual implica que no es posible rechazar la hipótesis nula y de este modo la serie es estacionaria o posee tendencia estacionaria. Complementando ambos test se puede concluir que la serie de diferencias de la serie original de saldos del segmento 4 es estacionaria, lo cual da el paso para aplicar los modelos ARIMA al cumplir la condición necesaria de estacionariedad.

3.6.1. ARIMA sin estacionalidad

De forma análoga a lo realizado en la sección 3.3.1. ARIMA sin estacionalidad, el siguiente paso es encontrar los parámetros *p* y *q*, pues gracias a la diferenciación el parámetro *d* es igual a 1. Para lograr lo anterior se utilizan las funciones ACF y PACF, las cuales se muestran a continuación:

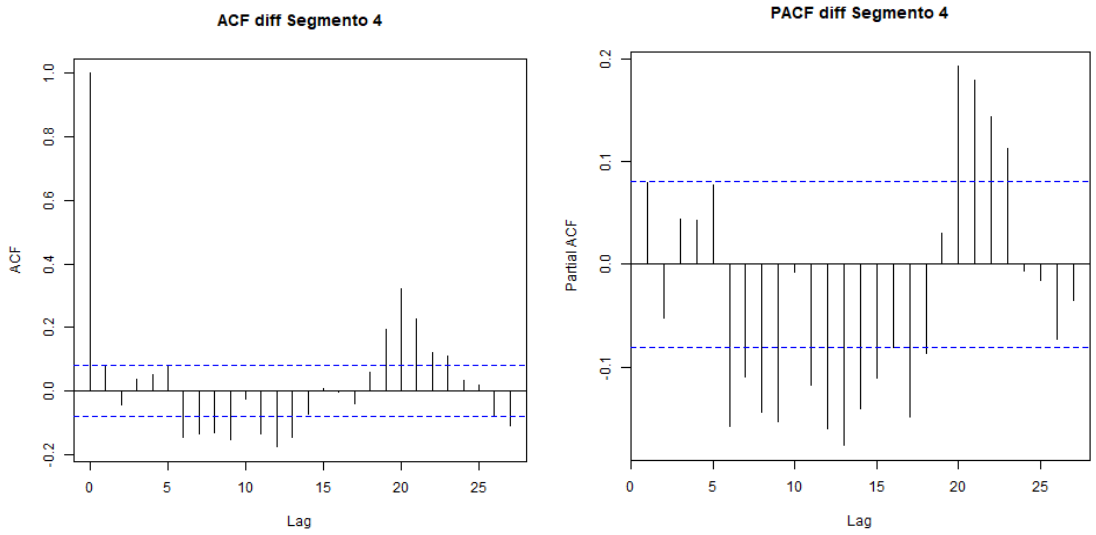

Ilustración 56. Gráficos ACF y PACF Segmento 4, primeras diferencias

De los gráficos anteriores se pueden comentar las siguientes observaciones:

- Ambos gráficos comparten un comportamiento similar, en donde bajo el *lag* 5 solo el *lag* 1 y el *lag* 5 están al borde de la significancia, mientras que sobre el *lag* 5 hay muchos *lags* significativos y con algún comportamiento estacional, ya que estos *lags* significativos se concentran en los múltiplos de 10.

Debido a este comportamiento inusual hasta ahora, en esta sección se considerarán solo los *lags* 1 y 5 tanto para la función ACF como para la PACF, dejando el potencial comportamiento estacional para secciones más adelantes de este trabajo. Así, de acuerdo con las significancias escogidas, se evalúan las posibles combinaciones para AR(1) y AR(5) respecto a su componente

autorregresiva, y MA(1) y MA(5) respecto a su componente de media móvil. Así, los parámetros ARIMA a evaluar son los siguientes:

- ARIMA(1,1,1)
- ARIMA(1,1,5)
- ARIMA(5,1,1)
- ARIMA(5,1,5)

A continuación, se muestra una tabla en donde se muestran los criterios AIC,BIC y Error Cuadrático Medio para cada uno de los parámetros ARIMA indicados anteriormente.

ARIMA	AIC	BIC	RMSE
(1,1,1)	5.175	5.188	19,50
(5,1,1)	5.176	5.206	19,38
(1,1,5)	5.180	5.210	19,44
(5,1,5)	5.159	5.207	18,97

Tabla 10. Criterios AIC, BIC y RMSE. Segmento 4

De la tabla anterior se observa que no hay un ARIMA que sea el mínimo en los 3 criterios evaluados. Así, los posibles candidatos son:

- Menor AIC: ARIMA(5,1,5)
- Menor BIC: ARIMA(1,1,1)
- Menor RMSE: ARIMA(5,1,5)

Considerando al BIC como el indicador principal, el ARIMA a escoger sería el ARIMA(1,1,1). A diferencia de los segmentos anteriores, en donde un análisis conjunto entre AIC y BCI llevó a escoger los parámetros ARIMA que tuvieran el menor AIC, en este caso al cambiar de ARIMA (1,1,1) a ARIMA(5,1,5) se pierde más información, por lo cual se selecciona el ARIMA(1,1,1). Esta elección implica escoger el RMSE más alto, sin embargo, es algo lógico ya que la menor cantidad de parámetros hacen que este criterio sea mayor, pero en este caso el criterio más relevante es el BIC.

A modo comparativo y de incluir una segunda opinión dentro de la selección de parámetros, se ejecuta la función auto.arima. El resultado es el siguiente:

```
Series: segmento_ts_model[[4]]
ARIMA(2,1,2) with drift

Coefficients:
         ar1     ar2      ma1      ma2   drift
      0.4477  0.3563  -0.4204  -0.5290  1.0619
s.e.  0.9855  0.8989   0.9304   0.9075  0.2116

sigma^2 estimated as 366.8:  log likelihood=-2572.5
AIC=5157.01   AICc=5157.15   BIC=5183.28

Training set error measures:
                     ME     RMSE      MAE         MPE     MAPE      MASE        ACF1
Training set -0.2103869 19.05519 13.37485 -0.1142921 1.572751 0.9715046 0.01860776
```

Ilustración 57. Resultados auto.arima, Segmento 4

Se observa que los parámetros difieren en su componente autorregresiva y de media móvil en un grado, lo cual es relativamente cercano a lo obtenido mediante la metodología propuesta en este trabajo. Adicionalmente, los criterios AIC y BIC son menores a todos los posibles encontrados con la metodología propuesta, pero en general los valores son bastante cercanos. Dicho lo anterior, los parámetros escogidos siguen siendo el del ARIMA(1,1,1), ya que como se indicó anteriormente, este análisis es solo un complemento y no es vinculante en la decisión de los parámetros.

Por último, se realiza el análisis sobre los residuos, de tal forma de corroborar si cumplen con las características que lo hacen ser ruido blanco. Este análisis se muestra a continuación:

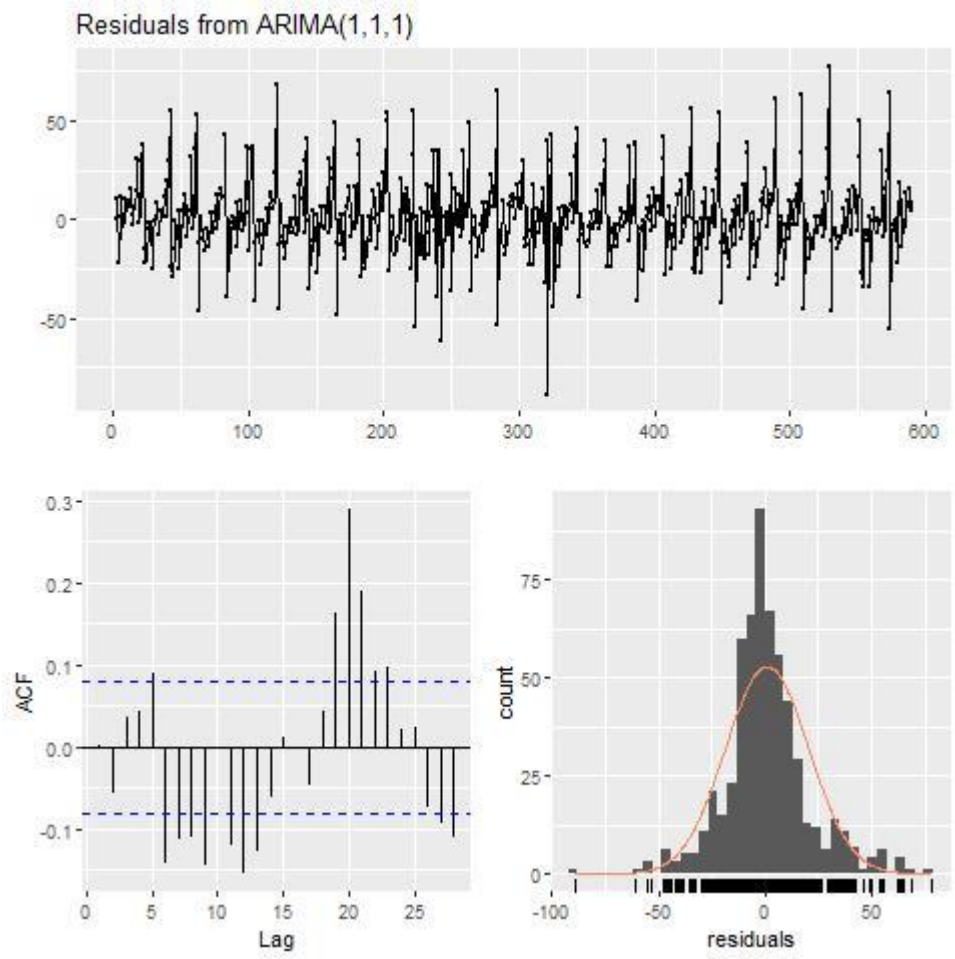

Ilustración 58. Análisis de residuos. Segmento 4

Del gráfico anterior se puede observar lo siguiente:

- Los residuos están centrados en cero, con un comportamiento estacional de los residuos, muy similar al comportamiento de la serie original de saldos.
- Los residuos siguen una distribución muy similar a la distribución normal, donde la gran mayoría de los datos está muy concentrada en cero.
- Del gráfico ACF se observa el comportamiento observado en el análisis previamente realizado, *lags* en forma de onda y en su mayoría significativos, implicando que hay información en los residuos que debiese estar en el modelo, por lo cual nos indica que la serie de residuos no es un ruido blanco.

Dicho lo anterior, se realiza el test Ljung-Box, para confirmar que los residuos no se comportan como ruido blanco. Los resultados son los siguientes:

```
        Ljung-Box test

data:  Residuals from ARIMA(1,1,1)
Q* = 47.826, df = 8, p-value = 1.067e-07

Model df: 2.   Total lags used: 10
```

Ilustración 59. Test Ljung-Box. Segmento 4

Se observa que el p-valor recién tiene su decimal distinto de cero en la séptima posición, un valor casi 0 y muy bajo comparado con el nivel de significancia de 0.05. Esto nos indica que se rechaza la hipótesis nula, aceptando la hipótesis alternativa de que los residuos no son independientes y por ende no son ruido blanco. Este resultado está en línea con lo obtenido a lo largo de todo el estudio de este segmento, por lo cual se hará necesario evaluar nuevos parámetros o variables a incorporar al modelo, de tal forma de rescatar toda la información necesaria y que los residuos solo sean ruido blanco.

