Instituto de Estadística
Universidad de la República
Montevideo-Uruguay

Pasantía de grado
Licenciatura en Estadística

DENSIDAD DE COTIZACIÓN: ESTUDIO DE COMPORTAMIENTO Y PREDICCIÓN

María Elisa Bertinat Tulipano
María Fernanda González Pérez
Junio 2009
Tutor: Ramón Álvarez

CONTENIDO

ÍNDICE DE TABLAS

TABLA DE ILUSTRACIONES

RESUMEN

El propósito de esta investigación es explicar y predecir a corto plazo la densidad de cotización de los afiliados a República AFAP. La densidad de cotización es un indicador de frecuencia de aportación de los afiliados que representa la proporción de aportes que realizan los individuos en un intervalo de tiempo determinado.

Este trabajo pretende brindar un mayor entendimiento de la densidad de cotización, considerando su evolución y la relación con otras variables que caracterizan a los afiliados. Respecto a esto último se concluye que el giro de actividad, la franja salarial por la que aporta y la edad están asociados a la densidad de cotización.
Por ejemplo, los afiliados más jóvenes tienen mayor posibilidad de mejorar sus nivel salarial y su estabilidad laboral con el transcurso del tiempo. El giro de actividad del afiliado y la franja salarial a la que pertenece están asociados a la estabilidad laboral y repercuten por lo tanto en el valor y la evolución de la densidad de cotización. Por ejemplo, personas con salarios elevados del sector público presentan comportamiento de aportación más regular que otras de bajo salario del sector construcción.
Además el valor futuro de la densidad de cotización está condicionado por sus antecedentes.

Durante el proceso de investigación se evalúan diferentes alternativas metodológicas para modelar la densidad de cotización en las que se aplican modelos lineales de regresión sobre diferentes conjuntos de datos. Una particularidad de estos modelos es la incorporación de un componente dinámico al considerar la variable de respuesta en un momento posterior al de las explicativas.
Los mejores resultados de predicción se obtienen cuando se realiza una agrupación de los afiliados en base a sus antecedentes de aportación. Puntualmente se definen tres grupos en función de la variación de la densidad de cotización en un período previo al utilizado para construir los modelos: uno en el que la variación es positiva, otro con variación negativa y un tercero con variación despreciable.
El planteo consiste en identificar el grupo al que pertenece el afiliado y aplicar el modelo correspondiente sólo en caso de que los antecedentes de aportación sean crecientes o decrecientes. En el tercer grupo se decide no aplicar el modelo y utilizar como estimación puntual el último valor disponible, dado que los cambios en la densidad de cotización no ameritan cálculos de predicción.

El mayor alcance de la metodología se logra cuando se aplica a los afiliados cuyo comportamiento en el tiempo de la densidad de cotización es fluctuante. Esta propiedad permite focalizar el análisis en un conjunto de afiliados del cual se tiene mayor incertidumbre respecto al valor futuro. Para estos casos se obtienen estimaciones de un año dado que superan ampliamente en precisión el dato de la densidad de cotización en años inmediatos anteriores.

La metodología está concebida para estimar la densidad de cotización futura con expresiones construidas a partir de datos del presente y el pasado. Para citar un ejemplo, es posible estimar la densidad de cotización de de 2010 empleando los

modelos construidos a partir de datos del período 2004 – 2008. Los grupos se determinan en función de la evolución de la densidad de cotización entre 2006 – 2008 y las variables explicativas estan conformadas con los valores a 2008.
En función de los resultados obtenidos, el modelo seleccionado como referencia considera el período 2003-2007 con rezagos de dos años. Este modelo es el que contempla los datos recientes y además arroja resultados de predicción satisfactorios.

Los modelos construidos a partir de un intervalo de tiempo determinado son aplicables a otros períodos. No obstante, se considera necesario actualizar periódicamente los modelos para mantener las bondades explicativas y predictivas.
La actualización permite incorporar información y mejorar los resultados al aumentar el rezago utilizado para la agrupación. En la aplicación actual el rezago se establece en 2 años evitando que el contexto de crisis del año 2002 distorsione el desempeño del modelo. Es importante evaluar la coyuntura económica al momento de la actualización dado que la metodología está concebida para un escenario de estabilidad.
La información disponible a 12 años del inicio del régimen mixto junto con la evolución de la densidad de cotización y la influencia del contexto económico provocan que sólo puedan realizarse estimaciones a corto plazo, entre dos y 4 años.

Además de las predicciones a corto plazo mediante modelos lineales se obtiene una función de probabilidad para la densidad de cotización. En particular la distribución asociada es una Beta en la que la media es el promedio de las predicciones de los afiliados del grupo y la dispersión se calcula por máxima verosimilitud. Esto proporciona una visión más global del comportamiento de cada grupo agregando a la estimación puntual datos de su distribución, conociendo cuál es la probabilidad con la que se registran determinados rangos de valores del recorrido.

Dentro de las posibles aplicaciones de los cálculos de predicción puede mencionarse el cálculo jubilatorio y su introducción en distintos estudios de cobertura del sistema provisional, entre otros.

CAPITULO 1: INTRODUCCIÓN GENERAL

1.1. INTRODUCCIÓN

En 1829 surge en Uruguay la primera Ley sobre pasividades con cobertura para militares, viudas y huérfanos. En 1967, luego de un período de 138 años se crea el Banco de Previsión Social (B.P.S.), un organismo que centraliza las diferentes cajas de prestaciones existentes hasta el momento.
Con la Ley 16713 de reforma de la seguridad social en 1996 se determina un régimen mixto compuesto por dos pilares de ahorro, uno de reparto o de "solidaridad intergeneracional" administrado por el B.P.S. y otro de ahorro individual a cargo de las Administradoras de Fondos de Ahorro Previsional (A.F.A.P.). De esta manera se establece el sistema de seguridad social que rige en la actualidad.

Según la Organización Internacional de Trabajo (O.I.T.) la seguridad social se define como: "la protección que la sociedad proporciona a sus miembros, mediante una serie de suposiciones públicas contra los infortunios económicos y sociales que – de lo contrario- serían ocasionados por la introducción o reducción considerable de ingresos a raíz de contingencias como la enfermedad, maternidad, accidentes de trabajo y enfermedades profesionales, desempleo, invalidez, vejez y muerte".

En Uruguay el sistema de seguridad social comprende prestaciones contributivas, como la jubilación común, y no contributivas como pensiones y asignaciones familiares. En las primeras el aporte realizado a partir de los salarios permite el acceso a una prestación al momento del retiro. Las segundas están concebidas para amparar a la población ante situaciones como invalidez, vejez, maternidad, etc. Este trabajo se enfoca en el pilar de ahorro individual que funciona exclusivamente a partir de prestaciones contributivas.

La principal fuente de financiación de los beneficios es el aporte obligatorio de los trabajadores uruguayos a la seguridad social, el que representa el 15% del salario nominal. De acuerdo a la reglamentación vigente los trabajadores pueden elegir entre distribuir dicho aporte entre BPS y AFAP o destinarlo al BPS exclusivamente. Solamente están obligados a afiliarse a una AFAP aquellos trabajadores menores de 40 años al 1º de abril de 1996 con salario nominal superior los $19.805, o aquellos que hayan ingresado al mercado laboral después del 1 de abril de 1996 en una actividad amparada por el BPS sin importar la edad. El mecanismo de distribución del aporte para quienes se afilian a una AFAP está regulado por el Artículo 8 de la Ley 16.713. Aquellos afiliados que optan por este artículo destinan la parte correspondiente a la AFAP cada vez que aportan a la seguridad social, los que no lo hacen aportan solo cuando superan el tope salarial de $19.805.

Las prestaciones contributivas a las que puede acceder un trabajador son jubilación común, jubilación por edad avanzada, jubilación parcial por ahorro, jubilación por

incapacidad total, subsidio transitorio por incapacidad parcial, subsidio por desempleo y pensión de sobrevivencia.

Según la Ley 18.395 se configura jubilación por incapacidad total cuando al trabajador le sobreviene una incapacidad absoluta y permanente para el desempeño de todo trabajo, debiendo cumplir además con los siguientes requisitos: en el caso de los trabajadores en actividad dos años de servicio para los mayores de 25 años y para los menores seis meses de servicio inmediatamente previos a la incapacidad. En el caso de los trabajadores sin actividad las condiciones son 10 años de servicio y residencia en el país.
Tienen derecho a recibir la prestación de subsidio transitorio por incapacidad parcial los trabajadores que padecen una incapacidad absoluta y permanente para el empleo o profesión habitual, sobrevenida en actividad o en períodos de inactividad compensada cualquiera sea la causa que la haya originado siempre que acredite: 2 años de servicio y en el caso de los menores de 25 años, seis meses de servicio. Para el subsidio especial por desempleo las condiciones son 58 años de edad, 28 años de servicios y un año de desempleo forzoso. Vale recordar que los subsidios no son permanentes sino que tienen un tiempo estipulado, en el caso del subsidio por desempleo la duración es dos años, ya que luego de este período el individuo cumple con los requisitos necesarios para configurar la causal por jubilación común que se detalla más adelante.
Las pensiones de sobrevivencia son prestaciones que cubren la contingencia de muerte del afiliado y se otorga a personas que tengan vinculación con el fallecido, con criterios establecidos en la Ley 16.713.

Las condiciones para configurar causal de jubilación por edad avanzada en vigencia desde febrero de 2009 son: 70 años de edad y 15 de servicio, 69 de edad y 17 de servicio o 68 de edad y 19 de servicio. A partir de enero de 2010 se admitirán además de las anteriores: 67 años de edad y 21 de servicio, 66 de edad y 23 de servicio y 65 años de edad y 25 de servicio.
En la jubilación común los requisitos mínimos son 60 años de edad y 35 años de servicios registrados en la historia laboral; a partir de julio de 2009 la Ley 18.395 reduce la exigencia de años de servicio a 30. La causal por edad avanzada se configura sólo por BPS, con 19 años de aporte y 68 años de edad, o con 17 años de trabajo y 69 años de edad. Por último, a la jubilación parcial por ahorro, prestación brindada únicamente por las AFAP's, se accede con 65 años de edad sin requisitos de años de servicio. La nueva legislación introduce un cambio en el cómputo de años de servicio para las madres.

En monto de la prestación de los afiliados al sistema mixto consta una parte correspondiente a BPS y otra a la AFAP que se calculan de forma diferente. La cuota parte correspondiente a BPS se calcula en el caso de la jubilación común en base a una tasa de reemplazo aplicada al monto básico jubilatorio, que varía entre el 50% y el 82.5% dependiendo de la edad y la cantidad de años de servicio. De acuerdo al artículo 27 de la ley 16.713 el sueldo básico jubilatorio "…será el promedio mensual de las asignaciones computables actualizadas de los diez últimos años de servicios registrados en la historia laboral, limitado al promedio mensual de los veinte años de

mejores asignaciones computables actualizadas, incrementado en un 5% (cinco por ciento). Si fuera más favorable para el trabajador, el sueldo básico jubilatorio será el promedio de los veinte años de mejores asignaciones computables actualizadas…"

En el caso de las AFAP los afiliados generan a través de sus aportes una renta vitalicia, calculada en base a lo ahorrado por el individuo, indicadores de esperanza de vida y tasas de interés técnico. De esta forma, la parte de jubilación del afiliado que corresponde a la capitalización individual depende del aporte salarial, la rentabilidad, la frecuencia de aporte, la edad y la esperanza de vida entre otros elementos. Los factores edad y frecuencia de aportación determinan además el momento en que se configura la causal. En este contexto, la regularidad de aportación se convierte en un factor preponderante en el ámbito de la seguridad social uruguaya. La disminución de los años de servicio establecida por la Ley 18.395 no quita validez al estudio de frecuencia de aportación, por el contrario, se considera pertinente evaluar su efecto en la cobertura de la seguridad social.

La importancia de la frecuencia de aportación tiene otros antecedentes además del mecanismo de funcionamiento de las AFAP, uno de ellos es la introducción del Registro de Historia Laboral que permite el conocimiento de los antecedentes de aportación de los trabajadores y desplaza el testimonio de terceros. Estos cambios dan lugar a diversas investigaciones en las que se evalúa su impacto en la cobertura de seguridad social.

En síntesis, el comportamiento de aporte de los afiliados a la seguridad social es de interés tanto para las administradoras de fondo como para los trabajadores. Este comportamiento se resume a partir de la variable densidad de cotización, cuya funcionalidad motiva su estudio en profundidad y la búsqueda de estimaciones futuras.

El presente informe se desarrolla en el marco del Convenio República AFAP S.A[1] – IESTA como pasantía final de las estudiantes María Elisa Bertinat y María Fernanda González bajo la tutoría del profesor Ramón Álvarez. El propósito de la investigación es identificar un modelo explicativo y predictivo de la densidad de cotización futura de los afiliados a RAFAP.

En este capítulo se presentan los antecedentes, la metodología de trabajo y la estructura general del documento.

1.2. OBJETIVO

1.2.1. Objetivo General

El objetivo general es la obtención de estimaciones de densidad de cotización de los afiliados a corto plazo, entendiendo como corto plazo un tramo no superior a los cuatro años.

[1] De aquí en adelante RAFAP

1.2.2. Objetivo Específico

- Construir un modelo predictivo de la densidad de cotización futura a partir de variables socioeconómicas y demográficas de los afiliados y antecedentes de aportación.
- Explicar e interpretar la relación existente entre las características del afiliado y la frecuencia de aporte.

1.3. ANTECEDENTES

Existen numerosos estudios relacionados a la seguridad social que tratan el tema de la densidad de cotización. Álvaro Forteza en su artículo "El acceso a la jubilación o pensión en Uruguay..." [5] utiliza la densidad de cotización para estimar el porcentaje de trabajadores que lograrán jubilarse en determinado rango de edad, bajo las exigencias de 35 años de servicio. En su artículo analiza el comportamiento de densidad de cotización de los trabajadores registrados en BPS y estima funciones de distribución para los aportes a lo largo del ciclo de vida. En la metodología utiliza supuestos poco convenientes como independencia en el tiempo de la probabilidad de aportar, y permanencia en un estrato salarial durante el ciclo de vida.
En un trabajo del siguiente año: "Seguridad social y género en Uruguay……" [6] Forteza junto con otros autores levantan el supuesto de independencia de los sucesos de aportación. Esta nueva propuesta considera dependencia de aportación de meses consecutivos.

Otro estudio en el que se introdujo el tema de la aportación en Uruguay fue realizado por Gabriel Lagomarsino y Bibiana Lanzilotta: "Densidad de aportes a la seguridad social…".[13] Para explicar la cantidad de aportes realizados entre enero de 1997 y diciembre de 2003 los autores emplean un modelo econométrico basado en variables de conteo con distribución Binomial Negativa. Esta técnica tiene la ventaja de captar la naturaleza discreta y no negativa de la cantidad de aportes, pero asume independencia de los sucesos de aportación. El estudio se centra en la interpretación del comportamiento de la densidad de cotización y no en la predicción. La metodología de este artículo es similar a la aplicada por Bertranou y Sánchez en su trabajo sobre la densidad de aportes a la Seguridad Social en la Argentina: "Características y determinantes de la densidad de aportes…" [3]

En el año 2006 Jimena Pardo, directora de RAFAP, elaboró el documento "El mercado de las AFAP en Uruguay: caracterización y proyección de los resultados representativos de la empresa líder". [14] El objetivo es estudiar el comportamiento de los afiliados a RAFAP en lo referente a la aportación y otras variables de interés. En función de estos resultados se realizan estimaciones sobre el porcentaje de afiliados que lograrían jubilarse. La metodología consiste en la definición de un grupo representativo de individuos en función de características socio-económicas y la estimación del porcentaje de jubilaciones. Se asume que no varían las características de los afiliados entre el momento del cálculo y el retiro.

Lagomarsino, Bertranou y Sánchez, aplican modelos de conteo, en los que la variable estudiada es la cantidad de aportes en un periodo de tiempo determinado. Estas distribuciones se adaptan a variables discretas no negativas, característica de la cantidad de aportes. En las investigaciones de Lagomarsino y Bertranou y Sánchez, la distribución asignada a la cantidad de aportes es una Binomial Negativa, donde el éxito es aportar y la probabilidad de éxito es la densidad de cotización.
Forteza desarrolla un procedimiento para la estimación de funciones de distribución de conteo de períodos de aportes a lo largo del ciclo de vida de los trabajadores; considerando dos escenarios, una más optimista, que va desde 1996 a 1998, y otro desde 1996 a 2005 que incluye la crisis. Específicamente estima para ambos escenarios el porcentaje de trabajadores que tienen S meses de servicio con E años de edad tomando en cuenta todas las combinaciones posibles. Para cada individuo tipo estima la probabilidad de computar el mes m para su jubilación. Esta probabilidad se estima con la media de aportes de cada grupo para cada edad, lo que permite el estudio de la evolución durante el ciclo de vida. La limitación más importante presente en dichas estimaciones es que se ha supuesto que la probabilidad de que un mes sea de servicio es independiente de lo que ocurre en el resto de los meses.

El objetivo de la presente investigación no se limita a interpretar el comportamiento de la densidad de cotización sino que el principal cometido es la predicción de la variable a corto plazo. La fase exploratoria, a diferencia de los antecedentes, es un insumo necesario pero no el fin de este trabajo. No se asume además la independencia intertemporal de los aportes, sino que se contemplan los antecedentes de aportación de los afiliados. El principal elemento que distingue a esta propuesta es la introducción de la dimensión tiempo y el énfasis en el valor futuro de la densidad de cotización.