3.6.2. ARIMA estacional o SARIMA

De la misma forma que en la sección 3.3.2. ARIMA estacional o SARIMA, se evalúa la significancia de agregar al modelo las variables dummies asociadas a los días calendario previamente definidos. Los resultados son los siguientes:

```
Call:
arima(x = segmento_ts_model[[4]], order = c(1, 1, 1), xreg = base_model[, c(8,
    12, 22, 32, 37, 38)], method = "ML")

Coefficients:
         ar1      ma1  Dummy_dia_1  Dummy_dia_5  Dummy_dia_15  Dummy_dia_25  Dummy_dia_30  Dummy_dia_31
      0.5219  -0.4344       3.7475      -4.4538        2.4255        4.4939       29.3325       47.3055
s.e.  0.2023   0.2130       3.2586       2.5367        2.5964        2.5944        2.8887        4.5378

sigma^2 estimated as 294.1:  log likelihood = -2509.63,  aic = 5037.27

Training set error measures:
                    ME    RMSE     MAE        MPE     MAPE      MASE         ACF1
Training set 0.8676275 17.1336 11.8675 0.07012216 1.389475 0.8620157 -0.002076938
```

```
z test of coefficients:

             Estimate Std. Error z value  Pr(>|z|)
ar1           0.52186    0.20232  2.5794  0.009898 **
ma1          -0.43438    0.21303 -2.0390  0.041448 *
Dummy_dia_1   3.74750    3.25856  1.1500  0.250125
Dummy_dia_5  -4.45382    2.53670 -1.7558  0.079130 .
Dummy_dia_15  2.42546    2.59639  0.9342  0.350218
Dummy_dia_25  4.49393    2.59440  1.7322  0.083244 .
Dummy_dia_30 29.33253    2.88866 10.1544 < 2.2e-16 ***
Dummy_dia_31 47.30552    4.53780 10.4248 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ilustración 60. Estadísticos y p-valor variables dummy estacionales. Segmento 4

De los resultados anteriores, se observa que las dummies asociadas al día 30 y 31 son significativas al 99,9% y que las dummies asociadas a los días 5 y 25 son significativas al 90%. Respecto al coeficiente de estas variables, las dummies del día 25, 30 y 31 son positivos, mientras que el coeficiente de la dummy del día 5 es negativo. Considerando que este segmento está compuesto por cuentas corrientes de personas naturales, con cuentas en moneda CLP, es un resultado lógico respecto a la hipótesis planteada inicialmente, la cual indica que en fechas históricamente de pago de sueldos (fin de mes y día 25 con menos significancia) aumenta el saldo en la cuenta corriente de las personas contratadas e inmersas en el mundo laboral. Mientras que en fechas históricamente asociadas al pago de deudas (día 5 del mes), existe una fuga en el saldo de estas cuentas.

3.6.3. ARIMA con variables exógenas o ARIMAX

De la misma forma que en la sección 3.3.3. ARIMA con variables exógenas o ARIMAX, se procede a incorporar las variables macroeconómicas ya mencionadas, de tal de forma de evaluar la significancia de agregarlas al modelo ya seleccionado. Los resultados son los siguientes:

```
Call:
arima(x = diff(segmento_ts_model[[4]]), order = c(1, 1, 1), xreg = cbind(depos_CLP,
    desempleo, IPC, IMACEC), method = "ML")

Coefficients:
         ar1      ma1  depos_CLP  desempleo       IPC   IMACEC
      0.1003  -1.0000     3.1165   -17.4415  -12.0644  -2.5028
s.e.  0.0414   0.0059     8.8582     6.2469    9.6884   0.8667

sigma^2 estimated as 371.4:  log likelihood = -2577.08,  aic = 5168.17

Training set error measures:
                    ME     RMSE      MAE      MPE     MAPE      MASE         ACF1
Training set 0.9399249 19.25468 13.46505 82.03909 147.5485 0.7566104 0.0008352926
```

```
z test of coefficients:

             Estimate  Std. Error   z value  Pr(>|z|)
ar1         0.1002831   0.0414197    2.4211  0.015472 *
ma1        -0.9999999   0.0059045 -169.3620 < 2.2e-16 ***
depos_CLP   3.1165195   8.8581923    0.3518  0.724971
desempleo -17.4414897   6.2469259   -2.7920  0.005238 **
IPC       -12.0643548   9.6884355   -1.2452  0.213046
IMACEC     -2.5027841   0.8666685   -2.8878  0.003879 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ilustración 61.Estadísticos y p-valor variables exógenas. Segmento 4

De los resultados anteriores se observa que las variables desempleo e IMACEC son significativas al 99% de confianza, donde la variación de un punto porcentual del desempleo hará caer en promedio 17 mil millones de pesos los saldos de este segmento, mientras que la misma variación en el IMACEC hará disminuir 2 mil 500 millones de pesos.

3.6.4. Proyección de los modelos

Siguiendo la misma estructura y metodología que en la sección 3.3.4. Proyección de los modelos, se muestran los gráficos proyectados a continuación:

- ARIMA(1,1,1) sin estacionalidad ni variables exógenas (con drift)

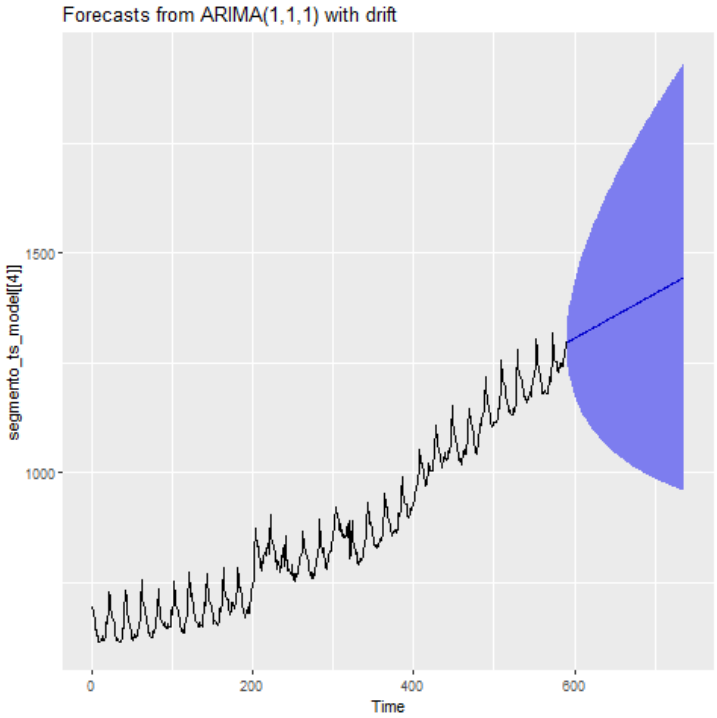

Ilustración 62. Proyección ARIMA no estacional. Segmento 4

- ARIMA(1,1,1) con δ_5, δ_{25}, δ_{30} y δ_{31}, variables exógenas: desempleo e IMACEC (con drift)

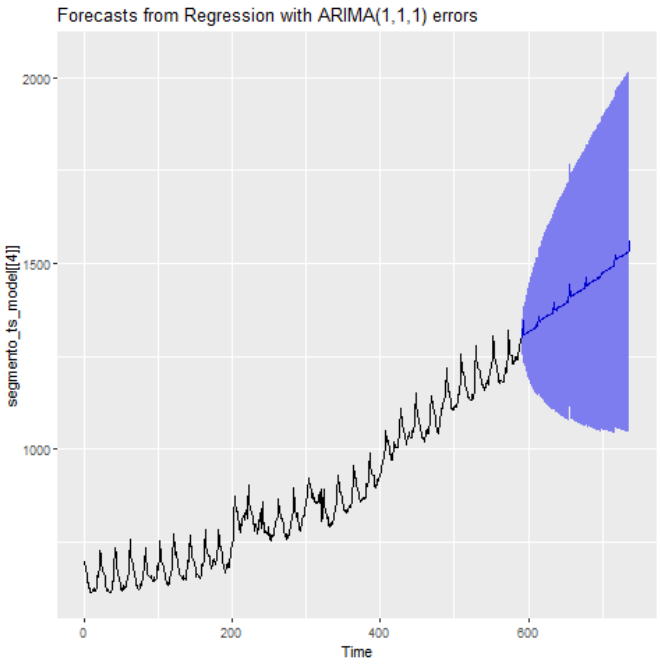

Ilustración 63. Proyección ARIMA con variable exógena. Segmento 4

Luego de esta observación gráfica de las proyecciones, se procede a comparar estos valores predichos versus la data de validación. Los resultados se muestran a continuación:

ARIMA	RMSE
(1,1,1)	56,57
(1,1,1) con δ_5, δ_{25}, δ_{30} y δ_{31}, + Desempleo, IMACEC	92,20

Tabla 11. Resultados RSME sobre data de validación

3.7. Resultados Segmento 5

Como ya se mencionó anteriormente, el segmento 5 corresponde a la agrupación de todos los clientes dentro de la cuenta IFRS 2100110 (Cuentas corrientes de personas naturales), que tienen sus saldos en moneda USD, sin diferenciación por remuneración de cuenta y pertenecientes a la clasificación interna de Personas naturales y PYME.

De aquí en adelante, todo el análisis y modelamiento se realiza sobre la data de modelamiento del segmento 5.

En la sección 3.1. Exploración preliminar se puede observar que el comportamiento de la serie posee un muy leve crecimiento en la primera parte del horizonte temporal y un crecimiento agresivo durante la última parte. De forma análoga a la realizado en la sección 3.3. Resultados Segmento 1, se utilizan los test KPSS y DFA para evaluar estadísticamente la presencia o no de estacionariedad. Los resultados son los siguientes:

```
###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression none


Call:
lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7446 -0.2123 -0.0545  0.1193  6.6989

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
z.lag.1     0.0041539  0.0009405   4.417 1.15e-05 ***
z.diff.lag -0.1320403  0.0368580  -3.582 0.000363 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7511 on 733 degrees of freedom
Multiple R-squared:  0.0368,    Adjusted R-squared:  0.03417
F-statistic:    14 on 2 and 733 DF,  p-value: 1.076e-06


Value of test-statistic is: 4.4168

Critical values for test statistics:
      1pct  5pct 10pct
tau1 -2.58 -1.95 -1.62
```

Ilustración 64. Test ADF Segmento 5, serie original

```
######################
# KPSS Unit Root Test #
######################

Test is of type: mu with 6 lags.

Value of test-statistic is: 8.7682

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.347 0.463  0.574 0.739
```

Ilustración 65. Test KPSS Segmento 5, serie original

Para este segmento, el estadístico del test ADF es 4.4148 y el estadístico del test KPSS es 8.7682. Comparándolos contra los valores críticos se tiene que el estadístico del test ADF es mayor al de los valores críticos, por lo cual no es posible rechazar la hipótesis nula. Respecto al

estadístico del test KPSS, este es mayor que los valores críticos, por ende, se rechaza la hipótesis nula en favor de la hipótesis alternativa. Así, se prueba que la serie no es estacionaria pues se comprueba la existencia de una raíz unitaria.

Dado el resultado anterior, se hace necesario diferencia5 la serie de saldos de este segmento para comprobar si de esta forma la serie se vuelve estacionaria. La nueva serie de saldos del segmento 5 diferenciada posee la siguiente forma:

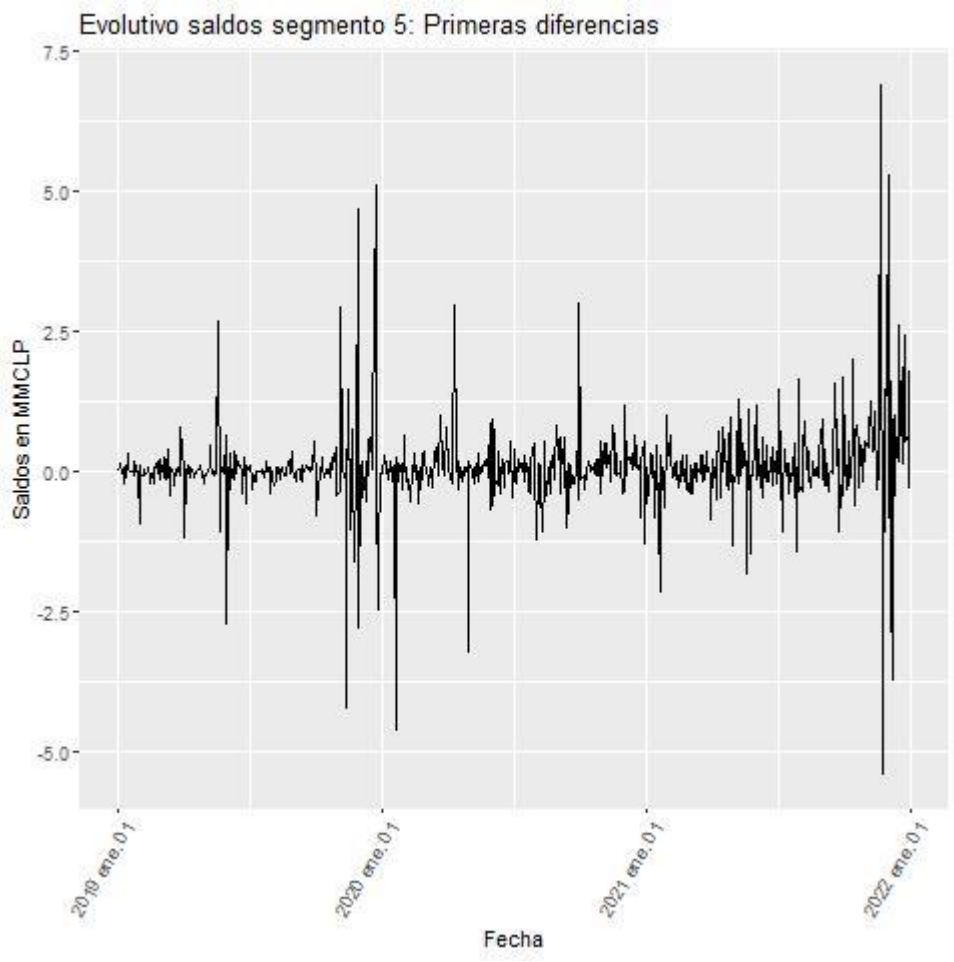

Ilustración 66. Evolutivo del segmento 5 en sus primeras diferencias

Del gráfico anterior, se observa que la serie de las diferencias está centrada en cero, con algunas variaciones más amplias y persistentes en la última parte de la data, casi dejando de estar centrada. Dicho lo anterior, la serie de diferencias pareciera cumplir con las condiciones de la estacionariedad la mayor parte del tiempo, con algunos problemas al final de la data que es necesario revisar más en detalle.

Para confirmar el comportamiento analizado gráficamente, se vuelven a realizar los test ADF y KPSS, pero ahora sobre la serie de diferencias. Los resultados son los siguientes:

```
###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression none


Call:
lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6281 -0.0976  0.0359  0.2239  6.9337

Coefficients:
           Estimate Std. Error t value Pr(>|t|)
z.lag.1    -1.14160    0.05512 -20.711   <2e-16 ***
z.diff.lag  0.03215    0.03708   0.867    0.386
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7611 on 732 degrees of freedom
Multiple R-squared:  0.5516,    Adjusted R-squared:  0.5504
F-statistic: 450.3 on 2 and 732 DF,  p-value: < 2.2e-16


Value of test-statistic is: -20.7106

Critical values for test statistics:
      1pct  5pct 10pct
tau1 -2.58 -1.95 -1.62
```

Ilustración 67.Test ADF Segmento 5, serie diferenciada

```
#######################
# KPSS Unit Root Test #
#######################

Test is of type: mu with 6 lags.

Value of test-statistic is: 1.313

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.347 0.463  0.574 0.739
```

Ilustración 68.Test KPSS Segmento 5, serie diferenciada

Con la diferenciación de la serie, se genera una modificación importante en los estadísticos KPSS y ADF. En el caso del test ADF, su valor es de -20.7106, menor que todos los valores críticos (o, dicho de otra forma, es más negativo que los valores críticos) por lo cual se rechaza la hipótesis nula, en favor de la alternativa la cual indica que la serie es estacionaria. Respecto al test KPSS, el valor del estadístico es de 1.313, mayor que los valores críticos para todos los niveles de significancia, lo cual implica que se rechaza la hipótesis nula y de este modo la serie posee una raíz unitaria. En este caso hay una disparidad entre los resultados de ambos test, provocado probablemente por el comportamiento de crecimiento acelerado en la última parte de la data. Para asegurar la estacionariedad con ambos test, se vuelve a diferenciar la serie:

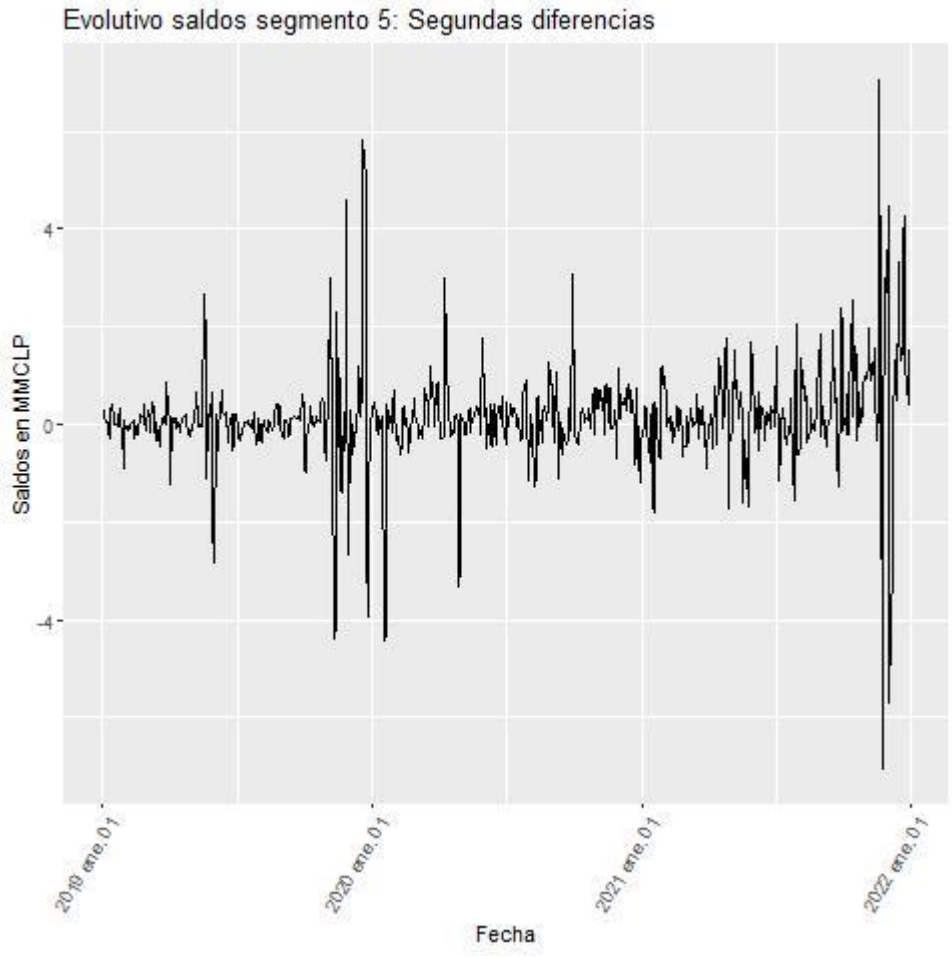

Ilustración 69. Evolutivo del segmento 5 en sus segundas diferencias

Del gráfico anterior, se observa que la serie de las segundas diferencias tiene un comportamiento muy similar a la serie de las primeras diferencias. Para confirmar si las series tienen el mismo comportamiento, se vuelven a realizar los test ADF y KPSS, pero ahora sobre la serie de segundas diferencias. Los resultados son los siguientes:

```
###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression none


Call:
lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)

Residuals:
    Min      1Q  Median      3Q     Max
-6.0308 -0.1487  0.0655  0.3200  6.9898

Coefficients:
           Estimate Std. Error t value Pr(>|t|)
z.lag.1    -0.77425    0.03702  -20.92   <2e-16 ***
z.diff.lag  0.35595    0.03462   10.28   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8639 on 731 degrees of freedom
Multiple R-squared:  0.375,     Adjusted R-squared:  0.3733
F-statistic: 219.3 on 2 and 731 DF,  p-value: < 2.2e-16


Value of test-statistic is: -20.9162

Critical values for test statistics:
      1pct  5pct 10pct
tau1 -2.58 -1.95 -1.62
```

Ilustración 70.Test ADF Segmento 5, serie diferenciada dos veces

```
######################
# KPSS Unit Root Test #
######################

Test is of type: mu with 6 lags.

Value of test-statistic is: 1.4059

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.347 0.463  0.574 0.739
```

Ilustración 71.Test KPSS Segmento 5, serie diferenciada dos veces

Con esta nueva diferenciación de la serie, los estadísticos de ADF y KPSS parecen no variar mucho respecto a los valores que tenían en la primera diferenciación. El estadístico de ADF pasa de -20.7106 a -20.9162 y el estadístico de KPSS pasa de 1.313 a 1.4059, lo cual no modifica ninguna de las conclusiones obtenidas sobre las hipótesis realizadas a la serie de las primeras diferencias. Esto lleva a evaluar una metodología alternativa para alcanzar la estacionariedad, ya que la diferenciación no permite cumplir este objetivo.

Para transformar series de tiempo que posean el comportamiento de la serie de este segmento en particular, se utiliza la denominada transformación de Box-Cox, la cual cumple la siguiente expresión dependiendo del valor de λ:

$$\begin{cases} \dfrac{x^{\lambda}-1}{\lambda} \; si \; \lambda \neq 0 \\ \log(x) \; si \; \lambda = 0 \end{cases}$$

Para encontrar este valor del parámetro λ, se utiliza la función boxcox, del paquete MASS en RStudio, el cual estima el parámetro de transformación mediante estimación de máxima verosimilitud. El resultado es el siguiente:

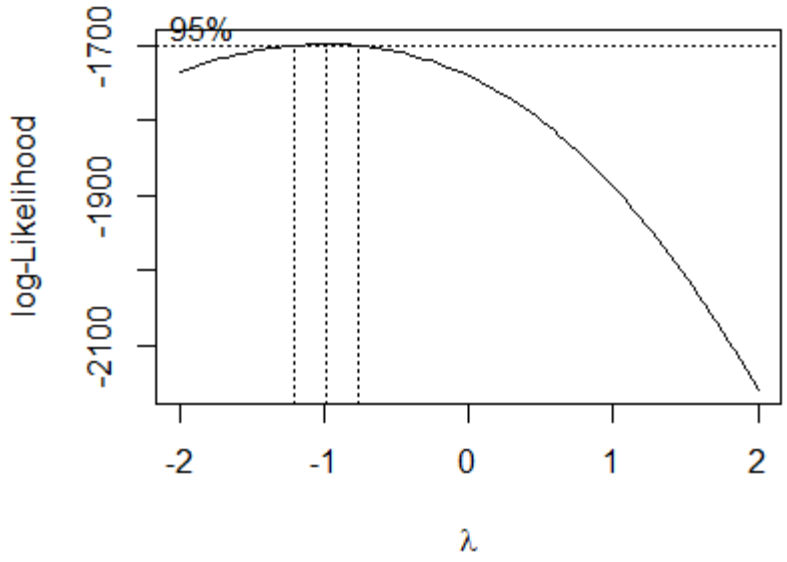

Ilustración 72. Gráfico Box-Cox y su respectivo lambda

Del gráfico anterior es importante notar que las líneas punteadas verticales indican el parámetro estimado (línea central) y el intervalo de confianza al 95% (líneas laterales), así dado que el estimador corresponde a -1, la transformación de Box-Cox que se aplica es la siguiente:

$$x_{Box-Cox} = \frac{x^{\lambda} - 1}{\lambda} = \frac{x^{-1} - 1}{-1} = 1 - \frac{1}{x}$$

Donde $x_{Box-Cox}$ indica la nueva serie de tiempo transformada mediante Box-Cox.

A continuación, se presenta la nueva forma de la serie transformada y la aplicación de los test ADF y KPSS:

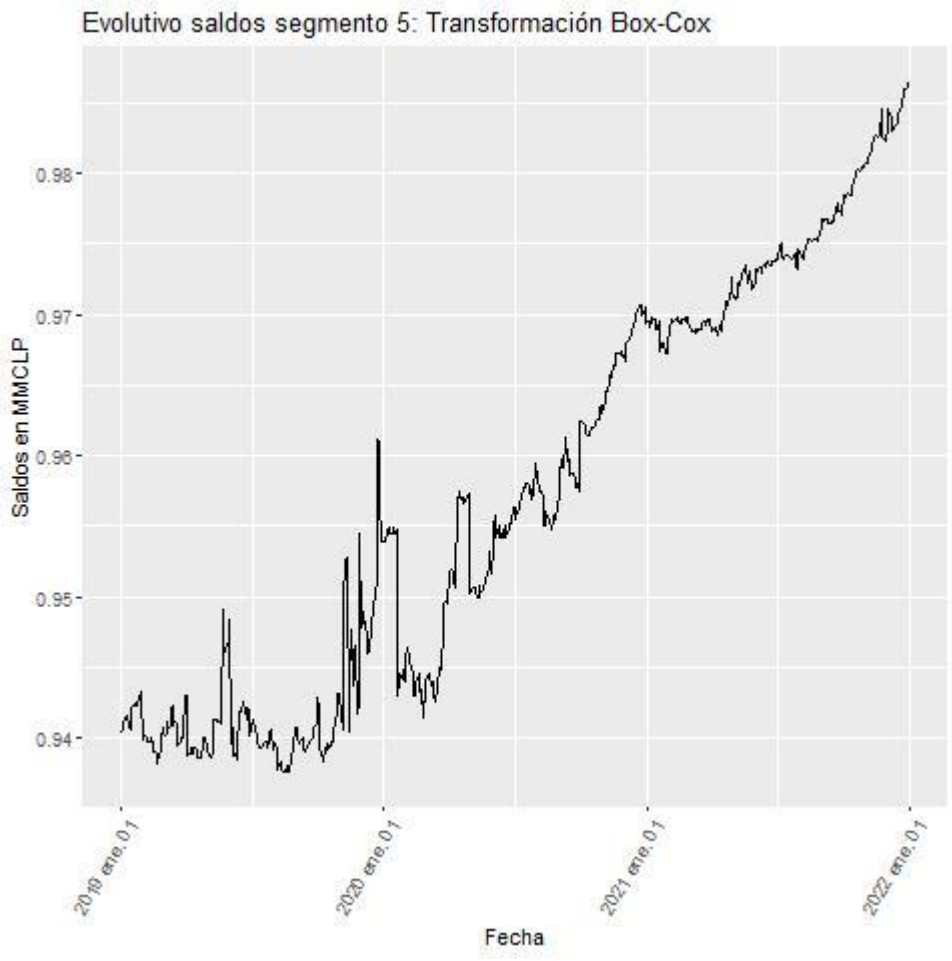

Ilustración 73. Evolutivo del segmento 5, serie transformada

```
Test regression none


Call:
lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0119924 -0.0002268 -0.0000026  0.0002302  0.0122737

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
z.lag.1     7.161e-05  5.029e-05   1.424  0.15486
z.diff.lag -1.041e-01  3.674e-02  -2.833  0.00474 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.001304 on 733 degrees of freedom
Multiple R-squared:  0.01304,   Adjusted R-squared:  0.01035
F-statistic: 4.843 on 2 and 733 DF,  p-value: 0.008135


Value of test-statistic is: 1.4241

Critical values for test statistics:
      1pct  5pct 10pct
tau1 -2.58 -1.95 -1.62
```