CAPÍTULO 2: MARCO TEÓRICO

2.1. MODELOS DE REGRESIÓN LINEAL

En los modelos de regresión se pretende modelar la relación entre variables a partir de un método matemático. Un modelo provee una herramienta teórica que permite entender mejor un fenómeno de interés, describirlo, predecir valores desconocidos e identificar las relaciones entre variables observadas.

El modelo para la i-ésima observación se expresa como:

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i, i = 1,2,...,n. \quad (1)$$

donde ε_i es el término del error, y_i es la variable de respuesta o dependiente y $x_i = (x_{i1} \quad x_{i2} \quad \cdots \quad x_{iK})$ es el vector de variables explicativas o predictoras para la i-ésima observación. Los parámetros β se estiman considerando la información de las *n* observaciones y de las variables explicativas asociadas.

Una observación importante es que el término regresión lineal refiere a un modelo que es lineal en los parámetros. Por lo tanto las expresiones $E(Y_i / x_i) = \beta_0 + \beta_1 x_{i1}^2, i = 1,2,...,n$ y $E(\log(Y_i) / x_i) = \beta_0 + \beta_1 (1/x_{i1}), i = 1,2,...,n$ son regresiones lineales.
Por ejemplo algunas transformaciones sobre la variable de respuesta son logit y logarítmica. En la primera el modelo resultante tiene como variable dependiente el logit de y_i, es decir $\log(y_i/(1-y_i))$. En el caso de la transformación logarítmica se modela $\log(y_i)$ lo que genera que la variable dependiente tome valores no negativos.

En los modelos de regresión lineal la variable dependiente y el error se asumen como variables aleatorias y las variables predictoras son conocidas y constantes. Otros supuestos adicionales para el término del error y para las observaciones son los siguientes:

1. $E(\varepsilon_i) = 0, i = 1,2,...,n \; E(y_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}, i = 1,2,...,n$
2. $Var(\varepsilon_i) = \sigma^2, i = 1,2,...,n \; Var(y_i) = \sigma^2, i = 1,2,...,n$
3. $Cov(\varepsilon_i, \varepsilon_j) = 0, i, j = 1,2,...,n, i \neq j \; Cov(y_i, y_j) = 0, i, j = 1,2,...,n, i \neq j$

Los supuestos entonces plantean (1) que el modelo es correcto, (2) que la varianza de las variables dependientes es constante e independiente de las variables explicativas y (3) que las observaciones no están correlacionadas entre sí.

El modelo puede reescribirse en forma matricial, reordenando las n ecuaciones como sigue:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & \cdots & x_{2k} \\ \vdots & \vdots & & & \vdots \\ \vdots & \vdots & & & \vdots \\ 1 & x_{n1} & \cdots & \cdots & x_{kk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \vdots \\ \varepsilon_n \end{pmatrix}; \; y = X\beta + \varepsilon \qquad (2)$$

La matriz X , de dimensión (n) x (k+1), contiene los datos experimentales, β es el vector de coeficientes de (k+1) elementos, y y ε, son vectores de dimensión (n) x (1) que contienen la variable dependiente y el error respectivamente.

Cuando no hay supuestos distribucionales los coeficientes de la regresión se obtienen por estimación de mínimos cuadrados, lo que se corresponde con la fórmula: $\hat{\beta} = (X'X)^{-1}X'y$. El parámetro de la varianza de las variables de dependientes se obtiene como un promedio: $\frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - x'_i\,\hat{\beta}_i)^2$, con $\hat{\beta}_i$ el vector de coeficientes estimados asociado a la i-ésima observación.

Cuando se incorpora el supuesto de normalidad en las observaciones el mecanismo de estimación de parámetros difiere. Los supuestos de normalidad $Y \sim N_n(X\beta, \sigma^2 I)$, o $\varepsilon \sim N_n(0, \sigma^2 I)$ permite obtener los parámetros por máxima verosimilitud.

2.1.1. PRUEBAS DE HIPÓTESIS

Las pruebas de hipótesis buscan evidencia en los datos de una muestra para refutar o no una afirmación acerca de un parámetro poblacional. En el problema se plantean dos hipótesis complementarias las que se denotan hipótesis nula e hipótesis alternativa.
La prueba de hipótesis es una regla que permite identificar para qué valores muestrales la hipótesis nula se rechaza o no.

En el caso de los modelos de regresión lineal la prueba de hipótesis para todos los regresores el objetivo es determinar la relevancia del conjunto de variables predictoras para explicar la variable independiente. En caso de no rechazar la hipótesis nula se acepta el modelo.
De manera análoga las pruebas para subconjuntos de coeficientes testean la validez de un subconjunto de variables para explicar la variable de interés, dado que se incluyen en el modelo el resto de las variables. Para desarrollar las pruebas se asume distribución normal de la variable dependiente: $Y \sim N_n(X\beta, \sigma^2 I)$.

Antes de definir los estadísticos asociados a la prueba es necesario definir la suma de cuadrados de las observaciones y una particularidad. Sea la suma de

cuadrados: $y'(I - 1/n\, J)y$, puede particionarse en dos términos $\hat{\beta}_1' X' X\hat{\beta}_1 + \sum_{i=1}^{n}(y_i - \bar{y})^2 - \hat{\beta}_1' X' y$ donde el primer término es la suma de cuadrados explicada por el modelo, SCE, y los dos últimos son la suma de cuadrados de los residuos,SCR. Se prueba que si la variable dependiente se distribuye normal multivariada $Y \sim N_n(X\beta, \sigma^2 I)$ entonces SCE/σ^2 y SCR/σ^2 tienen distribución $\chi^2(k, \lambda_1)$ [2] y χ^2 (n-k-1) y son independientes.

La prueba de significación del modelo tiene como hipótesis nula Ho) β_j =0 para todo j, y como alternativa H1) $\beta_j \neq 0$ para algún j. El estadístico asociado a la prueba de significación es: $F = \dfrac{SCR/k}{SCE/(n-k-1)}$, el que bajo Ho) se distribuye F (k, n-k-1). Se rechaza Ho) si $F > F_{\alpha,k,n-k-1,\alpha}$, donde el valor de referencia es la probabilidad acumulada en el extremo superior de la distribución central F.
En forma equivalente puede considerarse el p-valor, tal que se rechaza Ho) si el p-valor es menor que el nivel α de la prueba.
La prueba de significación de la variable j-ésima es Ho) $\beta_j = 0$. Este es un caso particular del anterior en donde k es igual a 1. La prueba evalúa si la j-ésima variable se correlaciona en forma lineal con la variable dependiente y el estadístico asociado es, $t_j = \dfrac{\hat{\beta}_j}{s g_{jj}^{1/2}}$ con g_{jj} el j-ésimo elemento de la diagonal de $(X'X)^{-1}$.El estadístico bajo Ho) cierta y con el supuesto de normalidad de las variables se distribuye t-student (n-k-1), rechazando Ho) si $|t_j| > t_{\alpha/2,n-k-1}$

2.1.2. INTERVALOS DE CONFIANZA PARA LOS PARÁMETROS

Un intervalo de confianza para un parámetro es cualquier par de funciones L(x) y U(x) de una muestra que satisfacen $L(x) \leq U(x)$, con x una realización de la variable aleatoria. Se define la probabilidad de cobertura del intervalo $[L(x), U(x)]$ como la probabilidad de que el intervalo aleatorio contenga al parámetro de interés.

Existen diferentes métodos para obtener los intervalos, uno de ellos es a través de la región de rechazo de la prueba de significación. Se presenta el siguiente ejemplo aplicado a los coeficientes estimados de un modelo de regresión lineal.
Para los parámetros β_j, donde la distribución asociada es $t_j = (\hat{\beta}_j - \beta_j)/s\sqrt{g_{jj}}$, la t de student, el intervalo de confianza se obtiene despejando de la región de rechazo

[2] $\lambda_1 = \beta_1' Xc' Xc\beta_1 / 2\sigma^2$

en este caso definida por $\{\beta : |\hat{\beta}_j - \beta_j| > t_{\alpha/2,n-k-1}.s.\sqrt{g_{jj}}\}$, lo que da lugar a un intervalo de confianza: $[-t_{\alpha/2,n-k-1}.s.\sqrt{g_{jj}}, t_{\alpha/2,n-k-1}.s.\sqrt{g_{jj}}]$.Este intervalo contiene con probabilidad (1-α) 100 al verdadero parámetro.

2.1.3. MEDIDAS DE AJUSTE DEL MODELO: COEFICIENTE DE CORRELACIÓN MÚLTIPLE

Otro indicador para los modelos es el coeficiente de correlación múltiple: R^2. Este coeficiente brinda una medida del ajuste del modelo, indicando la capacidad de las variables predictoras para explicar la variable dependiente. El coeficiente de obtiene del cociente:

$$R^2 = \frac{SCE}{SCT}. \tag{3}$$

El indicador pertenece al intervalo [0,1], aumenta al incrementar el número de predictores, es invariante ante cambios de escala de y entre otras cosas. Valores del coeficiente cercanos a 1 indican que el modelo es adecuado. Sin embargo como el indicador depende del tamaño de k es posible que si k es una proporción grande de n el coeficiente sea alto aunque el modelo no sea bueno.

Para superar esto se define un indicador ajustado, que tiene el mismo recorrido que el anterior, y que además de ser más preciso permite comparar modelos que difieren en el número de variables:

$$R_a{}^2 = \frac{(n-1)R^2 - k}{n-k-1} \tag{4}$$

2.1.4. VALIDACIÓN Y DIAGNÓSTICO DEL MODELO

En este punto se muestran algunos aspectos que permiten chequear la validez del modelo a partir del análisis de los residuos, detección de outliers y de observaciones influyentes.

En los modelos usualmente se asume $E(\varepsilon) = 0$, $Cov(\varepsilon) = \sigma^2 I$, y una vez estimados los parámetros es posible estimar la componente no observable:

$\hat{\varepsilon} = y - X\hat{\beta} = y - \hat{y}$

Una propiedad de los errores estimados es la varianza no constante. Ésta se calcula como: $Var(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ii})$, donde h_{ii} es el elemento de la diagonal de la matriz H; con $H = \frac{1}{n}J + X_c(X'_c\, X_c)^{-1}X'_c$, J una matriz cuadrada de dimensión n formada por unos, $X_c = (I - \frac{1}{n}J)X$ e I la matriz identidad.

El elemento h_{ii} de H es menor igual que uno, con lo cual la varianza del residuo será tanto más pequeña cuanto más cerca de uno esté h_{ii}.

Existen diferentes mecanismos para corregir la varianza no constante de los residuos, por ejemplo dividiendo cada valor entre su desvío. Otro método considera una estimación para la varianza en la que excluye la i-ésima observación, con lo que se obtienen los residuos estudentizados: $t_i = \frac{\hat{\varepsilon}}{s_{(i)}\sqrt{1-h_{ii}}}$, donde $s_{(i)}$ es el desvío estimado con (n-1) de las n observaciones. Puede probarse que si la i-ésima observación es un outlier existe más posibilidad de ser detectado con esta última técnica de estandarización.

Los n valores del error estimado se aplican en diversos gráficos y pruebas con los que se contrasta la validez del supuesto distribucional. Uno de los gráficos considera los valores de $\hat{y}$ en las abscisas y de $\hat{\varepsilon}$ en las ordenadas y es de esperar que de cumplirse el supuesto, el gráfico no muestre ningún patrón de comportamiento.

En algunos casos los modelos se comportan correctamente para la mayoría de los datos pero algunos residuos son mucho mayores que otros, lo que puede deberse a la presencia de observaciones atípicas ("outliers"). Un outlier puede ser el resultado de un error de medición, provenir de otra población o ser simplemente una observación inusual para el conjunto de estudio.

Una opción para evaluar la influencia de estas observaciones en el modelo es comparar dos modelos, uno con las n observaciones y otro que excluya los outliers.
Para chequear la existencia de outliers puede considerarse un gráfico de $\hat{\varepsilon}$ versus $\hat{y}$, o comparar los residuos estimados en el modelo completo con los estimados por un modelo obtenido sin considerar los outliers.
Si no hay explicación para los valores que se registran una solución puede ser descartar esas observaciones para construir el modelo, o pueden utilizarse métodos robustos para corregir las observaciones en lugar de eliminarlas.

Además pueden existir observaciones influyentes, que son aquellos puntos que tienen mayor impacto en las estimaciones. Este tipo de observaciones pueden ser buenas o malas para el modelo, en algunos casos puede reducir la varianza de las estimaciones y en otros puede afectar drásticamente el ajuste del modelo.
Una herramienta para identificar estas observaciones es una medida de influencia: h_{ii}, si es alta entonces la observación es influyente.
Otra forma de evaluar si una observación es influyente es comparar el modelo completo con uno que excluya esa observación, cuyas estimaciones denotamos con $\hat{\beta}_{(i)}$.Para ello se considera la distancia de Cook definida como: $D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})' X' X (\hat{\beta}_{(i)} - \hat{\beta})}{(k+1)s^2}$. Si la distancia es grande la observación tiene influencia en las estimaciones.

En la etapa de diagnóstico del modelo se desarrollan un conjunto de pruebas que permiten evaluar el cumplimiento de supuestos, la presencia de outliers y de observaciones influyentes. Las pruebas consideradas en esta etapa son: la prueba de

bondad de ajuste de Kolomogorov-Smirnov, la prueba de normalidad de Shapiro-Wilks, la prueba para detectar heteroscedasticidad de Breusch-Pagan y la prueba para detección de outliers en base al ajuste de Bonferroni.

La prueba de Kolmogorov-Smirnov es una prueba no paramétrica utilizada para evaluar la bondad de ajuste de dos distribuciones, inspirada en el teorema fundamental de la estadística de Glivenko-Cantelli.

Dada una muestra X1,X2, · · · ,Xn de tamaño n proveniente de una distribución F , la prueba de hipótesis asociada es $H_0 : F = F_0(x); H_1 : F \neq F_0$ para una cierta distribución $F_0(x)$.
El estadístico asociado a la prueba es la diferencia máxima entre ambas distribuciones: $D = \sup_x |F_n(x) - F_0(x)|$, donde $F_0(x)$ es la función correspondiente a la población de interés especificada en la hipótesis nula y $F_n(x)$ es la función de distribución muestral $F_n(x) = \sum_{i=1}^{n} 1_{\{X_i, +\infty\}}(x) = \sum_{i=1}^{n} 1_{[-,\infty,x]}(X_i)$. Si la distancia entre ambas distribuciones tiende a cero entonces no se rechaza que la distribución de las observaciones es normal. En el caso particular de los modelos de regresión lineal se evalúa si la distribución de los residuos es normal, $F_0(x) = \Phi(x)$.

La prueba de normalidad de Shapiro-Wilk evalúa si una muestra aleatoria, $(x_1, x_2, ..., x_n)$ de tamaño n proviene de una población con distribución normal.
El método considera la muestra ordenada, $(x_{(1)}, x_{(2)}, .., x_{(n)})$, donde $x_{(i)}$ es el i-ésimo valor de la muestra ordenada, y calcula las diferencias entre: el último y el primero, el segundo y el penúltimo, etc. El estadístico de contraste asociado se calcula como:
$\frac{(n-1)}{n} \frac{(\sum_{i=1}^{h} a_{in}(x_{(n-i+1)} - x_{(i)}))^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$, donde $h = \begin{cases} n/2, n\_par \\ n-1/2, n\_impar \end{cases}$ y a_{in} es un coeficiente de corrección.

En cuanto a la prueba de Breusch-Pagan, se aplica para detectar la heteroscedasticidad de los residuos, es decir para detectar aquellos casos en que la varianza de los residuos depende de las variables independientes.
La prueba aplica la regresión lineal al cuadrado de los residuos estimados del modelo original y evalúa si las variables independientes son significativas en ese nuevo modelo. Por ejemplo si el modelo es $y = X\beta + \varepsilon$, entonces para la prueba se modela $\varepsilon^2 = X\beta + \eta$. Si este modelo es significativo entonces se rechaza la hipótesis de residuos con varianza homoscedástica.

Por última la prueba para detectar outliers se basa en el ajuste de Bonferroni. Este ajuste corrige el p-valor del error estimado más grande, con lo que si $p' = \Pr(t_{n-k-2} > e_{\max})$ entonces el p-valor de Bonferroni es $p = 2np'$.

2.2. ESTIMACIÓN PUNTUAL POR MÁXIMA VEROSIMILITUD

La estimación puntual[3] por máxima verosimilitud es una técnica popular que considera como estimador del parámetro de interés aquel que maximiza la función de verosimilitud, $\hat{\theta}(X)$ .Intuitivamente este valor es aquel con el que la muestra es más verosímil de ser sorteada.

Si $f(x/\theta)$ es la función de densidad de la distribución y $x=(x_1,x_2,\ldots,x_n)$ una realización de la variable aleatoria X, la verosimilitud se define como la función del parámetro θ tal que $L(\theta/x)=f(x/\theta)$. Bajo el supuesto de que la muestra es independiente e idénticamente distribuida se cumple además:

$$L(\theta/x)=L(\theta_1,\theta_2,\ldots,\theta_k/x_1,x_2,\ldots,x_n)=\prod_{i=1}^{n} f(x_i/\theta_1,\theta_2,\ldots,\theta_k) \qquad (5)$$

Lo que distingue a la verosimilitud de la función de densidad es el conjunto de variables que se consideran fijas. La función de verosimilitud permite analizar la posibilidad de que una muestra dada sea observada bajo diferentes valores de θ.