Ilustración 74.Test ADF Segmento 5, serie transformada

```
#######################
# KPSS Unit Root Test #
#######################

Test is of type: mu with 6 lags.

Value of test-statistic is: 10.2916

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.347 0.463  0.574 0.739
```

Ilustración 75.Test KPSS Segmento 5, serie transformada

De los resultados obtenidos de los test (estadístico de ADF 1.4241 y KPSS 10.2916), se concluye nuevamente que la serie no es estacionaria, sin embargo, observando gráficamente la nueva serie transformada, se ve claramente que la transformación modificó la componente no lineal que se observaba en la parte final del gráfico, imperando un crecimiento lineal. Dicho lo anterior, es útil evaluar una diferenciación de la serie transformada, debido a que el comportamiento es similar al comportamiento original de las series de saldos de los segmentos anteriormente modelados. Así, la nueva gráfica de la serie diferenciada es:

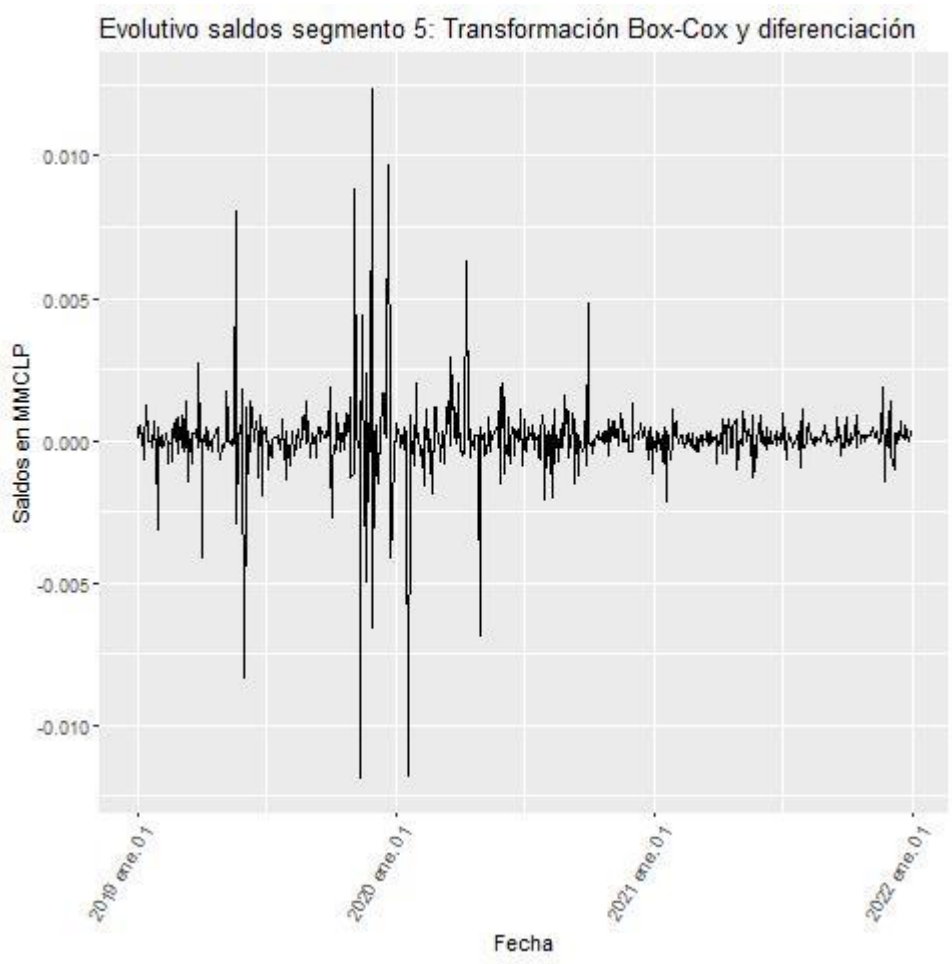

Ilustración 76. Evolutivo del segmento 5, serie transformada y diferenciada

Del gráfico anterior se observa que ahora si la serie pareciera estar centrada en cero durante todo el horizonte de evaluación, quitando el efecto de tendencia que se observaba en la parte final del gráfico. Para confirmar este comportamiento, se realiza nuevamente el test ADF y el test KPSS, los resultados son los siguientes:

```
###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression none


Call:
lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0119153 -0.0001688  0.0000684  0.0003241  0.0123530

Coefficients:
           Estimate Std. Error t value Pr(>|t|)
z.lag.1    -1.19482    0.05466 -21.858   <2e-16 ***
z.diff.lag  0.08466    0.03683   2.299   0.0218 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.001302 on 732 degrees of freedom
Multiple R-squared:  0.554,     Adjusted R-squared:  0.5528
F-statistic: 454.7 on 2 and 732 DF,  p-value: < 2.2e-16


Value of test-statistic is: -21.8583

Critical values for test statistics:
      1pct  5pct 10pct
tau1 -2.58 -1.95 -1.62
```

Ilustración 77.Test ADF Segmento 5, serie transformada y diferenciada

```
#######################
# KPSS Unit Root Test #
#######################

Test is of type: mu with 6 lags.

Value of test-statistic is: 0.0663

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.347 0.463  0.574 0.739
```

Ilustración 78.Test KPSS Segmento 5, serie transformada y diferenciada

Con la diferenciación de la serie transformada, se obtienen los resultados esperados en los estadísticos KPSS y ADF, dado el análisis gráfico realizado anteriormente, es decir como ahora el valor del estadístico de ADF es -21.8583 y el valor del estadístico de KPSS es de 0.0663, se puede

concluir que la serie de diferencias de la serie transformada mediante Box-Cox es estacionaria, lo cual da el paso para aplicar los modelos ARIMA al cumplir la condición necesaria de estacionariedad.

3.7.1. ARIMA sin estacionalidad

De forma análoga a lo realizado en la sección 3.3.1. ARIMA sin estacionalidad, el siguiente paso es encontrar los parámetros *p* y *q*, pues gracias a la diferenciación el parámetro *d* es igual a 1. Para lograr lo anterior se utilizan las funciones ACF y PACF, las cuales se muestran a continuación:

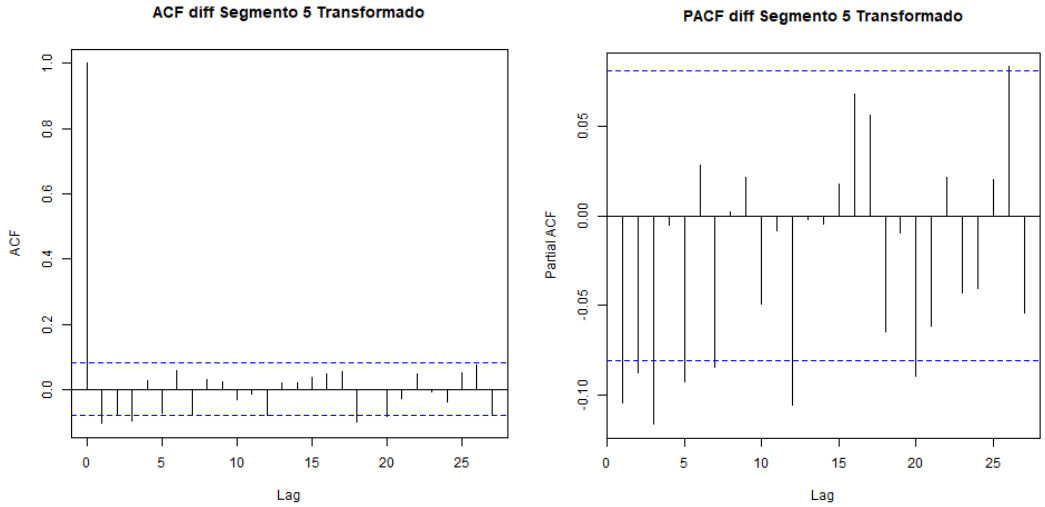

Ilustración 79. Gráficos ACF y PACF Segmento 5, transformada y primeras diferencias

De los gráficos anteriores se pueden comentar las siguientes observaciones:

- Respecto al gráfico de ACF, existen *lags* significativos en el *lag* 1 y lag 3, existiendo otros *lags* cercanos a la significancia, pero para este caso se considerarán los ya mencionado.
- Respecto al gráfico de PACF, los *lags* significativos son el *lag* 1, *lag* 2, *lag* 3 y *lag* 5. Adicionalmente se observan lags significativos múltiplos de 5.

De acuerdo con las significancias identificadas, se evalúan las posibles combinaciones para AR(1), AR(2), AR(3) y AR(5) respecto a su componente autorregresiva, y MA(1) y MA(3) respecto a su componente de media móvil. Así, los parámetros ARIMA a evaluar son los siguientes:

- ARIMA(1,1,1)
- ARIMA(1,1,3)
- ARIMA(2,1,1)
- ARIMA(2,1,3)
- ARIMA(3,1,1)
- ARIMA(3,1,3)
- ARIMA(5,1,1)
- ARIMA(5,1,3)

A continuación, se muestra una tabla en donde se muestran los criterios AIC,BIC y Error Cuadrático Medio para cada uno de los parámetros ARIMA indicados anteriormente.

ARIMA	AIC	BIC	RMSE
(1,1,1)	-6.010,53	-5.997,40	0,001442
(2,1,1)	-6.012,88	-5.995,37	0,001437
(3,1,1)	-6.018,62	-5.996,74	0,001427
(5,1,1)	-6.019,50	-5.988,86	0,001421
(1,1,3)	-6.020,26	-5.998,37	0,001424
(2,1,3)	-6.020,19	-5.993,93	0,001422
(3,1,3)	-6.019,92	-5.989,29	0,001420
(5,1,3)	-6.017,33	-5.977,94	0,001419

Tabla 12. Criterios AIC, BIC y RMSE. Segmento 5

De la tabla anterior se observa que no hay un ARIMA que sea el mínimo en los 3 criterios evaluados, pero si el ARIMA(1,1,3) tiene menor AIC y BIC simultáneamente, entonces como estos criterios son los principales a considerar, el ARIMA(1,1,3) es el escogido. Respecto al RMSE, todos los valores son muy similares, por lo cual no es un criterio relevante para discernir entre los distintos modelos.

A modo comparativo y de incluir una segunda opinión dentro de la selección de parámetros, se ejecuta la función auto.arima. El resultado es el siguiente:

```
Series: segmento_ts_model[[5]]
ARIMA(1,1,1) with drift

Coefficients:
         ar1      ma1   drift
      0.6995  -0.8138  0.0324
s.e.  0.1493   0.1234  0.0157

sigma^2 estimated as 0.378:  log likelihood=-547.75
AIC=1103.51   AICc=1103.58   BIC=1121.02

Training set error measures:
                      ME      RMSE       MAE          MPE     MAPE     MASE         ACF1
Training set 0.0001252237 0.6127032 0.298816 -0.09648976 1.350676 1.034093 -0.00307326
```

Ilustración 80. Resultados auto.arima, Segmento 5

Se observa que la mayoría de los parámetros y criterios difieren de los encontrados en este trabajo. Una posible razón es que el código aplicado no realizara internamente la transformación de la variable para asegurar la estacionariedad, con lo cual se hace difícil comparar ambos resultados y por ende no se considerará en esta ocasión. Así, el modelo seleccionado no se ve alterado y seguirá siendo el ARIMA(1,1,3)

Por último, se realiza el análisis sobre los residuos, de tal forma de corroborar si cumplen con las características que lo hacen ser ruido blanco. Este análisis se muestra a continuación:

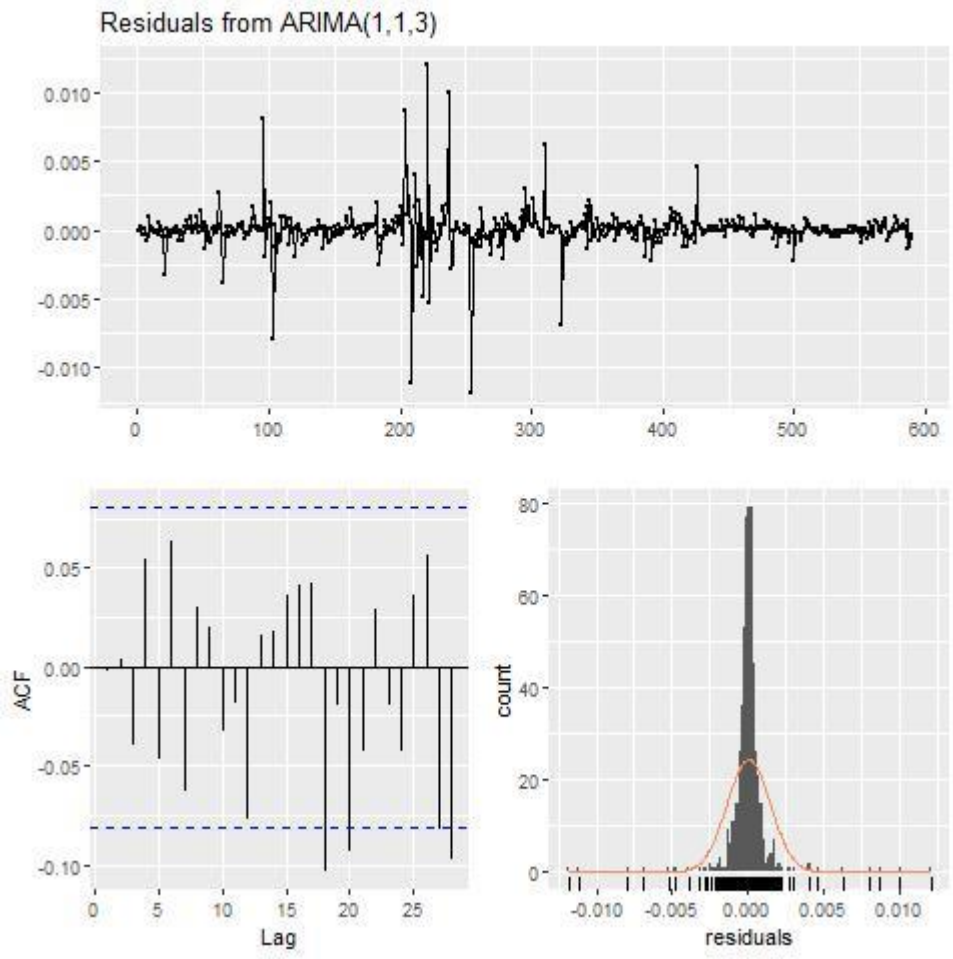

Ilustración 81. Análisis de residuos. Segmento 5

Del gráfico anterior se puede observar lo siguiente:

- Los residuos están centrados en cero, con una varianza relativamente constante a lo largo del tiempo, excepto en un par de fechas muy puntuales.
- Los residuos siguen una distribución muy similar a la distribución normal, donde la gran mayoría de los datos está muy concentrada en cero.
- Del gráfico ACF se observa que no existen *lags* significativos hasta recién cercano al *lag* 20, en el cual hay un par de *lags* significativos, los cuales no deberían tener una incidencia en el comportamiento de los residuos y en la conclusión de este.

Dicho lo anterior, se realiza el test Ljung-Box para confirmar que los *lags* significativos encontrados no alteran el comportamiento de los residuos. Los resultados son los siguientes:

```
        Ljung-Box test

data:  Residuals from ARIMA(1,1,3)
Q* = 10.212, df = 6, p-value = 0.116

Model df: 4.   Total lags used: 10
```

Ilustración 82. Test Ljung-Box. Segmento 5

Se observa que el p-valor es de 0.116, mayor que el nivel de significancia del 0.05, lo cual indica que la hipótesis nula no puede ser rechazada, implicando que los residuos se distribuyen independientemente, y por ende son ruido blanco.

3.7.2. ARIMA estacional o SARIMA

De la misma forma que en la sección 3.3.2. ARIMA estacional o SARIMA, se evalúa la significancia de agregar al modelo las variables dummies asociadas a los días calendario previamente definidos. Los resultados son los siguientes:

```
Call:
arima(x = segmento_ts_model[[5]], order = c(1, 1, 3), xreg = base_model[, c(8,
    12, 22, 32, 37, 38)], method = "ML")

Coefficients:
          ar1     ma1      ma2      ma3  Dummy_dia_1  Dummy_dia_5  Dummy_dia_15  Dummy_dia_25
      -0.7340  0.6310  -0.1576  -0.1366       0.0304      -0.0416       -0.0129       -0.0250
s.e.   0.1604  0.1634   0.0523   0.0431       0.1224       0.0994        0.1015        0.1013
      Dummy_dia_30  Dummy_dia_31
           -0.0097        0.2397
s.e.        0.1110        0.1734

sigma^2 estimated as 0.3743:  log likelihood = -546.38,  aic = 1114.77

Training set error measures:
                     ME     RMSE       MAE       MPE    MAPE     MASE         ACF1
Training set 0.04139629 0.611278 0.3012617 0.1150996 1.35416 1.042557 -0.003423116
```

```
z test of coefficients:

              Estimate Std. Error z value  Pr(>|z|)
ar1          -0.7340093  0.1604456 -4.5748 4.766e-06 ***
ma1           0.6310330  0.1634209  3.8614 0.0001127 ***
ma2          -0.1576235  0.0523271 -3.0123 0.0025930 **
ma3          -0.1366096  0.0431272 -3.1676 0.0015371 **
Dummy_dia_1   0.0303736  0.1224145  0.2481 0.8040409
Dummy_dia_5  -0.0415538  0.0994067 -0.4180 0.6759340
Dummy_dia_15 -0.0128569  0.1014845 -0.1267 0.8991868
Dummy_dia_25 -0.0250182  0.1012560 -0.2471 0.8048475
Dummy_dia_30 -0.0096502  0.1110030 -0.0869 0.9307221
Dummy_dia_31  0.2397455  0.1733596  1.3829 0.1666839
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ilustración 83. Estadísticos y p-valor variables dummy estacionales. Segmento 5

De los resultados anteriores se observa que ninguna variable dummy es significativa para el modelo, por lo cual no es relevante incluirlas en el modelamiento de este segmento.

3.7.3. ARIMA con variables exógenas o ARIMAX

De la misma forma que en la sección 3.3.3. ARIMA con variables exógenas o ARIMAX, se procede a incorporar las variables macroeconómicas ya mencionadas, de tal de forma de evaluar la significancia de agregarlas al modelo ya seleccionado. Los resultados son los siguientes:

```
Call:
arima(x = diff(segmento_ts_model[[5]]), order = c(1, 1, 3), xreg = cbind(depos_USD,
    desempleo, IPC, IMACEC), method = "ML")

Coefficients:
         ar1      ma1     ma2      ma3  depos_USD  desempleo     IPC   IMACEC
      0.7017  -1.8113  0.8126  -0.0006    -0.1675    -0.1498  0.3392  -0.0318
s.e.  0.5704   0.5982  0.7201   0.1319     0.1341     0.2000  0.3097   0.0273

sigma^2 estimated as 0.3746:  log likelihood = -548.71,  aic = 1115.41

Training set error measures:
                     ME      RMSE       MAE MPE MAPE      MASE         ACF1
Training set 0.02795995 0.6115529 0.2982403 NaN  Inf 0.6370666 -0.006227243
```

```
z test of coefficients:

             Estimate  Std. Error z value Pr(>|z|)
ar1         0.70170485  0.57038076  1.2302 0.218607
ma1        -1.81134470  0.59820813 -3.0280 0.002462 **
ma2         0.81256408  0.72012659  1.1284 0.259167
ma3        -0.00063383  0.13190389 -0.0048 0.996166
depos_USD  -0.16749542  0.13408847 -1.2491 0.211613
desempleo  -0.14984552  0.20001777 -0.7492 0.453760
IPC         0.33915508  0.30969017  1.0951 0.273454
IMACEC     -0.03181955  0.02728507 -1.1662 0.243538
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ilustración 84.Estadísticos y p-valor variables exógenas. Segmento 5

De los resultados anteriores se observa que ninguna variable exógena es significativa. Por lo cual no hay suficiente evidencia como para incorporarlas al modelo.

3.7.4. Proyección de los modelos

Siguiendo la misma estructura y metodología que en la sección 3.3.4. Proyección de los modelos, se muestran los gráficos proyectados a continuación. Cabe mencionar que para este segmento no hay variables estacionales ni exógenas significativas, por lo cual solo se considera el modelo ARIMA determinado en primera instancia:

- ARIMA(1,1,3) sin estacionalidad ni variables exógenas (con drift)

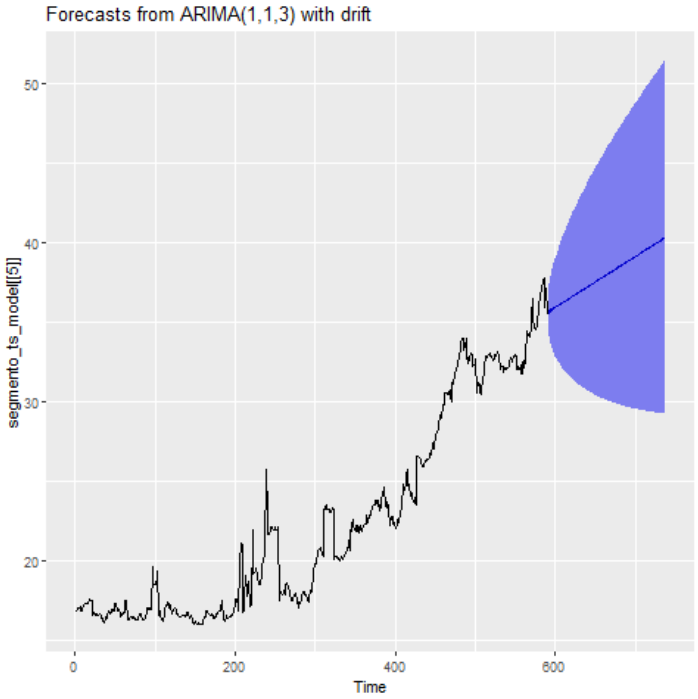

Ilustración 85. Proyección ARIMA no estacional. Segmento 5

Luego de esta observación gráfica de las proyecciones, se procede a comparar estos valores predichos versus la data de validación. Los resultados se muestran a continuación:

ARIMA	RMSE
(1,1,3)	12,26

Tabla 13. Resultados RSME sobre data de validación

3.8. Resultados Segmento 6

Como ya se mencionó anteriormente, el segmento 6 corresponde a la agrupación de todos los clientes dentro de la cuenta IFRS 2100204 (Cuentas de depósitos a la vista), que tienen sus saldos en moneda CLP, sin diferenciación por remuneración de cuenta y pertenecientes a la clasificación interna de Personas naturales, PYME e Instituciones no financieras.

De aquí en adelante, todo el análisis y modelamiento se realiza sobre la data de modelamiento del segmento 6.

En la sección 3.1. Exploración preliminar se puede observar que el comportamiento de la serie posee un crecimiento constante durante la mayor parte del horizonte de trabajo, con un leve decaimiento en la parte final de este. De forma análoga a la realizado en la sección 3.3. Resultados Segmento 1, se utilizan los test KPSS y DFA para evaluar estadísticamente la presencia o no de estacionariedad. Los resultados son los siguientes:

```
###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression none


Call:
lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)

Residuals:
     Min       1Q   Median       3Q      Max
-12.9293  -2.3038  -0.7181   1.7045  22.7769

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
z.lag.1    0.0004888  0.0003496   1.398   0.1625
z.diff.lag 0.0675836  0.0371900   1.817   0.0696 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.746 on 733 degrees of freedom
Multiple R-squared:  0.007553,  Adjusted R-squared:  0.004845
F-statistic: 2.789 on 2 and 733 DF,  p-value: 0.06212


Value of test-statistic is: 1.3981

Critical values for test statistics:
      1pct  5pct 10pct
tau1 -2.58 -1.95 -1.62
```

Ilustración 86. Test ADF Segmento 6, serie original

```
######################
# KPSS Unit Root Test #
######################

Test is of type: mu with 6 lags.

Value of test-statistic is: 10.3803

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.347 0.463  0.574 0.739
```

Ilustración 87. Test KPSS Segmento 6, serie original

Para este segmento, el estadístico del test ADF es 1.3981 y el estadístico del test KPSS es 10.3803. Comparándolos contra los valores críticos se tiene que el estadístico del test ADF es mayor al de los valores críticos, por lo cual no es posible rechazar la hipótesis nula. Respecto al estadístico del test KPSS, este es mayor que los valores críticos, por ende, se rechaza la hipótesis nula en favor de la hipótesis alternativa. Así, se prueba que la serie no es estacionaria pues se comprueba la existencia de una raíz unitaria.