El cálculo del estimador de θ por máxima verosimilitud es un problema de optimización en que debe identificarse el máximo absoluto de $L(\theta/x)$; en el caso que dicha función es diferenciable el estimador coincide con las raíces de la derivada:

$$\frac{\partial L(\theta/x)}{\partial \theta_i}=0$$

Además, si $L(\theta/x)$ es diferenciable una simplificación del cálculo se logra con el logaritmo natural de la verosimilitud, al que se denomina log-verosimilitud. Esta transformación es posible ya que el logaritmo es una función estrictamente creciente en $(0,\infty)$ y esto implica que el máximo de $L(\theta/x)$ y de $\log(L(\theta/x))$ coinciden.

2.3. MODELOS DE CONTEO: DISTRIBUCIÓN POISSON Y BINOMIAL NEGATIVA

En los modelos de conteo la variable a explicar es discreta no negativa y su función de probabilidad es lo que se conoce como un proceso contador de datos. El modelo básico para este tipo de variables se basa en las distribuciones Poisson o Binomial Negativa.

El modelo Poisson para una variable de conteo Y en un periodo de tiempo determinado corresponde a la ecuación:

$$P(Y=k)=\frac{\exp(-\lambda)\lambda^k}{k!} \qquad (6)$$

[3] Un estimador puntual es cualquier función de la muestra, $W(x_1,x_2,\ldots,x_n)$.

y se basa en tres supuestos :

- La media condicional de Y está definida como la función log-lineal de x y β:
$$E(Y_i / x) = \exp(x_i\beta) = \lambda_i$$
- La distribución de Y dado x es Poisson con parámetro λ. Esto implica que la media es igual a la varianza lo que impone una característica de equi-dispersión en la variable de respuesta.
- Las observaciones son independientes e idénticamente distribuidas.

Este modelo captura la naturaleza de los datos y permite hacer inferencia sobre la probabilidad de ocurrencia del evento. Si bien el supuesto de independencia de los eventos en el tiempo resulta restrictivo para algunas aplicaciones, el modelo permite la existencia de dependencia dinámica entre los eventos sucesivos.

Por otra parte el modelo Binomial Negativo, que también se adapta a este tipo de datos, tiene menos restricciones. Esta distribución se obtiene a partir de una Poisson cuyo parámetro se distribuye como una variable aleatoria gamma:

$$f(y_i / \alpha, \lambda_i) = \frac{\Gamma(\alpha + y_i)}{\Gamma(\alpha)\Gamma(y_i + 1)}\left(\frac{\alpha}{\alpha + \lambda_i}\right)^{\alpha}\left(\frac{\lambda_i}{\alpha + \lambda_i}\right)^{y_i} \tag{7}$$

$$\lambda_i = \exp(X_i\beta - \varepsilon_i) = \exp(X_i\beta)u_i$$

Esta propuesta, a diferencia de la anterior, permite introducir un componente estocástico como término de error lo que captura la heterogeneidad y los errores de medida por componentes no observables. Una limitación en este modelo deriva del supuesto de que los sucesos son independientes entre sí.

2.4. MODELOS PARA PROPORCIONES: DISTRIBUCIÓN BETA

La familia de distribuciones Beta, Beta (p, q), es una familia continua en (0,1), indexada por dos parámetros enteros positivos, p y q:

$$f(y) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1} \tag{8}$$

donde $\Gamma(p)$ denota la función gamma:

$$\Gamma(p) = \int_0^{\infty} t^{p-1}e^{-t}\partial t \tag{9}$$

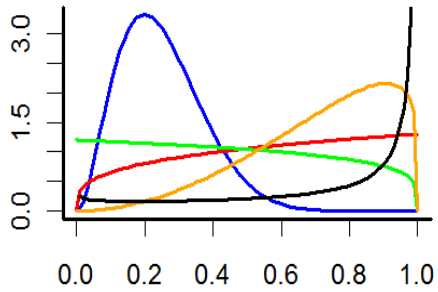
Su media y varianza son respectivamente:

$$E(y) = \frac{p}{p+q}, \tag{10}$$

$$V(y) = \frac{pq}{(p+q)^2(p+q+1)} \tag{11}$$

Por su recorrido esta distribución es muy usada para modelar proporciones. Además toma diferentes formas de acuerdo al valor que toman ambos parámetros lo que permite que se adapte a los datos. La distribución será estrictamente creciente si p>1, q=1; estrictamente decreciente si p=1 y q>1; con forma de "U" si p y q son menores a 1; o unimodal si ambos son mayores a 1.

GRÁFICO 1: Distintas formas de la distribución Beta.



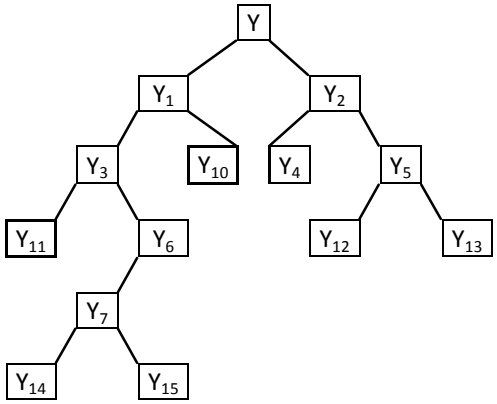
2.5. ÁRBOLES DE REGRESIÓN Y DE CLASIFICACIÓN

La técnica basada en estructuras de árbol es de carácter exploratorio no paramétrico y su fin es identificar reglas de clasificación y predicción sobre un conjunto de observaciones. Los datos constan de un conjunto de variables explicativas y una variable a explicar, si la variable de respuesta es categórica el árbol será de clasificación mientras que si es continua el árbol será de regresión. Las variables explicativas pueden ser tanto categóricas como numéricas.

En función del conjunto de variables predictoras el grupo conformado por los individuos objeto de estudio es particionado en subgrupos disjuntos con el fin de obtener una herramienta de clasificación(o predicción). La técnica permite además identificar un conjunto de variables a partir de las cuales es posible interpretar la estructura obtenida, como por ejemplo qué variables determinan la pertenencia de un individuo a una clase o qué variables hacen que la variable de respuesta tenga determinado valor.

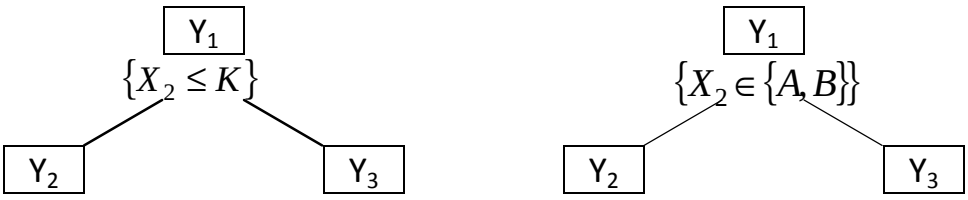
De acuerdo al tipo de partición los árboles pueden clasificarse en binarios o N-arios. En los árboles binarios la construcción en cada paso se obtiene por la partición sucesiva del conjunto en dos subconjuntos disjuntos. El primer elemento particionado se denomina nodo raíz y en cada partición el árbol se ramifica generándose nodos intermedios. Aquellos nodos que no son particionados se denominan nodos terminales. En los N-arios la partición en cada paso da lugar a N subconjuntos disjuntos.

En el caso presentado en la figura, el nodo raíz, Y, es particionado tal que $Y = Y_1 \bigcup Y_2$, y en forma análoga los nodos intermedios Y_1 y Y_2 se particionan tal que $Y_2 = Y_4 \bigcup Y_5$ y $Y_1 = Y_3 \bigcup Y_{10}$. Los nodos identificados con $Y_{1j,\, j=0,1,..,5}$ son terminales.



Los árboles de clasificación y regresión binarios en los que se centra el análisis de aquí en más son conocidos como árboles de tipo CART[4]. La estructura generada en los datos a partir de la técnica es del tipo jerárquica donde cada nodo, comenzando por el nodo raíz, es particionado en subconjuntos disjuntos. Cada partición tiene asociada una decisión con dos alternativas posibles basada en la información que provee una única variable.

En el caso de que la variable asociada al nodo es continua se compara el valor de cada individuo con un valor determinado, si el registro del individuo es menor sigue el camino de la izquierda, sino el de la derecha. Por ejemplo, suponga que X_2 es una variable numérica seleccionada por la técnica para particionar Y_1. Si $\{x; x_2 \leq K\}$ entonces $Y_2 = \{y; x_2 \leq K\}$ y si $\{x; x_2 > K\}$ entonces $Y3 = \{y; x_2 > K\}$. En cambio se la variable es categórica entonces lo que se evalúa en el nodo es a que categoría se asocia el individuo. Si las categorías de X_2 son A, B, C, D y E entonces $Y_2 = \{y; x_2 \in \{A, B\}\}$ y $Y_3 = \{y; x_2 \in \{C, D, E\}\}$



Junto con las reglas de decisión en cada nodo deben determinarse reglas para fijar el tamaño del árbol, la bondad de ajuste y el criterio de decisión que determine la categoría, o el valor que debe asociarse a cada individuo. La elección del conjunto de

[4] CART: Classification and regression trees

reglas se asocia al objetivo de obtener conjuntos homogéneos en cada nodo, buscando una buena representación del conjunto estudiado.

Una vez obtenido el árbol a través un proceso de aprendizaje puede ser utilizado como una herramienta de clasificación o predicción. Cada observación "recorre" el árbol comenzando por el nodo raíz y preguntando en cada nodo intermedio, de acuerdo a la respuesta el camino que sigue. La observación es clasificada o predicha cuando llega a un nodo Terminal.

2.5.1. ÁRBOLES DE REGRESIÓN

En los árboles de regresión, dado un conjunto de variables predictoras X y una variable de respuesta numérica, se define a la función de predicción como d(X). La construcción del árbol de regresión permite entender la relación existente entre la variable independiente y las predictoras y es además una herramienta que permite predecir el valor de y para nuevas observaciones. El conjunto de datos se particiona en cada nodo en dos subgrupos disjuntos hasta obtener los nodos terminales a los que se les asocia un único valor de predicción, y (t).

Específicamente el valor y (t) que se asocia al nodo t es aquel que minimiza el error cuadrático medio $R^*(d)$, calculado como $R^*(d) = E(Y - d(Y))^2$.
La estimación de $R^*(d)$ se define:

$$R(d) = \frac{1}{N}\sum_n (y_n - d(x_n))^2 \qquad (12)$$

Se prueba que el valor de y (t) que minimiza el error es el promedio de todos los y_n tal que x_n pertenece al nodo t, con N (t) el número de casos en el nodo:

$$\bar{y}(t) = \frac{1}{N(t)}\sum_{x_n \in t} y_n \ . \qquad (13)$$

$R\ (d)$ es una medida del nodo terminal t por lo que una medida global del árbol se obtiene promediando la precisión en cada uno de los nodos terminales:
$R(T) = \frac{1}{N}\sum_{t\in \tilde{T}}\sum_{x_n \in t}(y_n - \bar{y}(t))^2$, donde $\tilde{T}$ denota al conjunto conformado por los nodos terminales.

El árbol de regresión es formado iterativamente particionando cada nodo de tal forma de maximizar la reducción del error. La mejor partición s* de un nodo t es aquella s del conjunto de particiones S que logra la mayor reducción de R (T), es decir aquella que logra la mayor reducción del error. Más precisamente s* es $\max_{s\in S} \Delta R(s,T) = \max_{s\in S} \Delta(R(t) - R(t_D) - R(t_I))$,con t_D y t_I las particiones derecha e izquierda del nodo t respectivamente.

Uno de los problemas en la construcción de un árbol es determinar el número de particiones sobre el conjunto original. El objetivo es seleccionar aquella partición que genera un conjunto más "puro" que el anterior, por lo que un criterio para dejar de particionar al conjunto considera el decrecimiento de la impureza en el nodo: si la variación es menor que un valor predefinido entonces no se particiona. En el caso extremo, para los árboles de clasificación, un nodo conformado por individuos de una única clase es totalmente puro.

2.5.2. ÁRBOLES DE CLASIFICACIÓN

En los árboles de clasificación el mecanismo de participación es análogo. La diferencia esencialmente radica en los valores que asocian a los nodos terminales y en los indicadores que se consideran para evaluar el poder del modelo. Mientras que en los árboles de regresión el nodo terminal se asocia a un valor real calculado con una función de predicción, en los de clasificación se asocia a una etiqueta o categoría de la variable de estudio. De esta forma pueden existir más de un nodo con la misma etiqueta.

Las particiones en cada uno de los pasos, comenzando desde el conjunto global de datos, buscan aumentar la pureza de los nodos resultantes respecto al original. En este tipo de árboles la pureza del nodo se asocia a la proporción que hay en el nodo de cada clase. Si por ejemplo si una variable tiene las categorías A,B,C ,D, y en un nodo todas las observaciones son de la clase A la pureza es total y ésa clase es la que etiqueta al nodo. Si en cambio las clases están representadas en una misma proporción, ¼ en el ejemplo, es necesario buscar particiones que permitan que una clase predomine sobre las otras de tal manera de asociarla al nodo.

Formalmente si $p(j_0/t)=\max_j p(j/t)$ entonces la etiqueta que se asigna al nodo es j_0

2.6. MEDIDAS DE PRECISIÓN EN ESTIMACIÓN: TEST SAMPLE ESTIMATION Y CROSS VALIDATION

La precisión de la predicción del modelo puede evaluarse en diferentes conjuntos de datos. El método de Test Sample estimation considera una partición del conjunto de datos X en dos subconjuntos disjuntos, X_1 y X_2. Solamente los casos en X_1 se consideran para generar el modelo predictivo, y los casos de X_2 se utilizan para evaluar el poder predictivo. Si N_2 es el número de casos en X_2

$$R^{TS}(d)=\frac{1}{N_2}\sum_{y_n\in X_2}(y_n-d(x_n))^2 \qquad (14)$$

Comúnmente se toma 1/3 de los casos de la base para evaluar el poder predictivo y el modelo se genera con 2/3 de los casos. Los casos en ambos subgrupos se consideran independientes y provenientes de la misma distribución. Cuando el tamaño de la muestra es grande el método es eficiente.

Otro método aplicado para muestras pequeñas denominad “V-fols cross validation”, genera aleatoriamente subconjuntos de tamaños similar. Se denotan los subconjuntos como X_1, X_2……, X_v.
Para cada subconjunto se aplica el procedimiento usando todos los datos excepto los del subconjunto:X- X_j y se calcula el poder predictivo en los datos excluídos , es decir sobre X_j. El estimador de la precisión para el subconjunto, dado que se aplica test simple al subconjunto es:

$$R^{TS}(d^j) = \frac{1}{N_j} \sum_{y_n \in X_j} (y_n - d(x_n))^2 , \quad (15)$$

con N_j la cantidad de casos en X_j. Una medida global de la precisión está dada por un promedio de la medida de precisión obtenida para cada subconjunto:

$$R^{CV}(d) = \frac{1}{V} \sum_{j=1}^{V} R^{TS}(d^j) \quad (16)$$

2.7. VARIABLES CONSIDERADAS EN EL ESTUDIO

Las variables incluidas en el análisis consideran características sociodemográficos de los afiliados y características vinculadas a la regularidad de aportación y a la afiliación en RAFAP. Entre las sociodemográficas se incluyen edad, sexo, rama de actividad, nivel salarial, y en las segundas densidad de cotización, monto de aportes, antigüedad y origen de la afiliación.

Por Antigüedad se denota el tiempo transcurrido entre el momento de la afiliación a RAFAP y un período de tiempo posterior, en nuestro caso el momento más reciente.
La densidad de cotización es un indicador que se obtiene como la cantidad de aportes sobre la cantidad de meses de un período determinado.
La franja salarial designa el rango dentro del cual se ubica el salario por el que aporta el individuo, en pesos uruguayos.
El giro de actividad es precisamente la rama de actividad en la que se desempeña el individuo. Las categorías en las que se desagrega son Civil, Construcción, Doméstico, Industria y comercio y Rural.
Por último el origen de afiliación refiere a la forma mediante la cuál el individuo se afilia a RAFAP. Las categorías para esta variable son:

- Voluntario: Eligen la AFAP a la que quieren ser afiliados y permanecen en la misma.
- Oficio: La AFAP es asignada por BPS. Cuando se supera el tope salarial de $19.805 es obligatoria la afiliación a una AFAP, en este caso el individuo puede elegir su administradora y ser afiliado voluntario o el BPS le asigna una de oficio.
- Traspaso: Afiliados a otra AFAP anteriormente
- Reingreso de traspaso: refiere a individuos cuya afiliación a RAFAP fue interrumpida durante un período en el cual el individuo se traspasa a otra administradora.

En el Anexo 2 se detallan las categorías en las que se clasifican las distintas variables.

Presentadas las variables, en el capítulo a continuación se presenta el análisis exploratorio y las conclusiones obtenidas.