Dado el resultado anterior, se hace necesario diferenciar la serie de saldos de este segmento para comprobar si de esta forma la serie se vuelve estacionaria. La nueva serie de saldos del segmento 6 diferenciada posee la siguiente forma:

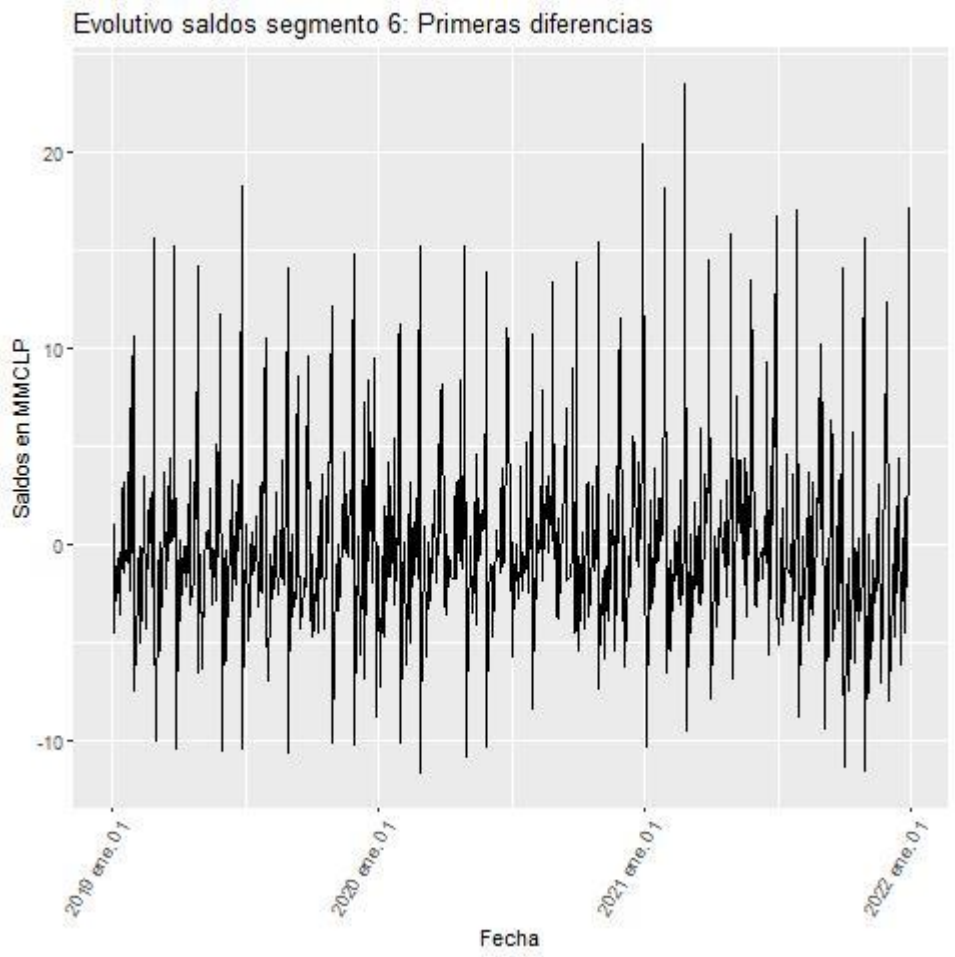

Ilustración 88. Evolutivo del segmento 6 en sus primeras diferencias

Del gráfico anterior, se observa que la serie de las diferencias está centrada en cero, con variaciones que poseen alguna forma estacional, pero lo cual no influye en el comportamiento general. Dicho lo anterior, la serie de diferencias pareciera cumplir con las condiciones de la estacionariedad.

Para confirmar el comportamiento analizado gráficamente, se vuelven a realizar los test ADF y KPSS, pero ahora sobre la serie de diferencias. Los resultados son los siguientes:

```
###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression none


Call:
lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)

Residuals:
     Min       1Q   Median       3Q      Max
-12.2328  -2.0255  -0.4917   2.0139  23.0037

Coefficients:
           Estimate Std. Error t value Pr(>|t|)
z.lag.1    -1.01583    0.05076 -20.011   <2e-16 ***
z.diff.lag  0.09266    0.03715   2.495   0.0128 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.735 on 732 degrees of freedom
Multiple R-squared:  0.4647,    Adjusted R-squared:  0.4633
F-statistic: 317.8 on 2 and 732 DF,  p-value: < 2.2e-16


Value of test-statistic is: -20.011

Critical values for test statistics:
      1pct  5pct 10pct
tau1 -2.58 -1.95 -1.62
```

Ilustración 89.Test ADF Segmento 6, serie diferenciada

```
#######################
# KPSS Unit Root Test #
#######################

Test is of type: mu with 6 lags.

Value of test-statistic is: 0.0752

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.347 0.463  0.574 0.739
```

Ilustración 90.Test KPSS Segmento 6, serie diferenciada

Con la diferenciación de la serie, se genera una modificación importante en los estadísticos KPSS y ADF. En el caso del test ADF, su valor es de -20.011, menor que todos los valores críticos (o, dicho de otra forma, es más negativo que los valores críticos) por lo cual se rechaza la hipótesis nula, en favor de la alternativa la cual indica que la serie es estacionaria. Respecto al test KPSS, el valor del estadístico es de 0.0752, menor que los valores críticos para todos los niveles de significancia, lo cual implica que no es posible rechazar la hipótesis nula y de este modo la serie es estacionaria o posee tendencia estacionaria. Complementando ambos test se puede concluir que la serie de diferencias de la serie original de saldos del segmento 6 es estacionaria, lo cual da el paso para aplicar los modelos ARIMA al cumplir la condición necesaria de estacionariedad.

3.8.1. ARIMA sin estacionalidad

De forma análoga a lo realizado en la sección 3.3.1. ARIMA sin estacionalidad, el siguiente paso es encontrar los parámetros *p* y *q*, pues gracias a la diferenciación el parámetro *d* es igual a 1. Para lograr lo anterior se utilizan las funciones ACF y PACF, las cuales se muestran a continuación:

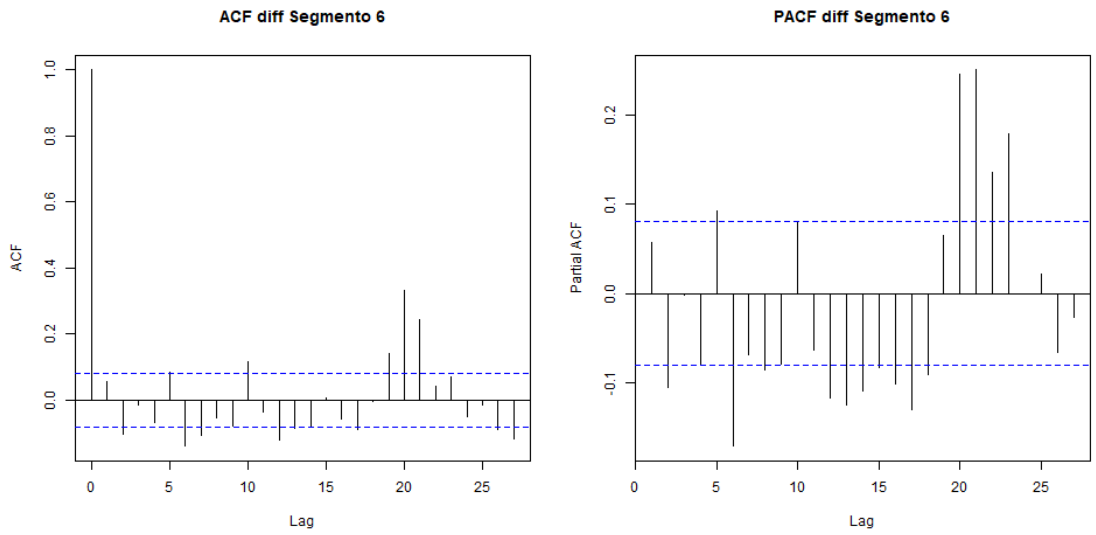

Ilustración 91. Gráficos ACF y PACF Segmento 6, primeras diferencias

- Ambos gráficos comparten un comportamiento similar, en donde bajo el *lag* 5 solo el *lag* 2 y el *lag* 5 están al borde de la significancia, mientras que sobre el *lag* 5 hay muchos *lags* significativos y con algún comportamiento estacional, ya que estos *lags* significativos se concentran en los múltiplos de 10.

En esta sección se considerarán solo los *lags* 2 y 5 tanto para la función ACF como para la PACF, dejando el potencial comportamiento estacional para secciones más adelantes de este trabajo. Así, de acuerdo con las significancias escogidas, se evalúan las posibles combinaciones para AR(2) y AR(5) respecto a su componente autorregresiva, y MA(2) y MA(5) respecto a su componente de media móvil. Así, los parámetros ARIMA a evaluar son los siguientes:

- ARIMA(2,1,2)
- ARIMA(2,1,5)
- ARIMA(5,1,2)
- ARIMA(5,1,5)

A continuación, se muestra una tabla en donde se muestran los criterios AIC,BIC y Error Cuadrático Medio para cada uno de los parámetros ARIMA indicados anteriormente.

ARIMA	AIC	BIC	RMSE
(2,1,2)	3.480	3.502	4,60

(5,1,2)	3.477	3.512	4,56
(2,1,5)	3.469	3.504	4,53
(5,1,5)	3.462	3.510	4,48

Tabla 14. Criterios AIC, BIC y RMSE. Segmento 4

De la tabla anterior se observa que no hay un ARIMA que sea el mínimo en los 3 criterios evaluados. Así, los posibles candidatos son:

- Menor AIC: ARIMA(5,1,5)
- Menor BIC: ARIMA(2,1,2)
- Menor RMSE: ARIMA(5,1,5)

Considerando al BIC como el indicador principal, el modelo a escoger sería el ARIMA(2,1,2), aunque también una opción razonable es elegir el ARIMA(5,1,5), sin embargo, en este caso se considerará el criterio BIC como el dominante, debido a la mayor penalización que realiza sobre de aumento de parámetros, buscando trabajar con la menor cantidad de parámetros posibles para eliminar el sobre ajuste del modelo. Así que por superioridad del criterio BIC sobre el AIC se escoge el modelo ARIMA (2,1,2).

A modo comparativo y de incluir una segunda opinión dentro de la selección de parámetros, se ejecuta la función auto.arima. El resultado es el siguiente:

```
Series: segmento_ts_model[[6]]
ARIMA(2,1,2) with drift

Coefficients:
         ar1     ar2      ma1      ma2   drift
      0.1950  0.5272  -0.1571  -0.7209  0.3696
s.e.  0.1041  0.0990   0.0855   0.0825  0.0833

sigma^2 estimated as 20.73:  log likelihood=-1726.21
AIC=3464.41   AICc=3464.55   BIC=3490.68

Training set error measures:
                     ME     RMSE      MAE          MPE      MAPE      MASE        ACF1
Training set -0.0167037 4.529962 3.164059 -0.02244156 0.6929369 1.002442 0.01147511
```

Ilustración 92. Resultados auto.arima, Segmento 6

En este caso se tiene que todos los parámetros del modelo coindicen. Adicionalmente, los criterios AIC y BIC son menores que los encontrados mediante la metodología propuesta, sin embargo, siguen siendo muy similares, por lo cual se sustenta aún más la elección del modelo ARIMA(2,1,2).

Por último, se realiza el análisis sobre los residuos, de tal forma de corroborar si cumplen con las características que lo hacen ser ruido blanco. Este análisis se muestra a continuación:

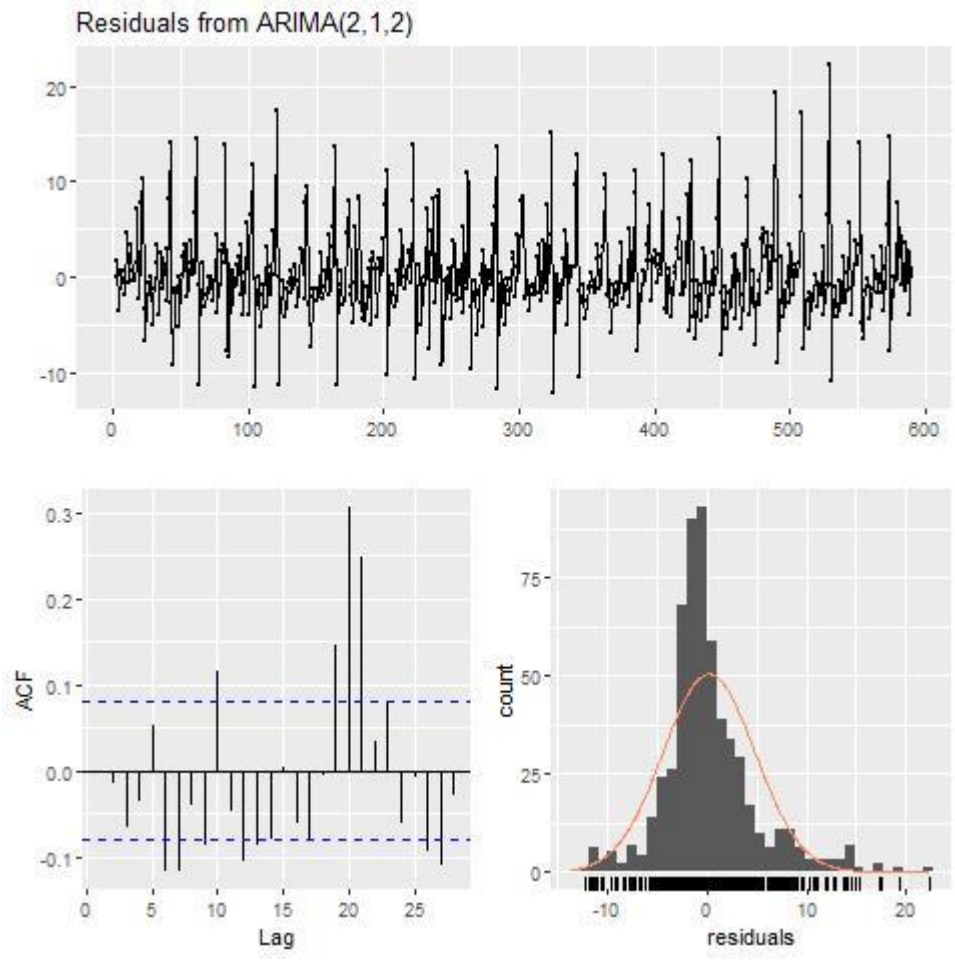

Ilustración 93. Análisis de residuos. Segmento 6

Del gráfico anterior se puede observar lo siguiente:

- Los residuos están centrados en cero, con un comportamiento estacional de los residuos, muy similar al comportamiento de la serie original de saldos.
- Los residuos siguen una distribución muy similar a la distribución normal, donde la gran mayoría de los datos está muy concentrada en cero, pero con una cola más relevante que la otra.
- Del gráfico ACF se observa el comportamiento observado en el análisis previamente realizado, *lags* en forma de onda y en su mayoría significativos, implicando que hay información en los residuos que debiese estar en el modelo, por lo cual nos indica que la serie de residuos no es un ruido blanco.

Dicho lo anterior, se realiza el test Ljung-Box, para confirmar que los residuos no se comportan como ruido blanco. Los resultados son los siguientes:

```
        Ljung-Box test

data:  Residuals from ARIMA(2,1,2)
Q* = 34.921, df = 6, p-value = 4.464e-06

Model df: 4.   Total lags used: 10
```

Ilustración 94. Test Ljung-Box. Segmento 6

Se observa que el p-valor recién tiene su decimal distinto de cero en la sexta posición, un valor casi 0 y muy bajo comparado con el nivel de significancia de 0.05. Esto nos indica que se rechaza la hipótesis nula, aceptando la hipótesis alternativa de que los residuos no son independientes y por ende no son ruido blanco. Este resultado está en línea con lo obtenido a lo largo de todo el estudio de este segmento, por lo cual se hará necesario evaluar nuevos parámetros o variables a incorporar al modelo, de tal forma de rescatar toda la información necesaria y que los residuos solo sean ruido blanco.

3.8.2. ARIMA estacional o SARIMA

De la misma forma que en la sección 3.3.2. ARIMA estacional o SARIMA, se evalúa la significancia de agregar al modelo las variables dummies asociadas a los días calendario previamente definidos. Los resultados son los siguientes:

```
Call:
arima(x = segmento_ts_model[[6]], order = c(2, 1, 2), xreg = base_model[, c(8,
    12, 22, 32, 37, 38)], method = "ML")

Coefficients:
         ar1     ar2      ma1      ma2  Dummy_dia_1  Dummy_dia_5  Dummy_dia_15  Dummy_dia_25
      0.0951  0.6398  -0.0467  -0.7436       2.1127      -0.1698        1.5049        0.9671
s.e.  0.1101  0.1000   0.0923   0.0834       0.7733       0.5954        0.6072        0.6055
      Dummy_dia_30  Dummy_dia_31
            7.7001       11.7950
s.e.        0.6888        1.1036

sigma^2 estimated as 15.83:  log likelihood = -1649.24,  aic = 3320.47

Training set error measures:
                    ME     RMSE      MAE        MPE      MAPE      MASE         ACF1
Training set 0.4624528 3.975764 2.691363 0.08984749 0.5851214 0.8526822 -0.006947796
```

```
z test of coefficients:

              Estimate Std. Error z value  Pr(>|z|)
ar1           0.095137   0.110128  0.8639  0.387654
ar2           0.639784   0.100047  6.3948 1.607e-10 ***
ma1          -0.046673   0.092315 -0.5056  0.613152
ma2          -0.743624   0.083424 -8.9138 < 2.2e-16 ***
Dummy_dia_1   2.112719   0.773292  2.7321  0.006293 **
Dummy_dia_5  -0.169818   0.595441 -0.2852  0.775494
Dummy_dia_15  1.504912   0.607205  2.4784  0.013196 *
Dummy_dia_25  0.967085   0.605507  1.5971  0.110232
Dummy_dia_30  7.700140   0.688833 11.1785 < 2.2e-16 ***
Dummy_dia_31 11.794996   1.103574 10.6880 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ilustración 95. Estadísticos y p-valor variables dummy estacionales. Segmento 6

De los resultados anteriores, se observa que las dummies asociadas al día 30 y 31 son significativas al 99,9%, que las dummy asociada al día 1 es significativa al 99% y la dummy asociada al día 15 es significativa al 95%. Respecto al coeficiente de estas variables, todos los coeficientes de las variables son positivos. Considerando que este segmento está compuesto por cuentas de depósitos a la vista, el resultado tiene lógica financiera debido a que en este segmento se encuentran personas naturales que no califican en la obtención de una cuenta corriente y reciben pagos (asociados a sueldos o devoluciones) desde instituciones bancarias en estas cuentas, quedando disponibles en las cuentas los últimos días del mes. Lo anterior también se ve reflejado en los coeficientes de las variables, en donde los coeficientes son más altos para las dummies asociadas al día 30 y 31 (7.7 y 11.7 respectivamente), mientras que considerablemente más pequeños los días 1 y 15 (2.1 y 1.5 respectivamente).

3.8.3. ARIMA con variables exógenas o ARIMAX

De la misma forma que en la sección 3.3.3. ARIMA con variables exógenas o ARIMAX, se procede a incorporar las variables macroeconómicas ya mencionadas, de tal de forma de evaluar la significancia de agregarlas al modelo ya seleccionado. Los resultados son los siguientes:

```
Call:
arima(x = diff(segmento_ts_model[[6]]), order = c(2, 1, 2), xreg = cbind(depos_CLP,
    desempleo, IPC, IMACEC), method = "ML")

Coefficients:
          ar1      ar2      ma1      ma2  depos_CLP  desempleo     IPC   IMACEC
      -0.6590  -0.0331  -0.2232  -0.7768     0.4461    -4.7161  4.6595  -0.5962
s.e.   0.0882   0.0467   0.0791   0.0788     2.0619     1.4476  2.2365   0.2006

sigma^2 estimated as 20.5:  log likelihood = -1725.49,  aic = 3468.98

Training set error measures:
                    ME     RMSE      MAE      MPE    MAPE      MASE          ACF1
Training set 0.2298225 4.523549 3.146547 108.6218 189.039 0.7239215 -0.003886177
```

```
z test of coefficients:

            Estimate Std. Error z value  Pr(>|z|)
ar1       -0.658953   0.088204 -7.4708 7.969e-14 ***
ar2       -0.033118   0.046668 -0.7096  0.477928
ma1       -0.223227   0.079113 -2.8216  0.004778 **
ma2       -0.776770   0.078834 -9.8532 < 2.2e-16 ***
depos_CLP  0.446121   2.061896  0.2164  0.828704
desempleo -4.716089   1.447619 -3.2578  0.001123 **
IPC        4.659462   2.236522  2.0834  0.037219 *
IMACEC    -0.596220   0.200641 -2.9716  0.002963 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ilustración 96.Estadísticos y p-valor variables exógenas. Segmento 4

De los resultados anteriores se observa que las variables desempleo e IMACEC son significativas al 99% de confianza y la variable IPC es significativa al 95%.

3.8.4. Proyección de los modelos

Siguiendo la misma estructura y metodología que en la sección 3.3.4. Proyección de los modelos, se muestran los gráficos proyectados a continuación:

- ARIMA(1,1,3) sin estacionalidad ni variables exógenas (con drift)

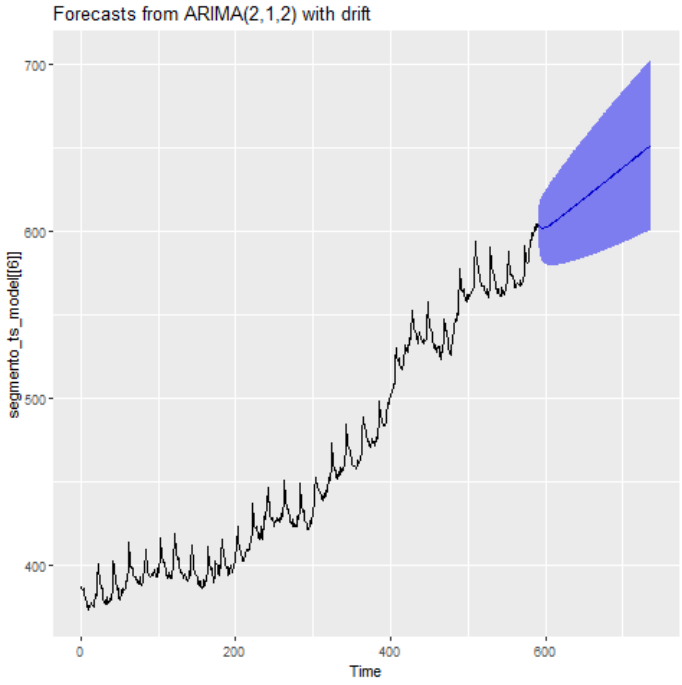

Ilustración 97. Proyección ARIMA no estacional. Segmento 6

- ARIMA(1,1,3) con δ_1, δ_{15}, δ_{30} y δ_{31} y variables exógenas: desempleo, IMACEC e IPC (con drift)

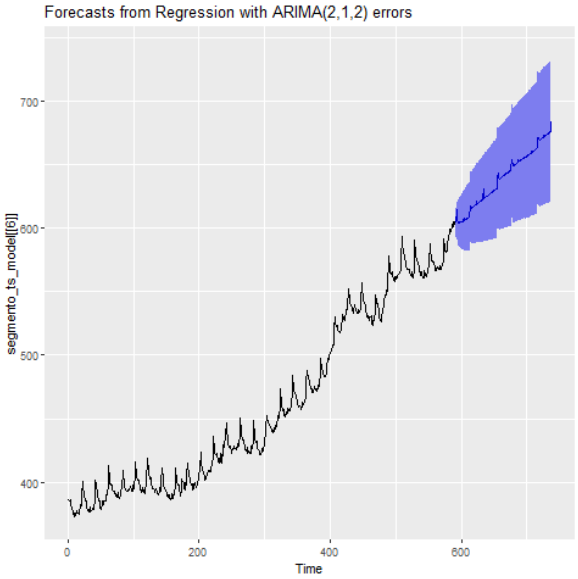

Ilustración 98. Proyección ARIMA con variable estacional y exógena. Segmento 6

Luego de esta observación gráfica de las proyecciones, se procede a comparar estos valores predichos versus la data de validación. Los resultados se muestran a continuación:

ARIMA	RMSE
(1,1,3)	32,10
(1,1,1) con δ_1, δ_{15}, δ_{30} y δ_{31}, + Desempleo, IMACEC, IPC	42,26

Tabla 15. Resultados RSME sobre data de validación

3.9. Consolidación de resultados

Con la finalidad de mejorar el entendimiento y facilitar la comparación de los resultados de todo el trabajo desarrollado en las secciones anteriores, se consolidan en esta sección.

A continuación, se muestran los ARIMAS no estacionales seleccionados para cada segmento, con el valor de sus respectivos criterios de elección:

Segmento	ARIMA seleccionado	AIC	BIC	RMSE
1	(1,1,5)	5.696	5.727	30,10
2	(6,1,1)	5.714	5.749	30,55
3	(5,1,1)	5.102	5.133	18,18
4	(1,1,1)	5.175	5.188	19,50
5	(1,1,3)	-6.020	-5.998	0,0014
6	(2,1,2)	3.480	3.502	4,60

Tabla 16. ARIMAS no estacionales, todos los segmentos

Respecto al análisis realizado incluyendo las dummies estacionales a los modelos anteriormente seleccionados, el resultado es el siguiente:

Segmento	Dummies significativas
1	δ_{25} al 99,9%
2	Ninguna
3	Ninguna
4	δ_{30} y δ_{31} al 99,9%; δ_5 y δ_{25} al 90%

5	Ninguna
6	δ_{30}y δ_{31} al 99,9%; δ_1 al 99%; δ_{15} al 95%

Tabla 17. Dummies significativas, todos los segmentos

Respecto al análisis de variables exógenas, a continuación, se muestran cuales fueron significativas y el valor de su estimador:

Segmento	Variables significativas	Valor estimador
1	Ninguna	No aplica
2	IPC al 90%	IPC: 26,49
3	IPC al 99%; desempleo e IMACEC al 95%; depósitos en USD al 90%	IPC: -24,88 Desempleo: -11,77 IMACEC: -1,57 Deposito USD: 6,88
4	Desempleo e IMACEC al 99%	Desempleo: -14,44 IMACEC: -2,50
5	Ninguna	No aplica
6	Desempleo e IMACEC al 99%; IPC al 95%	Desempleo: -4,71 IMACEC: -0,59 IPC: 4,65

Tabla 18. Variables exógenas significativas y estimadores, todos los segmentos

Finalmente, se muestra a continuación la validación de la predicción de los distintos modelos desarrollados durante todo este trabajo, versus la data de validación de cada segmento correspondiente, de tal forma de cuantificar la capacidad predictiva de los modelos. La tabla que agrupa estos resultados se muestra a continuación:

Segmento	ARIMA	RMSE
1	(1,1,5)	50,85
1	(1,1,5) con δ_{25}	50,47
2	(6,1,1)	87,79
2	(6,1,1) + IPC	100,50
3	(5,1,1)	46,54
3	(5,1,1) + IPC, Desempleo, IMACEC, depos_USD	38,65
4	(1,1,1)	56,57
4	(1,1,1) con δ_5, δ_{25}, δ_{30} y δ_{31}, + Desempleo, IMACEC	92,20
5	(1,1,3)	12,26
6	(1,1,3)	32,10

6	(1,1,1) con δ_1, δ_{15}, δ_{30} y δ_{31}, + Desempleo, IMACEC, IPC	42,26

Tabla 19. Resultado final sobre data de validación, todos los segmentos

Discusión y conclusiones

Una vez finalizado el análisis y ya con los resultados finales obtenidos para todos los segmentos evaluados, se procede a comentar y analizar este trabajo, dando énfasis a aquellos resultado que pudiesen no cumplir, a priori, con lo que se esperaba al inicio de esta memoria.