CAPITULO 3 ANÁLISIS EXPLORATORIO

En el análisis descriptivo se considera un enfoque bivariado, donde la variable de referencia es la densidad de cotización, y uno multivariado en el que se contempla la interacción de distintos factores, como giro de actividad, edad del afiliado entre otros.
Se describe el comportamiento del monto del aporte y de la densidad de cotización en una muestra de afiliados a RAFAP. Para seleccionarla se aplica muestreo sistemático fijando el tamaño[5] de muestra en 3000, lo que asegura precisión del 2% en la estimación de la media de la densidad de cotización para un nivel de confianza del 95%. Si bien el objeto del trabajo no es realizar estimaciones de parámetros poblacionales, la selección por medio de muestreo probabilístico proporciona un mecanismo para obtener una muestra representativa.
Ordenando el conjunto de datos con información de interés el método de muestreo aplicado asegura una cobertura de unidades de todos los tipos. En nuestro caso la muestra está integrada por individuos que tienen diferentes magnitudes o categorías de las variables giro, franja y sexo.

Para determinar la base de datos con la que se trabajará se impone además la condición de afiliación previa a enero 2000. Este es un plazo prudente para captar el comportamiento de aporte y asegurar la calidad de la información.

3.1. MONTO DEL APORTE

La evolución mensual del monto del aporte en el período 2000-2005, aunque registra un descenso en los años 2002 y 2003, tiene saldo favorable. Es posible detectar el efecto de la crisis económica en los distintos indicadores, siendo menos evidente en los de desvío.

CUADRO 1: Evolución del monto del aporte entre los años 2000 y 2005

Año	Mínimo	Cuartil 1	Mediana	Media	Cuartil 3	Máximo	Desvío
Marzo 2000	0	60	351	498	735	7300	621
Marzo 2001	0	0	353	522	775	5652	663
Marzo 2002	0	0	328	504	772	7252	666
Marzo 2003	0	0	298	486	764	4811	661
Marzo 2004	0	0	316	545	844	5264	746
Marzo 2005	0	0	365	617	939	6574	845

[5] La fórmula para obtener el tamaño de muestra es :

$$n = \frac{NZ^2_{0.975} S^2_y}{(N-1)e^2 + Z^2_{0.975} S^2_y}$$

Donde e es la precisión en la estimación, S^2_y es la varianza de la variable de interés, N es el tamaño de la población y $Z^2_{0.975}$ es el valor asociado a la distribución normal que acumula una probabilidad de 0.975.

Para determinar las causas de la disminución del monto de aporte se analiza la evolución de indicadores macroeconómicos en el período 2000 – 2005. Dado que la disminución puede ser resultado de la caída de salarios o del número de cotizantes, en particular se consideran índice medio de salarios y tasa de desempleo[6].

El efecto de la crisis se aprecia en las tasas de empleo y desempleo a partir de una caída de la actividad laboral entre 2001 y 2003. No obstante la crisis no afecta negativamente el nivel de ingreso ya que el índice medio de salarios es creciente en todo el período.

Al comparar las medidas resumen del monto de aporte para la cartera completa de afiliados y para el conjunto de cotizantes[7], se obtiene que para el primer grupo se registra una caída en 2002 y 2003 mientras que entre los cotizantes crece durante todo el período. Los resultados, consistentes con los indicadores de mercado, permiten concluir que la caída del monto de aportes se debe al aumento del desempleo y no al descenso de los salarios.

CUADRO 2: Evolución de la tasa de empleo y desempleo y del índice medio de salarios en Uruguay para el período 2000-2005

AÑO	Tasa de empleo	Tasa de desempleo	Índice medio de salarios[8]
2000	51,5	13,6	94,06
2001	51,4	15,3	97,88
2002	49,1	17	98,88
2003	48,3	16,9	104,06
2004	50,8	13,1	113,53
2005	51,4	12,2	124,31

CUADRO 3: Evolución del monto del aporte de los afiliados cotizantes entre los años 2000 y 2005

Año	Cuartil 1	Mediana	Media	Cuartil 3	Desvío
Marzo 2000	161	497	648	756	636
Marzo 2001	160	539	699	7779	682
Marzo 2002	158	558	704	807	692
Marzo 2003	155	582	721	818	692
Marzo 2004	161	624	802	871	783
Marzo 2005	171	649	872	939	888

La variable monto de aporte varía entre los $0 y $16.504 en el período, el valor promedio es $765 y desvío de $934. Se identifican como observaciones atípicas[9] aquellas que superan los 10.000 pesos por lo que se decide excluirlas de la muestra.

[6] Información extraída de www.ine.gub.uy

[7] Los cotizantes son los afiliados que realizan aportes en el mes de referencia.

[8] Base Diciembre de 2002

[9] Se considera valor atípico a aquellos que son numéricamente distintos del resto de los datos. Tomando como referencia el primer y tercer cuartil se considera valor atípico aquél que se aparta por lo menos 1.5 veces el rango intercuartílico (IQ) de éstos valores. Si Q_1 y Q_3 son respectivamente el primer y tercer cuartil y el rango intercuartílico es $(Q_3 - Q_1)$, un valor atípico es aquel menor que Q_1-1,5 $(Q_3 - Q_1)$ o mayor que Q_3+1,5. $(Q_3 - Q_1)$.

No se consideran otras opciones para el tratamiento de esos datos dado que la cantidad de casos era solo 0.7 % del total. Para este nuevo conjunto de datos la media del aporte asciende a $617 con un desvío estándar de $847 a marzo de 2005.

En cuanto a la relación del monto del aporte con otros variables se obtienen las siguientes conclusiones. La cartera de afiliados a RAFAP está compuesta en un 43% por mujeres y en un 57% por hombres. El comportamiento del aporte y su evolución no dependen del género, sin embargo los máximos se registran en los afiliados del sexo masculino.

La mitad de los afiliados tienen entre 33 y 44 años de edad, solo un 10% es menor a 29 años y otro 10% es mayor a 48. La correlación lineal entre el monto del aporte y la edad de los afiliados es positiva, lo que se vincula al reconocimiento de la experiencia en el puesto de trabajo, que se manifiesta en una mayor remuneración. En el período estudiado el aporte de los afiliados menores de 47 años crece mientras que en el resto es más notorio el efecto de la crisis económica.

El origen de la afiliación admite las categorías voluntario, oficio, traspaso y reingreso de traspaso[10]. En la cartera de RAFAP a enero de 2000 predominan los afiliados en forma voluntaria, alcanzando el 70% del total. El comportamiento de los aportes es similar para todos los orígenes y su evolución es creciente en todo el período.
Los indicadores de aporte del grupo de afiliados de oficio, como media, mediana, desvío estándar, son superiores al resto de las categorías dado que deben superar un tope salarial al momento de afiliarse. Resulta intuitivo inferir que este grupo, por su mecanismo de afiliación, es más homogéneo respecto al monto del aporte que los demás. Sin embargo la dispersión es semejante resultado que se debe a las condiciones impuestas por el Artículo 8 de la Ley 16.713[11]. Entre los afiliados de oficio hay un mayor porcentaje respecto al resto de las categorías que no opta por el artículo 8, lo que implica menor frecuencia de aportación y más dispersión de los montos. La explicación de ese mayor porcentaje es que los afiliados de oficio en general tienen menor interés en aportar a una administradora que el resto que se afilia en forma voluntaria.
En cuanto a los afiliados por reingreso de traspaso se generan dificultades para reconstruir la historia de densidad de cotización, motivo por el cual se excluyen de la muestra. La forma en que se registra la información hace que se conozcan todos los aportes realizados a RAFAP desde la primera afiliación pero la antigüedad sólo se contabiliza desde el último ingreso a la administradora.

La variable giro indica la rama de actividad por la que aporta el afiliado y la conforman las categorías industria y comercio, civil, rural, construcción y doméstico. Predominan industria y comercio y civil, representando en conjunto el 90% de la cartera, 60% y 30%

[10] Los afiliados de origen voluntario son aquellos que eligen su AFAP mientras que a los afiliados de oficio se la asigna BPS. El origen traspaso refiere a aquellos individuos afiliados a otras AFAP anteriormente. La categoría reingreso de traspaso refiere a individuos cuya afiliación a RAFAP fue interrumpida durante un período en el cual el individuo se traspasa a otra administradora.

[11] Quienes optan por éste vuelcan sus aportes a la AFAP cada vez que cotizan, mientras que los que no lo hacen aportan sólo cuando superan un tope exigido.

respectivamente. El bajo porcentaje de afiliados que integran los giros doméstico y construcción se explica por la distribución de los giros en el mercado y su alto grado de informalidad. Según un estudio realizado por el INE[12] en el conjunto de empleados informales se destacan las actividades construcción y doméstico por su alto nivel de no registro, alcanzando el 60%. Los giros industria y comercio y civil, además de ser los que predominan en la cartera, registran niveles superiores de aporte. En contraposición el giro doméstico presenta los valores más bajos, producto del nivel salarial del grupo. La evolución de los aportes entre 2000 y 2005 es diferente entre los giros. En industria y comercio y civil la evolución es creciente, mientras que en el doméstico es decreciente. En los giros rural y construcción existe una caída del aporte entre 2000 y 2003, producto de la crisis económica.

3.2. DENSIDAD DE COTIZACIÓN

La densidad de cotización es un indicador de la frecuencia de aportación de los afiliados, definido como la cantidad de aportes en meses que realizan los individuos en un intervalo de tiempo determinado también medido en meses.

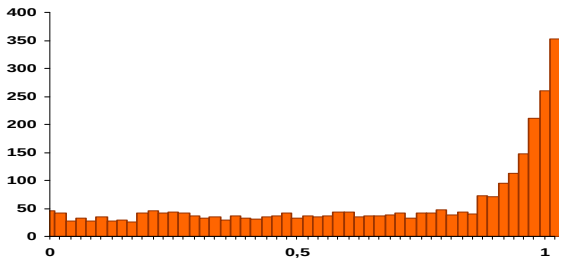
$$Densidad\ de\ Cotización = \frac{Cantidad\ de\ Aportes\ a\ la\ Cuenta\ entre\ t\ y\ t+p}{Cantidad\ de\ meses\ transcurridos\ entre\ t\ y\ t+p} \qquad (17)$$

En el estudio de esta variable se considera la interacción con factores como el sexo, el giro de actividad, el origen de la afiliación, la antigüedad en RAFAP, el salario y la edad del afiliado.

Con el propósito de identificar la relación intertemporal de los datos, se considera la evolución anual de estas variables en el período 2000 – 2007. Dado que la información es una serie mensual y la densidad de cotización es una variable acumulativa, se toma como medida resumen de cada año el valor a diciembre.

En el conjunto de afiliados a RAFAP la densidad de cotización es asimétrica y bimodal en 0 y 1. La distribución se concentra en el rango 0.9-1 lo que muestra una alta regularidad de aportación.

GRÁFICO 2: Histograma para densidad de cotización



[12] "Empleo informal en Uruguay"- 2007 basado en datos de la Encuesta Nacional de Hogares

CUADRO 4: Evolución de la densidad de cotización entre los años 2000 y 2005

Año	Cuartil 1	Mediana	Media	Cuartil 3	Desvío
2000	0.50	1.00	0.75	1.00	0.39
2001	0.50	1.00	0.74	1.00	0.38
2002	0.47	1.00	0.73	1.00	0.37
2003	0.44	0.98	0.71	1.00	0.37
2004	0.40	0.97	0.71	1.00	0.37
2005	0.40	0.96	0.71	1.00	0.36

El giro de actividad condiciona la densidad de cotización, siendo las categorías con mayor regularidad de aportación civil e industria y comercio, su densidad de cotización promedio es de 0.79 y 0.74 respectivamente. El resto de los giros son más irregulares con valores promedio que no supera el 0.70, resultado consistente con el nivel de informalidad y el carácter zafral de las actividades laborales con las que se asocian.

Giro de Actividad	Media	Mediana
Industria y Comercio	74.32	91
Civil	79.65	98
Aporte Rural	68.14	86
Construcción	45.29	41
Servicio Doméstico	60.34	59

La densidad de cotización es mayor para el sexo femenino, la mediana y la media son de 0.88 y 0.72, mientras que para el sexo masculino es 0.85 y 0.71.
La diferencia entre géneros puede inducir a la conclusión anticipada de mayor estabilidad laboral de la mujer en el mercado. Al respecto son válidas dos apreciaciones, no se pueden extender los resultados de la cartera de afiliados al mercado laboral y el comportamiento observado se debe al alto porcentaje de giro civil en el conjunto de mujeres.
Entre los aportes superiores sin embargo predominan los del género masculino.

CUADRO 5: Distribución de giro según sexo

	SEXO		
GIRO	F	M	Total
Industria y Comercio	60.15%	61.88%	61.12%
Civil	34.91%	24.41%	29.03%
Aporte rural	2.50%	6.07%	4.50%
Construcción	0.06%	7.04%	3.97%
Servicio doméstico	1.89%	0.10%	0.89%
Otros	0.49%	0.50%	0.50%
Total	100%	100%	100%

La relación entre densidad de cotización y edad es positiva, con una correlación de 0.15. Si bien este valor considerado en forma aislada no indica una dependencia importante, en el conjunto de datos la edad es uno de los factores más vinculado a la densidad de cotización. El sentido de la relación coincide con el comportamiento esperado a priori, los jóvenes presentan mayor inestabilidad en el mercado laboral (participación inestable, ejemplo contratos a término, pasantías) lo que repercute en la regularidad de aportación.

CUADRO 6: Densidad de cotización media por nivel de franja salarial

Franja etaria	Media
18 a 25	61.47
26 a 35	72.30
36 a 45	73.52
46 a 55	73.19
56 y más	73.21

Respecto a la relación con el origen, los afiliados de oficio presentan menor frecuencia de aportación y mayor heterogeneidad, lo que resulta particular dada la composición del grupo. La explicación reside en el porcentaje de afiliados que optan por el artículo 8, cuyo efecto fue explicado en el análisis de monto de aporte.

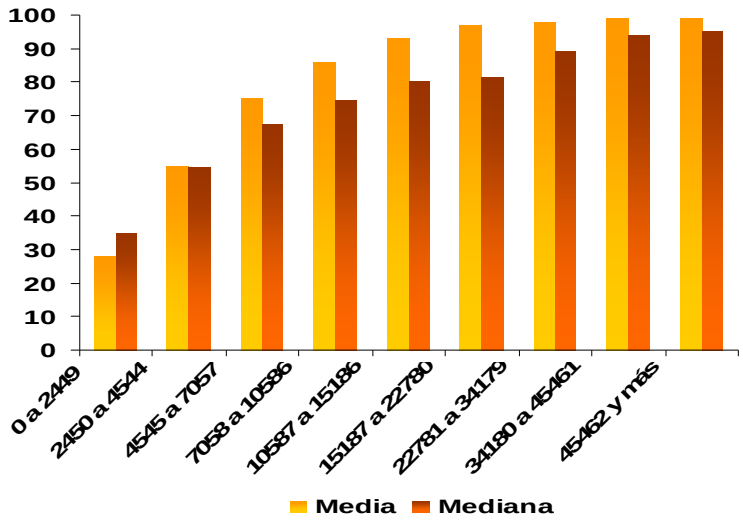
CUADRO 7: Densidad de cotización media por categorías de origen de afiliación

	Media
VOLUNTARIO	71.87
DE OFICIO	58.71
TRASPASO	76.12
REINGRESO DE TRASPASO	86.95

En la cartera la franja de renta, que representa el rango de salario por el que aporta el afiliado, predominan las franjas medias y bajas. Esta distribución no implica que RAFAP concentre individuos de bajos ingresos, por el contrario, los fondos indican que los niveles generales de aporte de sus afiliados son mayores que en el resto de las administradoras.

El salario es uno de los factores con mayor incidencia en la frecuencia de aportación según lo indica su correlación de 0.4. Vale señalar que este valor es alto si se compara con los resultados obtenidos en las demás variables. En resumen la dependencia entre salario y densidad de cotización es positiva, lo cual es consistente con el preconcepto de mayor estabilidad laboral para aquellos afiliados con mayores ingresos.

GRÁFICO 3: Media y Mediana de la densidad de cotización por franja de salario



Para complementar el análisis se estudia la interacción entre factores, por ejemplo, salario, edad y giro. El salario presenta crecimiento hasta los 45 años de edad en el giro civil, que se prolonga hasta los 60 en industria y comercio. En este último el crecimiento es más pronunciado. Para los giros: rural, construcción y doméstico existe cierta estabilidad del salario a lo largo del ciclo de vida, con una leve tendencia de crecimiento en el doméstico y decrecimiento en el rural. Para construcción no se distingue tendencia definida.
La relación entre densidad de cotización y salario para cada giro indica que la remuneración tiene mayor incidencia en la frecuencia de aportación de los trabajadores de los giros civil e industria y comercio.

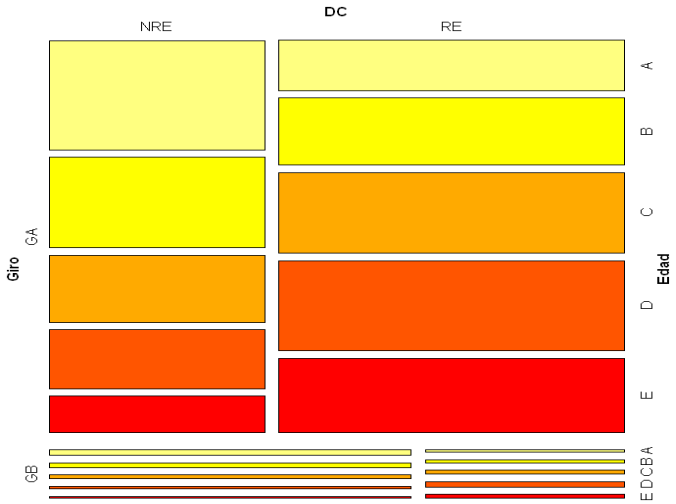
Respecto a la evolución de la frecuencia de aportación en el ciclo de vida según la actividad, se encuentran similitudes entre civil e industria y comercio. Ambos presentan tendencia de crecimiento de la densidad de cotización hasta los 40 años a partir de donde se estabiliza en valores cercanos a 1, lo que se vincula a la solidez que logran estos individuos en el mercado laboral. Las categorías construcción y doméstico presentan una leve tendencia de crecimiento que conserva durante todo el ciclo de vida. El giro rural presenta tendencia creciente más pronunciada que en el resto de los giros.