Uno de los resultados inesperados que rápidamente saltó a la vista en este trabajo fue la poca relevancia de las series de los depósitos a plazo en CLP y USD. Intuitivamente, la lógica financiera indica que, si aumentan consistentemente las tasas de estos depósitos, entonces es más atractivo invertir el dinero en ese instrumento que en las cuentas corrientes, las cuales no entregan intereses en su gran mayoría. Si bien ambos productos financieros no son completamente equivalentes, ya que el dinero en los depósitos a plazo no es de libre acceso del cliente hasta la fecha de vencimiento de este, de todas maneras, son productos relativamente simples de comprender y bastante comunes en el general de la población, por lo cual esto no sería una barrera. Una de las razones que puede explicar esta poca relevancia de la tasa de los depósitos, es que hasta el cierre del horizonte de evaluación de este trabajo las tasas estaban en alza, pero aún no alcanzaban un valor tan significativo, como lo es al cierre de este trabajo de memoria, por lo cual una forma de subsanar este resultado poco esperado es ampliando el horizonte de trabajo, de tal manera de incluir una mayor cantidad del crecimiento de estas series. Otra posible razón es que la transferencia de saldos desde las cuentas corrientes a los depósitos a plazo es más lenta de lo esperado, así hay que esperar más tiempo con tasas de depósitos altas para que se observe claramente una transferencia entre estos dos productos.

Un resultado esperado a priori y que se cumplió en este trabajo es la diferencia de impactos que tienen las variables estacionales y exógenas sobre los distintos segmentos. Si bien los productos son relativamente similares, ya que todos se enmarcan en el gran concepto de cuentas corrientes, las distintas segmentaciones capturan diferentes comportamientos y, a su vez, estos segmentos son impactados de distinta manera por las variables que se evaluaron en los modelos. A modo de ejemplo, se esperaba que el segmento 4 asociado a personas naturales y PYME en moneda CLP tuviera mucha influencia de las fechas de pago de sueldo (día 25 y fin de mes) y también del pago de deudas y tarjetas de crédito (usualmente asociado al día 5 de cada mes), mientras que segmentos de productos menos comunes (tales como cuentas remuneradas y/o cuentas en USD) no tuvieran un marcado comportamiento estacional intrames, lo cual se vio reflejado en la no existencia de variables *dummies* estacionales significativas.

Otro tema a discutir es la similitud de los resultados obtenidos mediante todo el trabajo estadístico realizado versus el entregado por el software estadístico, el cual en pocos segundos realiza el mismo trabajo y de forma muy similar. Una posible conclusión de estos resultados obtenidos es que no es necesario invertir tiempo y esfuerzo en comprobar las condiciones

estadísticas de las series con las que se trabajaron, si existe un software que lo realiza automáticamente. Sin embargo, cuando estos modelos se utilizan en instituciones bancarias y, en general, en cualquier ámbito laboral que tenga un impacto relevante es necesario e imperante sustentar de la forma más robusta posible todos los modelos que se implementen en las empresas, para así superar de buena manera las validaciones y auditorías tanto internas de la misma institución como externas, realizadas usualmente por el regulador.

Respecto a los resultados finales hay una serie de temas a analizar. Uno de ellos es que no todos los segmentos fueron susceptibles a la inclusión de variables estacionales o exógenas, mostrando mejores resultados solo al considerar sus componentes no estacionales, explicándose solo por sus datos anteriores, lo cual no estaba dentro de la planificación original al comenzar este trabajo. Una posible razón a este comportamiento es que la serie de saldos diarios no se ve afectada por modificación de variables tan ajenas al producto en sí, como lo son las variables macroeconómicas, y, por ende, la decisión de retirar el dinero de estas cuentas un día en particular no depende de variables tan generales. Así, una posible solución sería incluir variables más específicas relacionas con los clientes que pertenecen a los distintos segmentos, así como variables dummies que capturen si el cliente está desempleado o no, en vez de intentar explicarlo por la variable macroeconómica de la tasa de desempleo nacional. Para lograr lo anterior, es necesario abordar de otra forma la data disponible, ya que al trabajarla como serie de tiempo se pierde ese enfoque más centrado en el individuo en particular.

Otro de los resultados finales que pudiesen no cumplir con las expectativas iniciales, es que la performance en la proyección de aquellos modelos que incluyen las variables estacionales y exógenas resultó peor en algunos casos que el modelo original inicial sin más variables explicativas y solo considerando los sucesos pasados de la misma serie de saldos. La razón más directa que explica lo anterior es que las variables estacionales y exógenas que se incluyeron en el modelo no tengan valor explicativo sobre la serie de saldos de cuentas corrientes. Otra posible razón a este resultado es que el horizonte temporal considerado para este trabajo aún está siendo alterado por una serie de variables y sucesos externos, por ende, el comportamiento de la data de validación puede ser muy distinto a la data de modelación, y por ello se observa que los modelos aplicados sobre la data de entrenamiento se ajustan de mucha mejor manera que al compararlos contra la data de validación. Nuevamente ampliar el horizonte temporal parece ser una solución factible, esperando a que las condiciones externas de mercado se estabilicen, permitiendo así capturar el comportamiento de mejor manera.

A modo de conclusión, el trabajo desarrollado permite comprender de mejor manera cuáles son las variables más relevantes que afectan y explican el comportamiento de los distintos tipos de segmentos de clientes y productos asociados a las cuentas corrientes. También es necesario mencionar que realizar un completo análisis estadístico que permita justificar cada decisión y resultado obtenido fue fundamental en el desarrollo de este trabajo. Sin embargo, también es fundamental comprender y realizar un análisis financiero y de mercado de los resultados, ya que como finalidad de este trabajo es el aporte a la gestión diaria de la mesa de dinero del banco, la

cual por su naturaleza requiere que el sustento estadístico y los resultados que estos genera, posean sentido de negocio y estén acorde a la realidad del banco y del mercado en general.

Cabe mencionar que con el desarrollo de esta memoria se podrá incluir la metodología y la programación realizada en las labores diarias de la mesa de dinero, ayudando a la proyección diaria de cuentas corrientes que hasta la fecha de cierre de este trabajo no se posee. Además del aporte a la gestión que realiza este trabajo, también genera valor al reducir el riesgo de liquidez del banco, pues permitirá anticipar de mejor manera posibles pérdidas de liquidez, lo cual también tendrá como efecto secundario una estabilidad en los indicadores que monitorean las áreas de riesgo.

Finalmente, con la intención de mejorar lo realizado y desarrollar un trabajo a futuro sobre lo expuesto en esta memoria, se propone ampliar el horizonte temporal para abarcar más efectos macroeconómicos que al cierre de este trabajo aún están en desarrollo. Además, sería interesante aplicar distintas metodologías de predicción, de forma de encontrar un mejor modelo fuera de la familia de modelos ARIMA, o modificar el enfoque de series de tiempo por un enfoque de datos de panel, permitiendo analizar características de los clientes que no fueron consideradas en este trabajo. Por último, este trabajo puede dar paso a aplicar modelos predictivos a otros productos financieros que sean relevantes para las instituciones bancarias en general y que no tengan dentro de su gestión diaria, las cuales puedan afectar gravemente a la liquidez del mercado en situaciones de incertidumbre local e internacional.

Bibliografía

[1] Comisión para el Mercado Financiero (CMF) http://www.cmfchile.cl/portal/principal/605/w3-channel.html [En línea]

[2] Banco Central de Chile, Normas Financieras Capítulo III.B.12 [En línea]

[3] Instituto Nacional de Estadísticas INE [En línea]

[4] Superintendencia de Pensiones, Retiro de fondos de pensiones https://www.spensiones.cl/portal/institucional/594/w3-propertyvalue-10411.html [En Línea]

[5] Denis Kwiatkowski, Peter C.B. Phillips, Peter Schmidt, Yongcheol Shin. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?, Journal of Econometrics, 54, pp. 159-178, North-Holland

[6] Dickey, D.A. and W.A. Fuller, 1979. Distribution of the estimators for autoregressive time series with a unit root. Journal of the American Statistical Association, 74, pp. 427-31.

Anexo

La Tabla 83 del MSI muestra la forma de categorizar los distintos productos que ofrezcan las entidades bancarias y que son reportados en el reporte C46.

Código	Origen flujo	Descripción
001	Obligaciones a la vista – Minoristas (*)	Comprende las cuentas corrientes y otras obligaciones a la vista
002	Obligaciones a la vista – mayoristas (*)	Comprende las cuentas corrientes y otras obligaciones a la vista
015	Cuentas de ahorro a plazo giro incondicional	
016	Cuentas de ahorro a plazo giro diferido	
101	Operaciones de retro compra	
201	Uso de líneas de crédito y liquidez otorgadas por el Banco Central del país	Deben informarse solo los flujos de egreso vinculados al reembolso de montos ya utilizados.
202	Uso de líneas de crédito otorgadas por otros bancos del país	
203	Uso líneas de crédito obtenidas en el exterior	
204	Uso de otras líneas de crédito obtenidas en el país	
301	Depósitos y captaciones a plazo – minoristas (*)	
302	Depósitos y captaciones a plazo – mayoristas (*)	
303	Obligaciones por letras de crédito y bonos hipotecarios	
304	Obligaciones por bonos y efectos de comercio	

305	Otros préstamos del exterior no vinculados a líneas de crédito	Comprenden obligaciones con el exterior no contempladas en otros códigos. (**)
306	Obligaciones o compromisos con el Banco Central de Chile no vinculadas a líneas de crédito	Incluye cuentas y documentos por pagar, otros préstamos obtenidos en el país y cualquier otro pasivo o compromiso de pago con contrapartes locales, dentro o fuera de balance.(**)
307	Obligaciones o compromisos con otros bancos del país no vinculadas a líneas de crédito	
308	Otras obligaciones o compromisos en el país no vinculadas a líneas de crédito	
400	Fondos disponibles que no pertenezcan a encaje ni a la reserva técnica	Fondos disponibles en caja o depositados en la cuenta corriente del Banco Central de Chile; remesas en efectivo en tránsito entre oficinas de una misma empresa bancaria; remesas en efectivo al Banco Central de Chile; y efectivo en custodia en las bóvedas de las empresas especializadas de transporte de valores con las que se mantenga contratos vigentes de servicios.

401	Canje y overnight	
402	Encaje	
403	Inversiones financieras computables a valor razonable	Instrumentos a que se refiere el numeral 2.1 del Título III del Capítulo 12-20 de la RAN.
501	Operaciones de retro venta – otros bancos del país	
502	Operaciones de retro venta – contrapartes locales no bancarias	(**)
503	Operaciones de retro venta – contrapartes extranjeras	
600	Colocaciones con otros bancos del país, no vinculadas a líneas de crédito	Créditos otorgados a otros bancos o sus filiales.
601	Créditos comerciales no vinculados a líneas de crédito	Flujos asociados a colocaciones para fines productivos, incluyendo leasing comercial y *factoring*.
602	Créditos de consumo no vinculados a líneas de crédito	Incluye créditos de consumo y leasing de consumo
603	Créditos hipotecarios de vivienda	Incluye créditos y leasing para vivienda
700	Líneas otorgadas a otros bancos del país	Líneas de crédito o de liquidez otorgadas a otros bancos o a sus filiales.

701	Líneas de crédito y sobregiros – comerciales	Deben informarse separadamente los flujos de egreso asociados al uso estimado de cupos no utilizados y los flujos de ingreso vinculados al reembolso de cupos utilizados o de los cupos que se estime se utilicen en el futuro.
702	Líneas de crédito y sobregiros – consumo	
703	Otras líneas de crédito otorgadas	
800	Inversiones financieras computadas según flujo del emisor.	Instrumentos financieros no derivados mantenidos a término, entregados en garantía o sujetos a cualquier otro tipo de gravamen, destinados para la constitución de la reserva técnica, y aquellos vendidos con pacto de retrocompra.
801	Otras operaciones activas o compromisos no vinculados a líneas de crédito	Incluye flujos asociados a otras cuentas del activo o compromisos que significan flujos a favor, no contemplados en otros códigos.

900	Contratos de derivados	Deben informarse los montos estimados a pagar y a recibir por contratos con instrumentos derivados, conforme se establece en el numeral 2.2 del Título III del Cap. 12-20 de la RAN.