Los gráficos de mosaico permiten visualizar en forma sencilla las frecuencias asociadas a una tabla de contingencia. El ancho y el alto de cada diagrama muestran la frecuencia relativa de las variables: el tamaño de cada cuadrante es proporcional a la cantidad de observaciones en las distintas celdas de una tabla de frecuencia.
Se incluye un gráfico de mosaico con distribución de frecuencias por edad, giro y densidad de cotización que resume los resultados antes mencionados. Se define la categoría regular como Densidad de Cotización mayor a 0.8; se observa que con ese

valor o menos no se configura causal jubilatoria con los mínimos requisitos (35 años de aporte y 60 años de edad), ni siquiera comenzando a aportar a los 18 años.

En la figura se observa que las categorías civil e industria y comercio tienen aportación más regular que el resto (los rectángulos asociados a GA son más largos que los asociados a GB) y la regularidad aumenta con la edad sin distinción de giro(los rectángulos asociados son más altos a medida que la edad aumenta).

GRÁFICO 4: Mosaico para la distribución de frecuencias por edad, giro y densidad de cotización



Referencias

DC (densidad de cotización)
NRE (No Regulares) __ DC<80
RE (Regulares) __ DC>=80
Giro
GA (giro A) __ civil e industria y comercio
GB (giro B) __ rural, construcción, doméstico y Otros

Edad
A __ Hasta 31 años
B __ de 31 a 36 años
C __ de 37 a 41 años
D __ de 42 a 46 años
E __ más de 46 años

Como medida numérica complementaria al gráfico se utilizan los odds ratio. Los odds se obtienen como el cociente entre la probabilidad de éxito y la probabilidad de fracaso de un suceso determinado. Si π es la probabilidad de éxito entonces el odds es de la forma: $\pi/(1-\pi)$. Los odds son no negativos y su valor es mayor que uno si el éxito tiene más probabilidad que el fracaso, por ejemplo si el odds es 4 entonces esperamos ver 4 éxitos por cada fracaso.

El odds ratio es el cociente entre dos odds que refieren a sucesos distintos, llamémosle suceso 1 y 2: $\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$. Si el ratio es mayor que uno entonces el odds del éxito es mayor en el suceso uno que en el dos, por lo tanto en el suceso 1 es más plausible el éxito que en el 2.
En este caso el odds ratio representa la chance de no ser regular respecto a serlo, entendiendo por regular aquel comportamiento de aporte con densidad de cotización mayor o igual a 0.80. Los odds fueron calculados para los distintos rangos de edad, condicionado por giro, considerando el suceso "ser regular" como éxito. Se agrupan los giros en las categorías A y B y se compara el odds de regularidad dado que pertenece al giro B con el odds del giro A.
En el cuadro se presentan los odds del suceso "el afiliado es regular" en el giro de industria y comercio y civil respecto al odd del mismo suceso para el resto de los giros, distinguiendo según tramo de edad.

CUADRO 8: Valores de los Log Odds Ratio y sus intervalos de confianza para las dimensiones de tabla correspondientes a cada rango de edad

EDAD	LOG ODDS RATIO	ERROR ESTAND0AR	VALOR z	PR(>|z|)	SIGNIFICACIÓN	ODDS RATIO
Menos de 31	-1,1470	0,306	-3,75	$8{,}88 \times 10^{-5}$	***	0,3176
de 31 a 36	-1,1451	0,280	-4,08	$2{,}22 \times 10^{-5}$	***	0,3182
de 37 a 41	-1,2557	0,273	-4,60	$2{,}11 \times 10^{-6}$	***	0,2849
de 42 a 46	-0,7966	0,275	-2,90	$1{,}89 \times 10^{-3}$	**	0,4508
más de 46	-1,0961	0,308	-3,56	$1{,}87 \times 10^{-3}$	***	0,3342

Códigos de significación: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

De acuerdo al odds ratio si el afiliado se desempeña en el giro civil o industria y comercio tiene "más chance" de ser regular. En el tramo de edad 37 a 41 años se profundizan las diferencias de regularidad de aportación de los giros A y B a favor del A.

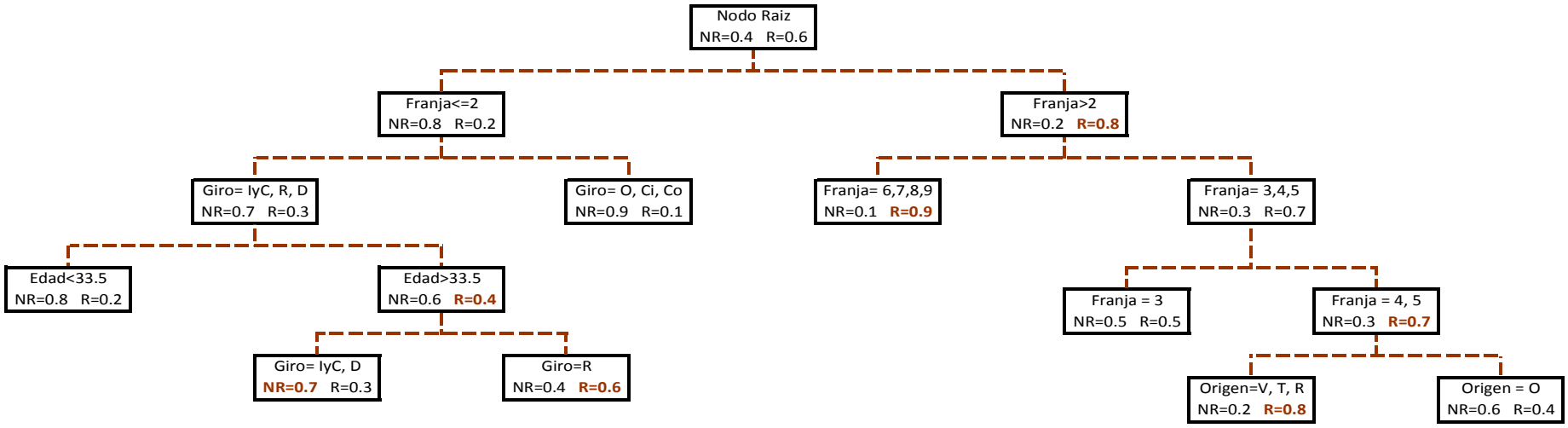
Para continuar con el estudio de interacción entre variables se realiza otro análisis multivariado a partir de árboles de clasificación. La variable de respuesta fue construida categorizando la variable densidad de cotización de 2005. Los valores menores a 0.8 constituyen la categoría de no regulares y los mayores o iguales a 0.8 la categoría regulares.

Las variables explicativas son sexo, giro de actividad, origen de la afiliación, franja de Renta, edad, antigüedad. Para las dinámicas se toma el valor registrado para diciembre del 2005. La variable con mayor poder explicativo es franja de renta que presenta una relación positiva con la proporción de cotizaciones. Otros factores que se asocian positivamente a la regularidad de aportación son edad y antigüedad en RAFAP. Respecto al origen de afiliación, los de oficio presentan un menor porcentaje de afiliados regulares, resultado coherente con lo obtenido en el análisis bivariado.

Uno de los resultados del análisis previo indica que las variables giro de actividad y densidad de cotización están asociadas. Sin embargo en el árbol de clasificación giro aparece con menor poder explicativo que franja de renta y comportamiento poco consistente con los resultados obtenidos previamente. La explicación de este último es la alta dependencia entre el giro de actividad y la franja de renta.[13] Las inconsistencias del comportamiento del giro se anulan si la variable franja se excluye del árbol.

[13] El resultado de la prueba de independencia chi cuadrado para giro y franja es p-valor 0 y un estadístico de 2285.

GRÁFICO 5: Árbol de clasificación para la variable densidad de cotización a 12/2005 categorizada (NR= No Regular, R= Regular)



Nota: Los valores que aparecen en los nodos representan el porcentaje de individuos en cada categoría (NR, R) para cada nodo.
El árbol consta de 11 nodos terminales, la deviance media de los residuos es 0.94. El porcentaje de error de clasificación es de 23%

La deviance es una medida de impureza aplicada a un nodo t, definida como $D(t) = -2\sum_{j=1}^{J} n_{ij}\ p(j/t)\log(p(j/t))$, con nj la cantidad reobservaciones de la categoría j, p (j /t) es la probabilidad de ser de la categoría j dado que está en el nodo t. j, J=cantidad de categorías.

Referencias:

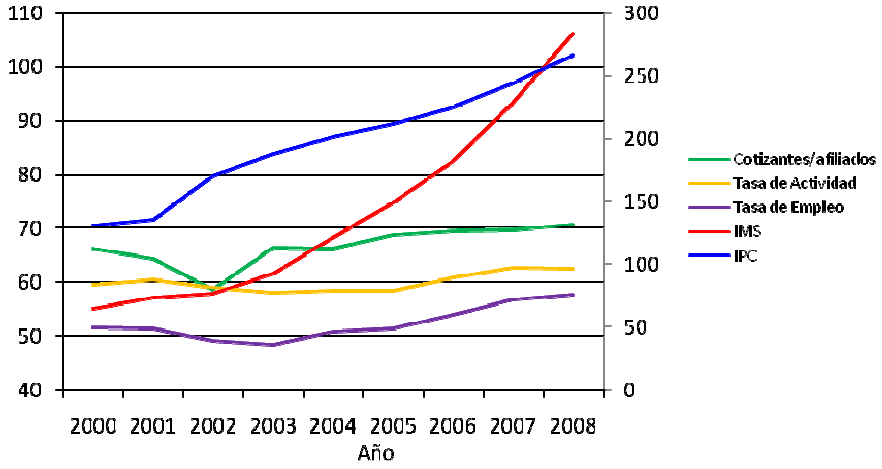
Giro:
Inc.: industria y comercio
D: doméstico
R: rural
Ci: civil
Co: construcción
O: Otros

Origen:
V: Voluntario
R: Reingreso de traspaso
T: Traspaso
O: Oficio

Franja
0: Sin información
1: 0 – 2449
2: 2450 – 4544
3: 4545 – 7057
4: 7058 – 10586
5: 10587 – 15186
6: 15187 – 22780
7: 22781 - 34179
8: 34180 - 45461
9: Más

Hasta este punto se analizó la relación entre la densidad de cotización y las características de los afiliados. No obstante existe dependencia entre el comportamiento de aportación y la realidad económica de la región que debe ser contemplada. Para representar el efecto del contexto económico en el fenómeno de aportación se comparan indicadores a nivel país de empleo, salario y precios con porcentaje de cotizantes en la cartera de RAFAP.

GRÁFICO 6: Evolución de indicadores período 2000-2008



El gráfico permite visualizar un comportamiento alineado de las variables económicas y la proporción de cotizantes. En todos los casos se evidencia el impacto de la crisis de 2002 y el crecimiento sostenido a partir de 2003.

Si bien se constata que el contexto económico influye en la aportación, en este estudio se contemplan exclusivamente las características del afiliado. Incorporar variables de tipo económico en los modelos utilizados implicaría realizar predicciones de las mismas y su complejidad excede a esta investigación. Más adelante se muestra cómo el efecto del contexto se concreta en el corto plazo pero se diluye en el largo. En la metodología aplicada se asume un escenario de estabilidad económica.

CAPITULO 4: METODOLOGÍA

El objetivo central en esta etapa del trabajo es identificar un modelo probabilístico de corto plazo capaz de predecir la densidad de cotización p años posteriores al último dato disponible, con p un número entero entre 1 y 4. El modelo se basa en la información disponible al momento t, tanto de la densidad de cotización como de otras características de los afiliados, para explicar valores de la densidad de cotización al momento (t+p). Como resultado de la aplicación se obtendrán estimaciones puntuales de la variable en (t+p) y además una distribución de probabilidad de la densidad de cotización futura a corto plazo asociada a cada afiliado.[14]

$$DC^{(t+p)} = g(X^t)$$
$$p \in [1,4],\ p \in N$$
$$DC^{(t+p)} \sim F(\mu, \sigma)$$

No se incluyen en forma explícita variables económicas por lo que para su aplicación y validez se asume un contexto económico estable. Si bien no se descarta que los factores económicos afecten la densidad de cotización, la complejidad de considerar variables macroeconómicas excede al alcance de esta investigación.
La incorporación de los antecedentes de aportación introduce un componente dinámico, que se traduce en modelos generados para la densidad de cotización en diferentes momentos y rezagos del tiempo. El análisis descriptivo constituye la principal herramienta para identificar las variables explicativas de los modelos y determinar la metodología apropiada para estimar la densidad de cotización futura a corto plazo.

4.1. MODELO MATEMÁTICO

En el proceso de investigación se evalúan distintas opciones de metodología para explicar y predecir la densidad de cotización futura como por ejemplo modelos de conteo y modelos para proporciones. Dichos modelos son aplicados por algunos autores consultados, como Lagomarsino, Bertranou y Sánchez. [3][13]
La variable cantidad de aportes no negativa y discreta se adapta a los modelos de conteo. El modelo Poisson, por ejemplo captura la naturaleza de los datos y permite hacer inferencia sobre la probabilidad de que ocurran k aportes. Una limitación de estos modelos está en el supuesto de que los eventos ocurren en forma independiente en el tiempo, lo que no se ajusta al comportamiento de aporte de los individuos. Si una persona que en el mes m tiene trabajo y aporta, tendrá más probabilidad de aportar en el mes m+1 que una que estaba desocupada en el mes m.

[14] En este documento los términos densidad de cotización en el momento t y densidad actual son equivalentes, al igual que densidad de cotización en el momento (t+ p) con densidad de cotización futura.

El modelo Binomial Negativo que también se adapta a la variable cantidad de aportes, es menos restrictivo que el anterior dado que permite introducir un componente estocástico que capta la heterogeneidad y los errores de medida por componentes no observables. La limitación nuevamente deriva del supuesto de que los sucesos son independientes entre sí.
Esos supuestos de independencia pueden ser evitados con otras técnicas, por lo que se decide explorar alternativas que permitan estimar y predecir la densidad de cotización en el corto plazo.

El modelo elegido se basa en la distribución Beta, $Beta \sim (\mu, \sigma^2)$, donde μ y σ^2 denotan su media y varianza respectivamente. Esta distribución se adapta a proporciones y permite enfocar el estudio en la variable densidad de cotización. Además de adaptarse a la forma unimodal y al recorrido en [0,1] de la densidad de cotización no impone restricciones de independencia a las variables.
Como la variable de estudio está definida en el dominio de los números reales y restringida al intervalo [0,1] la relación entre la media y las variables predictoras no es lineal.
Por la misma razón la varianza es heteroscedástica, acercándose a cero cuando la media se aproxima a los limites de su dominio, lo que indica que la varianza depende de la media. Debe tenerse en cuenta además la asimetría de la distribución en el caso de las proporciones.
El argumento para la heteroscedasticidad está en que la varianza de una proporción está dada por $V(t) = \mu(1-\mu)$, con μ la media de t. Cuando μ tiende a uno o a cero la varianza tiende a cero, mientras que la varianza es máxima cuando la media está en el medio del recorrido ($\mu = 1/2$).

Esta distribución está indexada por dos parámetros estrictamente positivos p y q, que determinan su forma y escala respectivamente. Ambos se relacionan con la esperanza ($E(y)$) y la varianza ($V(y)$) de la distribución de acuerdo a las siguientes ecuaciones:

$$E(y) = \frac{p}{p+q} \tag{18}$$

$$V(y) = \frac{pq}{(p+q)^2(p+q+1)} \tag{19}$$

En [16] se propone una reparametrización del modelo original que permite incluir la heteroscedasticidad, definiendo μ=p/(p+q) y Φ=p+q, con lo que la media y la varianza de la proporción son:

$$E(y) = \mu \tag{20}$$

$$V(y) = \frac{\mu(1-\mu)}{(1+\Phi)}. \tag{21}$$

donde Φ puede interpretarse como un parámetro de dispersión en el sentido que para valores dados de μ a mayor valor de Φ menor la varianza de y. Con estos parámetros la función de probabilidad queda expresada como:

$$f(y) = \frac{\Gamma(\Phi)}{\Gamma(\mu.\Phi)\Gamma((1-\mu).\Phi)} y^{\mu.\Phi-1}(1-y)^{(1-\mu).\Phi-1}, \text{ con } 0< \mu <1, \Phi >0. \tag{22}$$

Para estimar los parámetros se propone un modelo para la media de la forma $g(\mu) = \sum_{i=1}^{K} x_i \beta_i$, con $g(\mu) = \log(\mu/(1-\mu))$ la función link del logit, β un vector de parámetros de regresión y x_i las observaciones para las K variables, fijas y conocidas.
Este tipo de transformaciones se utilizan cuando la variable de respuesta está restringida al intervalo (0,1).La transformación además simplifica las estimaciones, construcción y evaluación del modelo.

Se explora además la alternativa de una transformación logarítmica en la densidad de cotización. Los resultados obtenidos son similares a los de la transformación logit por lo que no se considera en la aplicación. (Ver Anexo 5)

El parámetro de dispersión se estima mediante el método de máxima verosimilitud, donde la función de verosimilitud se expresa $\ell(\beta, \Phi) = \sum \ell_i(\mu_i, \Phi)$, con:

$$\ell_i(\mu_i, \Phi) = In\Gamma(\Phi) - In\Gamma(\mu_i\Phi) - In\Gamma((1-\mu_i)\Phi) + (\mu_i\Phi - 1)Iny_i + ((1-\mu_i)\Phi - 1)In(1-y_i) \quad (23)$$

Según los autores del material de referencia[16], Ferrari y Cribari-Nieto, algunas de las ventajas de la reparametrización son:

- La varianza es una función de la media por lo que no se impone homoscedasticidad.
- Los parámetros de la distribución Beta permiten modelar distintas formas de la distribución.

Antes de continuar son necesarias dos apreciaciones en cuanto a la estimación de los parámetros.
Una particularidad de la transformación logit es que se aplica a variables con recorrido en el intervalo (0,1), mientras que la densidad de cotización admite valores en [0,1]. En principio esto constituye una limitación, pero es necesario hacer algunas precisiones al respecto.
No corresponde que existan afiliados con densidad de cotización cero ya que deben aportar al BPS al menos una vez para afiliarse a una AFAP. En cuanto a la densidad de cotización 1, es razonable pensar que esos casos mantengan su comportamiento, y por lo tanto la densidad de cotización futura sea igual a la pasada. Para evaluar esta hipótesis se estudia el comportamiento de las tasas de variación respecto a otros años de los afiliados que a diciembre de 2005 presentan una densidad de cotización de 1. Se compara este resultado con el obtenido con los afiliados que no alcanzan una densidad de cotización 1 en el período 2000 – 2007 (este es el conjunto de datos con el que se construye el modelo).
En el siguiente cuadro los datos permiten verificar la conjetura de que los afiliados que alcanzan la máxima densidad de cotización en algún año presentan un comportamiento estable y no amerita que se investigue en profundidad su evolución en el tiempo. El cuadro número 9 demuestra la irregularidad en el comportamiento

del segundo grupo, en los que es conveniente estudiar la evolución de la densidad de cotización.

CUADRO 9: Comportamiento de la tasa de Variación de la densidad de cotización respecto a 2005 para los afiliados que alcanzan el valor máximo de densidad de cotización a 12/2005.

	Mínimo	Cuartil 1	Mediana	Media	Cuartil 3	Máximo
$V_{(2005,2003)}$	0	0	0	0	0	0
$V_{(2005,2004)}$	0	0	0	0	0	0
$V_{(2006,2005)}$	-0.067	0	0	-0.002	0	0
$V_{(2007,2003)}$	-0.157	0	0	-0.005	0	0

CUADRO 10: Características de la tasa de Variación de la densidad de cotización respecto a 2005 para los afiliados que no alcanzan el valor máximo de densidad de cotización entre 2000 y 2007.

	Mínimo	Cuartil 1	Mediana	Media	Cuartil 3	Máximo
$V_{(2005,2003)}$	-0.137	-0.014	0.004	0.001	0.016	0.122
$V_{(2005,2004)}$	-0.261	-0.026	0.006	-0.002	0.028	0.210
$V_{(2006,2005)}$	-0.106	-0.008	0.004	0.003	0.015	0.129
$V_{(2007,2003)}$	-0.199	-0.011	0.007	0.006	0.029	0.227

En cuanto a la obtención del valor máximo verosímil, dadas las dimensiones de la base y el número de operaciones, el cálculo computacional del máximo a priori no debería generar inconvenientes ni en tiempo de máquina ni en memoria. Sin embargo cuando se implementa el cálculo surge una limitación operativa: por las características de precisión de máquina el máximo entero que puede calcularse es 1.797693e+308 y entonces la función gamma sólo es aplicable para valores inferiores a 172. Esto restringe la búsqueda del óptimo en el intervalo (0,172) cuando en realidad el óptimo es un número real positivo que eventualmente puede superar este valor. Puntualmente, el parámetro de dispersión, equivale a la suma de los parámetros de la distribución Beta, p y q.

Si la densidad de cotización se interpreta como la probabilidad de aportar, entonces (p-1) representa la cantidad de éxitos (aporta) y (q-1) la cantidad de fracasos (no aporta). Con lo que (p+q-2) es la cantidad de oportunidades que el afiliado tiene para aportar, es decir la cantidad de meses transcurridos desde que se afilia hasta la fecha final del período que se considere. En este estudio la cantidad de intentos equivale a la antigüedad del afiliado en la administradora, la cual se crea en el año 1996. De esta forma la antigüedad en la actualidad no puede superar los 13 años, es decir los 156 intentos.

Para el año 2010 o 2011 algunos afiliados que son parte del sistema mixto desde el comienzo van a alcanzar valores de antigüedad para los que será necesario aplicar el

cálculo por aproximación. Puntualmente aquellos que alcancen los 14,5 años de antigüedad, 170 meses, p+q=172.

Para superar este inconveniente se consideran dos propiedades: la primera es la relación entre la función gamma y la función factorial: $\Gamma(\phi) = (\phi - 1)!, \forall \phi \in Z^*$; la segunda es una aproximación asintótica de la función factorial basada en la fórmula de Stirling $n! \approx n^n e^{-n} \sqrt{2\pi n}$, donde $\approx$ significa aproximadamente igual. Aplicando ambas se propone una aproximación a la log-verosimilitud:

$$\log L(\phi / \mu, y) \approx \sum_{i=1}^{n} \Delta + (\mu\phi - 1)\log(y_i) + ((1-\mu)\phi - 1)\log(1 - y_i) \qquad (24)$$

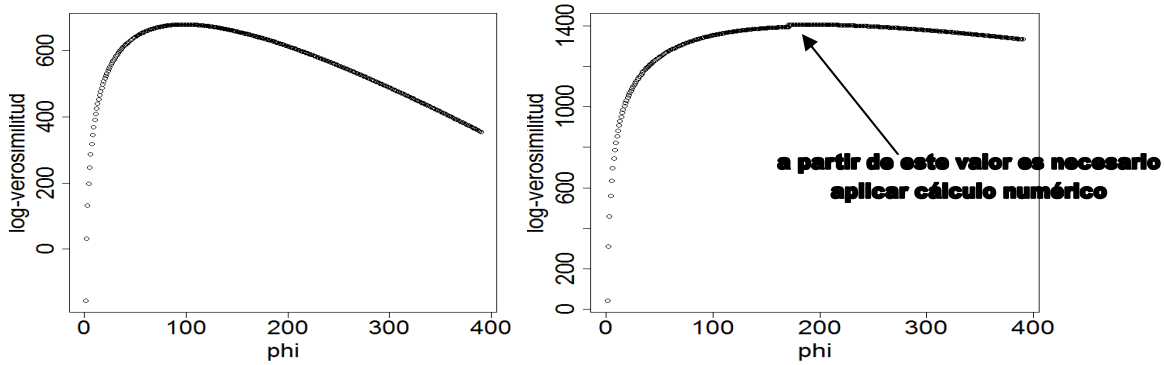
$$\Delta = \log\_ stg(\phi - 1) - \log\_ stg(\mu\phi - 1) - \log\_ stg((1-\mu)\phi - 1)$$

$$\log\_ stg(n) = n\log(n) - n + \frac{1}{2}\log(2\pi n)$$

De esta forma el cálculo del estimador puntual ϕ se basa en un algoritmo en el que se aplica el cálculo directo cuando el óptimo está en el intervalo [1,172] y la aproximación de Stirling cuando pertenece al intervalo $(172, \infty)$. Esta última sólo se aplica cuando el método directo no funciona, es decir cuando por defecto el cálculo define el máximo como el extremo superior del intervalo [1,172].

Para validar la aproximación obtenida se comparan los resultados de ambas estrategias en el valor límite. La verosimilitud estimada tiene un error relativo de 0.8% respecto al valor real. La oportunidad en las que deba usarse el cálculo aproximado depende de los datos con los que se trabaja, los gráficos a continuación muestran dos casos, sólo en el segundo es necesario aplicarla.

GRÁFICO 7: Curva de log-verosimilitud para aplicaciones de optimización. En el gráfico derecho son necesarias aproximaciones numéricas, en el izquierdo se obtiene por cálculo directo.



4.2. MODELO DEL DENSIDAD DE COTIZACIÓN MEDIA: CRITERIOS DE EVALUACIÓN

La evaluación del modelo considera tres aspectos: diagnóstico, capacidad predictiva y robustez respecto al momento del tiempo.
El diagnóstico contempla la significación de las variables y del modelo general, el ajuste y la consistencia de los coeficientes estimados con los resultados del análisis descriptivo. El poder explicativo de los modelos se mide a partir de pruebas de hipótesis y el estadístico R^2 ajustado, que permite comparar modelos con diferente cantidad de variables.
El análisis de capacidad predictiva se desarrolla sobre los datos con los que se construye el modelo y a partir de sample testing. Se analiza el poder predictivo a partir de un criterio básico de tolerancia y un Índice de Predicción (IP).

El criterio de tolerancia básico que se utiliza de aquí en adelante controla que la distancia entre la predicción y el valor real sea menor que la distancia entre el valor real y el que se considera como variable explicativa. Esta condición se impone como necesaria para seguir adelante con el modelo, si el último dato conocido se aproxima más al valor futuro real que la predicción esta última pierde sentido. Se define el índice de predicción (IP) de acuerdo a la siguiente expresión:

$$valor\_de\_referencia = \sum_{i=1}^{n} \frac{\left| DC_{(t+p)}(i) - DC_{(t)}(i) \right|}{n}$$

$$\bar{e} = \sum_{i=1}^{n} \frac{\left| DC_{(t+p)}(i) - \hat{DC}_{(t+p)}(i) \right|}{n} \qquad (25)$$

$$IP = valor\_de\_referencia - \bar{e}$$

donde i=1,.., n son los individuos, DC es la densidad de cotización, $\hat{DC}$ es la estimación de DC, t es el momento de tiempo considerado en las variables explicativas y p el rezago entre las variables explicativas y la de respuesta. La predicción es aceptable según nuestro criterio si IP es positivo.

Criterio de tolerancia de predicción:
IP>0 → Se acepta el modelo
IP≤0 → Se descarta el modelo

El alcance y la vigencia del modelo están sujetos a la condición de su no dependencia del momento del tiempo, si esto no se cumple conformaría una herramienta explicativa y no predictiva

La robustez del modelo respecto al momento del tiempo (no del rezago) se considera un aspecto fundamental dado que el fin de esta investigación es obtener una herramienta que describa el comportamiento de aportación de los afiliados, entendiendo por comportamiento la frecuencia con la que aportan.
Si el modelo no fuese estacionario entonces sería una herramienta descriptiva y su potencial para la estimación de la densidad de cotización se limitaría al período contemplado en el modelo y no la predicción de valores futuros.
En cambio si el modelo es una herramienta explicativa que sólo depende del rezago del tiempo pero no del período puede aplicarse tanto en instantes posteriores como anteriores a los usados para construir el modelo. Se busca es un modelo M que se obtiene con datos de t y t+k pero cuya aplicación puede extenderse a cualquier t, manteniendo el rezago k. Es decir, un modelo que dependa del rezago pero que sea robusto respecto al momento del tiempo.
En forma análoga si pensamos al modelo M como una función de la densidad de cotización en *t, D.C.*$_t$, la densidad de cotización en *t+k,* DC_{t+k} , y un conjunto de variables del afiliado en el momento k , llamémosle V.Afil $_k$ entonces:

$$M(DC_i, DC_{i+k}, V.Afil_k) = M(DC_j, DC_{j+k}, V.Afil_k); i \neq j$$

Un mecanismo para evaluar la robustez respecto al momento del tiempo es comparando la expresión de modelos con el mismo rezago pero para diferentes períodos. Los coeficientes estimados en un período deberían ser válidos al aplicarlos a datos de otro período si los rezagos coinciden. Otro mecanismo considera el poder predictivo de modelos construidos con datos de un período de tiempo y aplicados sobre una matriz de datos de un período posterior. Por ejemplo se genera con datos del período 2005-2007 y se aplica en datos del período 2007-2009.

4.3. MODELO DE LA DENSIDAD DE COTIZACIÓN MEDIA: ESTRUCTURA DE DATOS EN LA QUE SE APLICA

El modelo para estimar la densidad de cotización futura se aplica y evalúa en distintas estructuras de datos, lo que da lugar a dos aproximaciones.
En la primera aproximación se crean individuos tipo considerando las variables densidad de cotización actual, edad, franja de salario y giro de actividad; la cantidad de individuos tipo y la concentración de población en cada grupo depende de la categorización de las variables.
Las categorías se establecen buscando un equilibrio entre dos criterios, minimización de la pérdida de información y minimización de la cantidad de categorías. Este último evita que se conformen grupos con pocos afiliados. De acuerdo a esto se decide categorizar las variables de la siguiente manera:

CUADRO 11: Variables y sus respectivas categorías consideradas en el modelo

VARIABLE	CATEGORÍAS
Edad	18 a 22
	23 a 27
	28 a 32
	33 a 37
	38 a 42
	Mayor a 42
Franja de Salario	0 a 7057
	7058 a 22780
	Mayor a 22780
Giro de Actividad	Civil e Industria y Comercio
	Otros

VARIABLE	CATEGORÍAS
Densidad de Cotización	0 a 0.1
	0.1 a 0.2
	0.2 a 0.3
	0.3 a 0.4
	0.4 a 0.5
	0.5 a 0.6
	0.6 a 0.7
	0.7 a 0.8
	0.8 a 0.9
	0.9 a 1

Los individuos tipo quedan caracterizados por la densidad de cotización media y por la varianza del grupo y estos valores son el insumo para modelar los parámetros de la distribución. Por ejemplo para datos de los años 2008 y 2006 se obtienen 213 grupos (individuos tipo). El tamaño de los grupos es de 13 afiliados en promedio, y la dimensión varía entre 1 y 250 personas.

En el modelo de regresión lineal la variable dependiente es la transformación logit de la densidad de cotización futura promedio de los individuos tipo, y las variables explicativas son la edad promedio del grupo, el giro, la franja, y el logit de la densidad de cotización actual promedio del grupo. La transformación logit de la densidad de cotización como variable explicativa permite obtener resultados predictivos mejores a la aplicación sin transformar. La expresión del modelo es:

$$g(\mu_i) = \log\left(\frac{dc_{futura,i}}{1-dc_{futura,i}}\right) = \beta_{i,0} + \sum_{j=1}^{4} x_{ij}\beta_j$$
$$x_{i,1} = edad; x_{i,2} = giro; x_{i,3} = franja; x_{i,4} = \log it(dc_{actual}), \qquad (26)$$
$$i = 1,2,...,213.$$

donde μ_i es la media del i-ésimo individuo tipo, x_{ij} es la j-ésima variable explicativa del individuo i, y β_j es el parámetro correspondiente a la j-ésima variable explicativa.

Como resultado de la aplicación se obtienen tantas estimaciones de la media como individuos tipo con el que se estima el parámetro de dispersión por máxima verosimilitud para cada grupo. Las realizaciones de la muestra para este cálculo son los valores estimados de densidad de cotización de los afiliados que conforman el grupo. La Log-verosimilitud queda expresada de la siguiente forma:

$$\ell_k(\mu_i, \Phi_i) = In\Gamma(\Phi_i) - In\Gamma(\mu_i\Phi_i) - In\Gamma((1-\mu_i)\Phi_i) + (\mu_i\Phi_i - 1)Iny_{k,i} + ((1-\mu_i)\Phi_i - 1)In(1-y_{k,i})$$
$$k = 1,..,K = tamaño\_del\_grupo \qquad (27)$$

Donde μ_i y Φ_i son la media y el parámetro de dispersión del i-ésimo individuo tipo, y y_{ki} es el valor del k-ésimo afiliado que conforma al i-ésimo individuo. $\ell(\beta, \Phi_i)$ (27) es la log-verosimiltud para el individuo i.

Con esta aplicación no se logran resultados satisfactorios como se verá en el siguiente capítulo lo que motiva una segunda aproximación: el modelo de regresión lineal se aplica sobre los afiliados en lugar de individuos tipo.

Los resultados de predicción sobre la muestra completa para esta alternativa tampoco superan la tolerancia exigida (Ver Criterios de evaluación del modelo para la media). La explicación de este resultado reside en que no es posible captar las diferentes formas de evolución de la densidad de cotización en el tiempo. En la muestra se detecta tanto evolución positiva como negativa y el modelo debe captar esas fluctuaciones. El problema del modelo es el peso que tiene la densidad de cotización pasada para predecir, superando ampliamente a las demás variables explicativas.
Se considera entonces una agrupación de los afiliados en función de la evolución de la densidad de cotización en un tramo de tiempo anterior al utilizado para construir el modelo. Se define un indicador de variación, $V_{(t+p,\ t)}$ expresado como la diferencia entre la densidad de cotización en dos momentos del tiempo:

$$V_{(t+p,\ t)} = \text{densidad de cotización en } (t+p) - \text{densidad de cotización en } (t) \qquad (28)$$

En función de este indicador se forman los grupos de afiliados y se construye un modelo predictivo del valor esperado de la densidad de cotización futura para cada uno de ellos. Por ejemplo, para construir el modelo que estima el valor esperado de la densidad de cotización del 2007 en función de datos del 2005, se agrupan los afiliados de acuerdo a la diferencia entre la densidad de cotización de 2005 y de 2003, $V_{(2005,\ 2003)}$. Con esta metodología, que es la que finalmente se adopta como resolución al problema, se conforman 3 grupos en base a la magnitud de la variación.
Quienes tienen variación significativa y negativa, quienes tienen variación significativa y positiva, y un tercer grupo en el que no hay cambios de magnitud. La agrupación se define con dos puntos de corte en el recorrido del indicador, los que se obtienen aplicando árboles de regresión y criterios de tamaño y dispersión en los grupos.
La variable de respuesta del árbol es la variación entre las densidades de cotización que intervienen en el modelo predictivo, y la explicativa es el índice de variación considerado para agrupar. Tomando el ejemplo anterior, la variable de respuesta es $V_{(2007,2005)}$ y la variable explicativa es $V_{(2005,2003)}$.
Para seleccionar los puntos de corte se complementan los resultados de la técnica con dos requisitos: que los grupos cuente con un tamaño no menor al 10% de la muestra de afiliados, y que ambos puntos de corte sean del mismo orden en valor absoluto para evitar grupos con comportamiento atípico. El primero tiene prioridad sobre el segundo.

CAPITULO 5 RESULTADOS

En este capítulo se presentan los resultados obtenidos al aplicar las aproximaciones metodológicas detalladas anteriormente. Primero se exponen los resultados de la metodología aplicada a los individuos tipo, luego los de afiliados en forma individual y por último los correspondientes a la agrupación según antecedentes de aportación.

Con los datos de los individuos tipo se estiman los parámetros del modelo y por máxima verosimilitud se obtiene el parámetro de dispersión para cada grupo. A cada afiliado se le imputa como estimación puntual de densidad de cotización la obtenida para el grupo al cual pertenece. Se presenta un ejemplo de aplicación para datos de los años 2004 y 2006.

El modelo obtenido para la media es significativo, con un R^2 ajustado de 0.97. Las variables significativas al 5 % son la densidad de cotización del año 2004, la franja y el giro.

$$Logit(DC_i^{2006}) = 0.95 Logit(DC_i^{2004}) + 0.10 Franja\_2_i + 0.14 Franja\_3_i - 0.10 GiroA_i + \varepsilon_i \quad (29)$$

CUADRO 12: Coeficientes y medidas descriptivas del modelo de regresión por variable, para el modelo en el que se consideran individuos tipo

	Estimación	Error Std.	Valor t	Pr(>|t|)	
Constante	-0.079	0.079	-1.011	0.313	
LDC_{2004}	0.953	0.011	88.371	$<2.00x10^{-16}$	***
$Franja_{2004}.L$	0.101	0.039	2.566	0.011	*
$Franja_{2004}.Q$	0.139	0.054	2.607	0.009	**
Giro A	-0.104	0.037	-2.772	0.006	**
edad $_{2004}$	-0.001	0.002	-0.385	0.701	

Códigos de Significación: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.'

El poder predictivo del modelo es bueno, el error promedio[15] 0.03, y la optimización posterior en general no presenta dificultades. Los casos en los que la metodología se ve comprometida son aquellos en los que el grupo que define al individuo tipo es muy pequeño. Este inconveniente de todos modos podría subsanarse con una muestra más grande

El modelo tiene buen poder predictivo para los individuos tipo en los que la predicción aporta información disminuyendo el error promedio de 0.033 a 0.030. Los cuadros a continuación muestran los datos descriptivos de los errores considerados.

[15] El error medio es el promedio de las distancias en valor absoluto entre la densidad de cotización real y su predicción.

CUADRO 13: Error medio de predicción cuando se imputa el valor de la densidad de cotización que se corresponde con la variable explicativa del modelo

Mínimo	Cuartil 1	Mediana	Media	Cuartil 3	Máximo
0	0.0097	0.0262	0.0467	0.0712	0.2621

CUADRO 14: Error medio de predicción cuando se imputa como valor futuro el valor de la media grupal del individuo tipo

Mínimo	Cuartil 1	Mediana	Media	Cuartil 3	Máximo
0.0001	0.0167	0.0347	0.0520	0.0746	0.2485

CUADRO 15: Error medio de predicción cuando se imputa en los individuos tipo el valor de la densidad de cotización que es variable explicativa en el modelo

Mínimo	Cuartil 1	Mediana	Media	Cuartil 3	Máximo
0.0002	0.0076	0.0204	0.0331	0.0506	0.1738

CUADRO 16: Error medio si para predecir en los individuos tipo se utiliza la predicción generada con el modelo

Mínimo	Cuartil 1	Mediana	Media	Cuartil 3	Máximo
0.0001	0.0082	0.0207	0.0301	0.0385	0.1523

El problema de la metodología radica no en la calidad del modelo sino en el interés de la investigación. Mientras el modelo explica el comportamiento general de individuos tipo el interés de este trabajo está en los datos de los afiliados de manera individual. De esta forma es impreciso asignar a un afiliado las medidas resumen del grupo al cual pertenece en lugar de utilizar los datos del afiliado en el pasado.
La falla en este planteo se detecta cuando se pretende estimar la densidad de cotización para un afiliado en función del valor resumen del grupo. Si se compara el error de predicción obtenido para los afiliados con el valor que se toma de referencia el poder predictivo no es satisfactorio. La distancia media entre la variable explicativa y el valor explicado es de 0.047 mientras que la distancia media entre la predicción y la variable explicada es 0.052. En particular el poder predictivo es pobre en los casos en que la variación en el tiempo de la densidad de cotización es importante.
La agrupación de individuos si bien simplifica la aplicación y se adapta al modelo elegido genera como contrapartida pérdida de información. La variabilidad intragrupal provoca que el error de predicción para los afiliados no alcance la tolerancia exigida.

En la segunda aproximación se trabaja con los afiliados en forma individual. En este caso el ajuste de los modelos es aceptable, con R^2 ajustado superior a 0.9 en todas las

combinaciones evaluadas. Las variables explicativas significativas al 1% son la densidad de cotización previa y la franja de renta, a medida que el rezago entre variables explicativas y variable de respuesta aumenta son significativas además el giro de actividad y la edad.
A modo de ejemplo se presentan los resultados para el modelo cuya variable de respuesta es DC_{2007} y el rezago con las explicativas es de 2 años. En Anexo 2 se incluyen algunos ejemplos en los cuales la variable de respuesta es DC_{2007} y el rezago respecto a las explicativas es de 1 año y 4 años.

CUADRO 17: Coeficientes y medidas descriptivas del modelo de regresión por variable, para el caso en que se aplica un único modelo a todos los afiliados.

	Estimación	Error Std.	Valor t	Pr(>|t|)	
Constante	0.071	0.040	1.766	0.077	.
LDC_{2005}	0.987	0.004	2.304	$<2.00x10^{-16}$	***
$Franja_{2005}.L$	0.069	0.019	3.685	$2.3\ x10^{-4}$	***
$Franja_{2005}.Q$	-0.052	0.013	-3.876	$1.1\ x10^{-4}$	***
Giro	-0.033	0.024	-1.366	0.172	
edad $_{2005}$	$4.9\ x10^{-4}$	0.001	0.477	0.633	

Códigos de Significación: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

La predicción cumple con la condición de tolerancia mínima estipulada a priori (IP>0), sin embargo, nuevamente los resultados no colman las expectativas.(IP es 0.0025). Conjuntamente se relativiza el índice para medir el aporte de la predicción, dividiendo IP entre $Dis(DC_{(t+p)}, DC_{(t)})$, y se obtiene que la mejora del error de estimación no supera el 3% al utilizar el modelo.

Los resultados obtenidos motivan una tercera aplicación, en la que se decide trabajar con grupos generados a partir de la evolución de la densidad de cotización y se construye un modelo para cada uno de ellos. El diseño cumple con los criterios de bondad de ajuste y predicción definidos y es el modelo definitivo en la investigación.
Los modelos lineales se generan en forma independiente para cada grupo donde los afiliados que los conforman son las unidades de análisis. La agrupación se determina en función de la evolución de la densidad de cotización y las variables explicativas son las mismas que en las aplicaciones anteriores. Son necesarios dos coeficientes, uno negativo y uno positivo, que se ajusten a los criterios de no disparidad de los tamaños de los grupos y sean del mismo orden en valor absoluto.

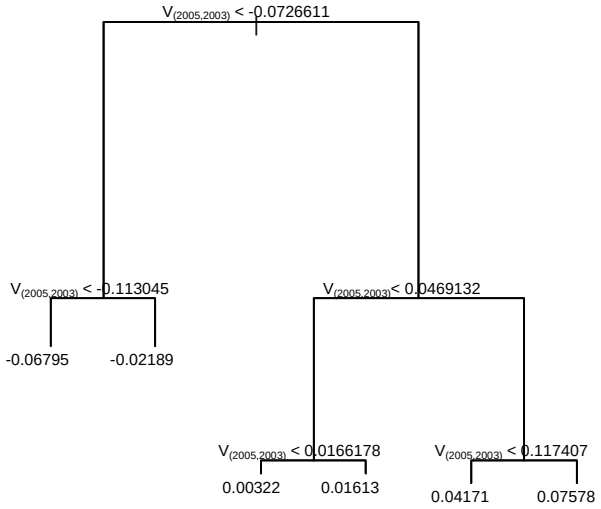
Los coeficientes se determinan aplicando árboles de regresión, con variable dependiente $V_{(t+p,\ t)}$ y variable independiente $V_{(t,\ t-q),}$ los valores p y q enteros positivos, p y q en el intervalo [1,4] p rezago del modelo y q rezago para la agrupación. $V_{(t+p,\ t)}$ representa la evolución entre los valores de densidad de cotización utilizados en el modelo de regresión, ($DC_{(t+p)}$ es la variable de respuesta y $DC_{(t)}$ la explicativa). Por otra parte $V_{(t,\ t-q)}$ considera los antecedentes, es decir la evolución de la densidad de cotización y es la variable utilizada para agrupar.

El árbol de regresión proporciona particiones homogéneas respecto a la evolución de la densidad de cotización entre t y (t+p) basado en el comportamiento de aportación entre (t-q) y t.
Con este procedimiento obtenemos un mejor ajuste en los modelos respecto a las propuestas anteriores. Con los árboles de clasificación se busca que el comportamiento intra- grupos en términos de $V_{(t+p,\ t)}$ sea homogéneo, alcanzando un mejor ajuste.

Algunas generalidades que se identifican en los coeficientes asociados a los árboles, y que permiten entender la aplicación se obtienen del estudio de sensibilidad de los árboles ante cambios en los rezagos y en los momentos del tiempo. Se detecta que a mayor q y por consiguiente mayor dispersión de la variable explicativa son mayores en valor absoluto los coeficientes que determinan los primeros nodos del árbol. El efecto del rezago se mantiene aún cuando no interviene el factor momento del tiempo. (Ver anexo 3) Cuando se incluye el año 2002 en el período para un mismo rezago la variable explicativa presenta mayor dispersión y son superiores en valor absoluto los coeficientes de los primeros nodos. Esta dependencia con el momento del tiempo se manifiesta únicamente con el año 2002, fenómeno notado en instancias previas que se atribuye al contexto de crisis existente. En consecuencia se considera conveniente emplear únicamente la información contenida en los períodos que inician en el año 2003 para construir los árboles y formar los grupos.
Para demostrar la influencia del momento del tiempo y de los rezagos se estudian algunos casos[16], se toma como caso de referencia el árbol en que la variable de respuesta es $V_{(2007,2005)}$ y la explicativa $V_{(2005,2003)}$. Se incluyen otros ejemplos en Anexo 3.

GRÁFICO 8: Árbol de regresión para datos del período 2003-2007.

- Variable de respuesta: $V_{(2007,2005)}$ Variable explicativa: $V_{(2005,2003)}$



[16] A todos los árboles construidos les impusimos las siguientes condiciones: cantidad mínima de nodos = 5, tamaño mínimo del nodo = 10, deviance mínima = 0.005

Medidas descriptivas consideradas para el árbol

- Número de nodos terminales: 6
- Deviance media de los residuos: 0.002125 = 5.195 / 2444

CUADRO 18: Distribución de los residuos

Mínimo	Cuartil 1	Mediana	Media	Cuartil 3	Máximo
-0.203	-0.015	7.84×10^{-6}	1.16×10^{-18}	0.0107	0.211

CUADRO 19: Resumen de la variable $V_{(2005,2003)}$

Media	Desvío
-0.002	0.067

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
-0.260	-0.095	-0.044	-0.016	3×10^{-4}	0.006	0.012	0.021	0.038	0.074	0.210

Dado que el principal propósito del uso de esta herramienta es la obtención de dos valores, uno positivo y otro negativo del mismo orden, la exigencia de ajuste y crecimiento del árbol se adecua al mismo. No fue necesario establecer niveles de deviance mínima menor ya que con los primeros nodos se identifican los coeficientes que determinan los grupos. De todas maneras la deviance de los residuos del árbol es aceptable si se compara con los niveles standard (0.005). La cantidad de nodos terminales es suficiente para obtener los coeficientes y puede aumentar o disminuir en función de la exigencia de impureza y tamaño de los nodos fijada para la construcción del árbol.

Analizados los resultados de los árboles se seleccionan los coeficientes que determinan los tres grupos: uno que comprende a aquellos afiliados que presentan decrecimiento en la densidad de cotización, otro con los que presentan un crecimiento y un tercero con los que no tienen grandes variaciones.
Los coeficientes seleccionados son los que se asocian a los nodos de la parte superior del árbol, cuyo poder de discriminación es mayor. En el primer ejemplo se eligen los valores que definen el primer y tercer nodo (-0.0726611, 0.0469132) de esta manera cada uno de los grupos contiene al menos al 10% de la población.

Si bien es necesario construir y estudiar un árbol para cada combinación de variables explicativas y de respuesta se observa que dado un mismo rezago los coeficientes y nodos que los contienen son similares para diferentes momentos del tiempo, a excepción del 2002. Este es un indicio de atemporalidad, fundamental en nuestra metodología.

Antes de continuar con la construcción del modelo resulta pertinente caracterizar a los grupos obtenidos. Se evalúa el comportamiento de variables como giro, origen, franja salarial, edad, antigüedad y densidad de cotización tomando como referencia al grupo más estable. Para simplificar la terminología en esta sección se denomina al

conjunto de afiliados cuya densidad de cotización presenta decrecimiento Grupo A, al que presenta crecimiento Grupo B y al que mantiene comportamiento estable Grupo C.
Comparando con el conjunto de referencia se obtiene que el grupo A se caracteriza por una edad promedio menor. Respecto al origen de afiliación, el peso relativo de la categoría traspaso es mayor y menor el de la categoría oficio. Este grupo además presenta los promedios más bajos de salario y densidad de cotización.
En cuanto al grupo B, los promedios de antigüedad, edad, densidad de cotización y salario son inferiores a los del conjunto de referencia. Además es superior la proporción de afiliados de oficio y menor la de voluntario.
El grupo C es el que presenta los niveles más altos de densidad de cotización lo que es consistente con la noción de que los afiliados que alcanzan los niveles más altos de densidad de cotización logran mayor estabilidad.
Algunas conclusiones obtenidas en el análisis exploratorio se retoman en este punto. La composición de los grupos observada es consistente con dicho análisis, el grupo C, caracterizado por un promedio mayor de densidad de cotización, presenta además indicadores de edad, salario y antigüedad superiores.

Obtenidos los coeficientes y construidos los grupos se cuenta con la estructura necesaria para generar los modelos predictivos. Los resultados del modelo para los distintos grupos se presentan para un caso particular: la variable explicada es la densidad de cotización del 2007, las variables explicativas corresponden al año 2005 y el árbol se estudia para el rezago 2003-2005. Se presentan otros casos en Anexo 4.

Respecto al diagnóstico del modelo de regresión resulta que se detectan observaciones outliers y no se cumple el supuesto de normalidad de los residuos. Por otra parte, no se identifican observaciones influyentes y no se rechaza la homoscedasticidad de los residuos estimados. Esto último se evalúa a partir de la prueba de varianza del error no constante (score test for non constant error variance).
Para la verificación del supuesto de normalidad se utilizan las pruebas de Kolmogorov Smirnov y Shapiro aplicadas sobre los residuos studentizados estimados. Con estos contrastes se rechaza la hipótesis de normalidad.
Una opción considerada para resolver la no normalidad de los residuos es eliminar observaciones atípicas del conjunto de datos utilizado para construir el modelo. Una vez puesta en práctica no se obtiene una mejora sustancial.
En base a estos resultados se decide no excluir observaciones atípicas y continuar con el modelo sin perder de vista que la no normalidad de los residuos afecta la interpretación de las pruebas de bondad de ajuste del modelo y los coeficientes. Todas las pruebas de bondad de ajuste se basan en el supuesto de normalidad, atendiendo a este problema se fijan niveles de significación más exigentes.

CUADRO 20: Resultados de diagnóstico para el modelo de referencia

GRUPO A	PRUEBA KOLMOGOROV-SMIRNOV	D: 0.3379
		p-value: < 2.2 e-16
	PRUEBA SHAPIRO-WILK	W: 0.863
		p-value: <2.2 e-16
	PRUEBA BONFERRONI (OUTLIERS)	Detecta outliers
	PRUEBA BREUSCH-PAGAN(VARIANZA DE LOS RESIDUOS)	Homoscedasticidad
		p-value: 0.096
	PRUEBA (INFLUYENTES)	No detecta
GRUPO B	PRUEBA KOLMOGOROV-SMIRNOV	D: 0.3663
		p-value: < 2.2 e-16
	PRUEBA SHAPIRO-WILK	W: 0.8755
		p-value: <2.2 e-16
	PRUEBA BONFERRONI (OUTLIERS)	Detecta outliers
	PRUEBA BREUSCH-PAGAN(VARIANZA DE LOS RESIDUOS)	Heteroscedasticidad
		p-value: 0.00
	PRUEBA(INFLUYENTES)	Detecta
GRUPO C	PRUEBA KOLMOGOROV-SMIRNOV	D: 0.3237
		p-value: < 2.2 e-16
	PRUEBA SHAPIRO-WILK	W: 0.8077
		p-value: <2.2 e-16
	PRUEBA BONFERRONI (OUTLIERS)	Detecta outliers
	PRUEBA BREUSCH-PAGAN(VARIANZA DE LOS RESIDUOS)	Homoscedasticidad
		p-value: 0.017
	PRUEBA(INFLUYENTES)	Detecta

CUADRO 21: Coeficientes y medidas descriptivas asociadas por variable y grupo.

	SIGNIFICACIÓN DE LAS VARIABLES	Estimación	Error Std.	Valor t	Pr(>|t|)	
GRUPO A	Constante	-0.187	0.078	-2.01	0.0169	*
	LDC_{2005}	0.942	0.021	43.678	<2e-16	***
	$Franja_{2005}$.L	0.0173	0.039	0.440	0.6601	
	$Franja_{2005}$.Q	-0.165	0.168	-0.985	0.325	
	Giro	0.0365	0.051	0.711	0.477	
	edad $_{2005}$	-3.0×10^{-3}	0.002	-0.162	0.8713	
GRUPO B	Constante	0.210	0.059	3.559	4.2e-4	***
	LDC_{2005}	0.916	0.014	65.577	2e-16	***
	$Franja_{2005}$.L	0.100	0.025	4.085	5.3e-5	***
	$Franja_{2005}$.Q	0.026	0.056	0.460	0.645	
	Giro	-0.060	0.039	-1.528	0.127	
	edad $_{2005}$	0.0008	0.002	0.524	0.600	
GRUPO C	Constante	0.0039	0.047	0.081	0.935	
	LDC_{2005}	0.989	0.004	207.017	2e-16	***
	$Franja_{2005}$.L	0.0713	0.022	3.292	0.001	**
	$Franja_{2005}$.Q	0.0608	0.029	2.041	0.041	*
	Giro	-0.0546	0.029	-1.867	0.062	.
	edad $_{2005}$	0.0011	0.001	0.944	0.345	

CUADRO 22: Resultados de ajuste de los modelos correspondientes a cada grupo

	ESTADÍSTICO F	P-VALOR	R^2 ajustado
GRUPO A	394, g.l. (5, 348)	2.2×10^{-16}	0,8477
GRUPO B	1026, g.l.(5,389)	2.2×10^{-16}	0.9286
GRUPO C	13.710 g.l. (5,1.695)	2.2×10^{-16}	0.9759

Según los contrastes realizados el ajuste del modelo y los errores de predicción son aceptables. Si bien el no cumplimiento del supuesto de normalidad de los residuos disminuye el potencial de la prueba de ajuste, el p-valor resultante es suficientemente bajo como para admitir el modelo. Algo similar sucede con las pruebas de significación de los parámetros.
En cuanto a la interpretación de los coeficientes se dificulta como consecuencia de la transformación de la variable dependiente. Las conclusiones respecto a la dependencia entre los predictores y la variable dependiente se deducen del análisis descriptivo.

El conjunto de variables significativas cambia en función del grupo que se considere. En el grupo A en general la única variable significativa es la densidad de cotización pasada, con coeficiente positivo y superior a 0.9. En el caso de los grupos B y C no hay un patrón de comportamiento, pero sí se identifican algunas coincidencias. En todos los casos más de una variable es significativa, además de la densidad de cotización pasada, la franja y el giro alcanzan niveles de significación satisfactorios. La edad del afiliado, en presencia de otras variables pierde poder explicativo.

En cuanto al poder predictivo se estudia en el conjunto completo de datos y por sample testing, utilizando un 70% de la base para entrenar el modelo y el 30% restante para estudiar la predicción. Si se comparan los resultados entre grupos se tiene que los mayores errores de predicción se registran en los grupos A y B. El grupo en el que los afiliados no registran variaciones importantes el error promedio es menor.

Para la comparación entre grupos, se relativiza el error utilizando el valor de la media y el desvío de la variable explicada. Se calcula qué porción de la media es el error ($(\bar{e}/\mu_{DC,(t+p)})$) y cuántas veces supera el desvío al error ($\sigma_{DC,(t+p)}/\bar{e}$), esto provee una medida general de precisión de la estimación. Para el ejemplo el error en el grupo A es un 11% de la media y 4 veces menor que el desvío. En los grupos B y C el error representa entre el 4 y 5% de la media, y es 8 y 13 veces menor que el desvío respectivamente.
En el grupo A, el error medio está en torno a 0.05, mejorando entre 20% y 30% al error que se obtendría al considerar como predicción la variable explicativa densidad de cotización. En el caso del grupo B el error promedio se reduce a valores cercanos a 0.03, con una mejora del 50% respecto al error de referencia. Por último en el grupo C el error promedio es de 0.02 y la mejora es de 8%. De acuerdo a lo anterior el grupo C es el que tiene mejor ajuste, lo que se explica entre otros factores por la heterogeneidad intragrupo. Mientras que en los grupos A y B el desvío estándar de $V_{(2007,2005)}$ es superior a 0.05 en el C es de 0.025.

De acuerdo a los resultados el modelo constituye una herramienta útil para explicar el comportamiento de aquellos afiliados atípicos. Estos casos representan en promedio un 30% del total de los afiliados estudiados y se caracterizan por variaciones de la densidad de cotización distinta al comportamiento común de la cartera.

Otro aspecto importante del modelo es la robustez respecto al momento del tiempo, la que se evalúa aplicándolo a datos pertenecientes a otros períodos y calculando el error de predicción resultante. Para todas las combinaciones el IP indica que la predicción se asemeja más al valor real futuro que el último dato disponible de dicha variable. La capacidad predictiva en cada grupo mantiene el comportamiento descrito para el modelo de referencia independientemente del período considerado.

CUADRO 23: Resultados predictivos por grupo de afiliados.

GRUPO A	Error medio($\bar{e}$)	$\mu_{DC,(t+p)}$	I.P.	APORTE	$(\bar{e}/\mu_{DC,(t+p)})$	$\sigma_{DC,(t+p)}$	$\sigma_{DC,(t+p)}/\bar{e}$
PREDICCIÓN (2007-2005-2003)	0.054	0.072	0.017	24%	12%	3.1	0.054
SAMPLE TESTING	0.052	0.068	0.016	23%	11%	3.3	0.052
2006-2004-2002	0.058	0.076	0.018	24%	11%	2.7	0.058
2005-2003-2001	0.065	0.078	0.013	17%	14%	3.2	0.065
2004-2002-2000	0.170	0.151	-0.019	-13%	27%	1.7	0.170
GRUPO B	Error medio($\bar{e}$)	$\mu_{DC,(t+p)}$	I.P.	APORTE	$(\bar{e}/\mu_{DC,(t+p)})$	$\sigma_{DC,(t+p)}$	$\sigma_{DC,(t+p)}/\bar{e}$
PREDICCIÓN (2007-2005-2003)	0.032	0.063	0.032	50%	5%	5.6	0.032
SAMPLE TESTING	0.036	0.070	0.034	48%	6%	5.2	0.036
2006-2004-2002	0.034	0.045	0.018	25%	4%	6.4	0.034
2005-2003-2001	0.043	0.072	0.029	41%	6%	4.3	0.043
2004-2002-2000	0.089	0.119	0.030	25%	15%	2.2	0.089
GRUPO C	Error medio($\bar{e}$)	$\mu_{DC,(t+p)}$	I.P.	APORTE	$(\bar{e}/\mu_{DC,(t+p)})$	$\sigma_{DC,(t+p)}$	$\sigma_{DC,(t+p)}/\bar{e}$
PREDICCIÓN (2007-2005-2003)	0.024	0.026	0.002	8%	3%	13.9	0.024
SAMPLE TESTING	0.025	0.028	0.003	9%	4%	13.0	0.025
2006-2004-2002	0.033	0.033	0.000	0%	6%	10.9	0.033
2005-2003-2001	0.028	0.030	0.002	6%	4%	12.0	0.028
2004-2002-2000	0.045	0.048	0.003	6%	7%	7.6	0.045

Por otro lado, al comparar los coeficientes de los diferentes modelos se obtiene que son similares en magnitud aunque no hay una combinación única de variables explicativas significativas. Estas diferencias en la expresión del modelo al comparar distintos períodos no descarta su carácter estacionario. Si bien los coeficientes no son idénticos, la interpretación es consistente y no es evidencia suficiente para concluir que el modelo no es estacionario.

Para continuar con el análisis, se genera un modelo para un período y se evalúa el poder predictivo en datos correspondientes a otro. En el ejemplo el modelo se genera para el período 2007-2005-2003 y se evalúa en los períodos 2004-2002-2000; 2005-2003-2001; 2006-2004-2002. Se presentan algunos resultados de predicción para estos casos.
El error de predicción para el período 2006-2004 y para 2005-2003 es en promedio de 0.06 y el modelo mejora la información que aporta la densidad de cotización pasada, mientras que para el 2002 -2000 el comportamiento es circunstancial, el modelo no aporta información y el error duplica al de los anteriores. Los resultados atípicos obtenidos en 2002 se atenúan cuando se consideran períodos más extensos que permitan revertir las variaciones bruscas generadas el contexto de crisis.

Los resultados obtenidos nuevamente sugieren que el comportamiento es estacionario, con errores de predicción similares a los obtenidos en el modelo al aplicar sample testing con los datos originales. Esto indica que el desempeño del modelo no depende del momento del tiempo en el que es construido.
Se considera apropiado utilizar los resultados de estimación de sample testing para comparar con la de los datos en los que se evalúa la atemporalidad. Es posible que el error de predicción se subestime si se aplica el modelo sobre los mismos datos con lo que fue construido, por lo tanto no se considera como parámetro de referencia.

Otra dimensión estudiada es la sensibilidad respecto a la amplitud de los rezagos. Como era esperado resulta favorable para el desempeño del modelo considerar rezagos más extensos para generar los grupos. Al aumentar el período de información los grupos obtenidos son menos heterogéneos y esto repercute en forma positiva en el poder predictivo. Si en cambio se considera el rezago entre variables explicativas y explicadas la conclusión es opuesta. Si el rezago es pequeño la evolución de la variable es menor y la incertidumbre en la variación se reduce, impactando en forma positiva en los resultados.

Conjuntamente con la robustez del modelo respecto al momento del tiempo debe analizarse la relación entre desvío y media de la densidad de cotización por año. Los resultados indican una tasa constante en el tiempo en todos los casos menos los de crisis o inestabilidad económica; en los que disminuye la media y aumenta la dispersión. Esto permite concluir que existe determinada regularidad en el comportamiento de la densidad de cotización en el tiempo (mirada desde el ángulo de relación desvío - media) que es trastocada en períodos adversos para la economía

CUADRO 24: Coeficiente de variación anual para la densidad de cotización ene el período 2000-2008.

Año	2000	2001	2002	2003	2004	2005	2006	2007	2008
Coeficiente de variación	0.44	0.45	0.51	0.47	0.48	0.48	0.48	0.47	0.47

Como última prueba de no dependencia respecto al momento del tiempo a partir del cual se construye el modelo se aplica alguno de ellos para datos con rezagos que contienen el año 2008 con el fin de analizar el poder predictivo en datos de períodos posteriores a los usados para generar el modelo. Este análisis de atemporalidad se distingue por incluir un año que no se contempló en ninguna de las etapas de la metodología, que es en definitiva la aplicación que se pretende para el modelo. Los modelos presentados hasta el momento fueron construidos y evaluados con datos del período 2003-2007.
La similitud entre los resultados predictivos obtenidos para los datos con los que se genera el modelo y para el 2008 constituye un nuevo indicio de no dependencia del momento del tiempo.

CUADRO 25: Medidas resumen del poder predictivo en los datos de origen de los modelos y en datos del 2008.

	MODELOS	DATOS	$\bar{e}$ *	VALOR DE REF.	I.P.	APORTE	$(\bar{e}/\mu_{DC,(t+p)})$	$\sigma_{DC(t+p)}/\bar{e}$
GRUPO A	2007-2005-2003	2007-2005-2003	0,054	0,072	0,017	24%	12%	3,1
		2008-2006-2004	0,043	0,065	0,022	34%	9%	3,9
	2007-2004-2003	2007-2004-2003	0,082	0,108	0,026	24%	17%	1,9
		2008-2005-2004	0,071	0,102	0,031	31%	15%	2,3
	2006-2004-2003	2006-2004-2003	0,057	0,084	0,026	31%	12%	2,6
		2008-2006-2005	0,037	0,073	0,036	49%	7%	4,1
GRUPO B		DATOS	$\bar{e}$	VALOR DE REF.	I.P.	APORTE	$(\bar{e}/\mu_{DC,(t+p)})$	$\sigma_{DC(t+p)}/\bar{e}$
	2007-2005-2003	2007-2005-2003	0,032	0,063	0,032	50%	5%	5,6
		2008-2006-2004	0,029	0,061	0,032	52%	5%	6,0
	2007-2004-2003	2007-2004-2003	0,058	0,076	0,018	24%	11%	3,9
		2008-2005-2004	0,046	0,095	0,049	52%	8%	3,5
	2006-2004-2003	2006-2004-2003	0,033	0,095	0,062	65%	5%	5,3
		2008-2006-2005	0,029	0,067	0,038	57%	5%	5,7
GRUPO C		DATOS	$\bar{e}$	VALOR DE REF.	I.P.	APORTE	$(\bar{e}/\mu_{DC,(t+p)})$	$\sigma_{DC(t+p)}/\bar{e}$
	2007-2005-2003	2007-2005-2003	0,024	0,026	0,002	8%	3%	13,9
		2008-2006-2004	0,022	0,024	0,002	6%	3%	15,1
	2007-2004-2003	2007-2004-2003	0,038	0,076	0,018	24%	11%	9,3
		2008-2005-2004	0,035	0,038	0,003	8%	5%	9,7
	2006-2004-2003	2006-2004-2003	0,025	0,027	0,002	7%	4%	13,4
		2008-2006-2005	0,022	0,024	0,002	7%	3%	15,1

Nota: La columna modelos identifica el período con el que se genera el modelo, la columna datos indica el período al que corresponden los datos sobre los que se aplica el modelo, y por consiguiente se estudia la predicción. $\overline{e}$ es el error medio

El resultado clave para evaluar el cumplimiento del objetivo de predicción futura a corto plazo y aplicación con nuevos datos, es la precisión para estimar la densidad de cotización en 2008 con datos de 2006 y un modelo construido con información del período 2004-2006. A partir de esta investigación se genera una herramienta que en la práctica permitirá estimar la densidad de cotización a 2010 con datos de 2008 y un modelo construido a partir de información de 2006-2008.

Una vez evaluado el poder predictivo y la atemporalidad de la metodología elegida, y para culminar con el análisis de desempeño se comparan los resultados de predicción con el modelo sin agrupación descartado al comienzo, unificando criterios de medición. Se calcula el error de predicción de ambas metodologías para el conjunto completo de datos y para cada grupo. En todos los casos la variable independiente pertenece al año 2007, las variables explicativas al año 2005 y la agrupación se construye para el período 2003-2005.

CUADRO 26: Comparativo de resultados según la metodología aplicada para la base global de afiliados

Metodología	DATOS	$\overline{e}$ *	VALOR DE REF.	I.P.	APORTE	$(\overline{e}/\mu_{DC,(t+p)})$	$\sigma_{DC(t+p)}/\overline{e}$
Con agrupación	base completa	0,030	0,039	0,009	23%	4%	11,4
Sin agrupación	base completa	0,037	0,039	0,002	6%	8%	4,6

CUADRO 27: Comparativo de resultados según la metodología para los grupos de afiliados obtenidos e función de la evolución de la densidad de cotización

METODOLOGÍA	DATOS	$\overline{e}$ *	VALOR DE REF.	I.P.	APORTE	$(\overline{e}/\mu_{DC,(t+p)})$	$\sigma_{DC(t+p)}/\overline{e}$
Con agrupación	Grupo A	0,054	0,072	0,017	24%	12%	3,1
	Grupo B	0,032	0,063	0,032	50%	5%	5,6
	Grupo C	0,024	0,026	0,002	8%	3%	13,9
Sin agrupación	Grupo A	0,077	0,072	-0,005	-7%	17%	2,2
	Grupo B	0,051	0,063	0,012	19%	8%	3,5
	Grupo C	0,024	0,026	0,002	7%	4%	13,9

En ambas pruebas los resultados predictivos están a favor del método de agrupación. La estimación es más precisa tanto en la base completa como en cada uno de los grupos. En particular en el grupo B en el que el modelo único no alcanza la tolerancia mínima, se alcanza un aporte del 25% respecto a la información disponible.

Los modelos obtenidos permiten mejorar la predicción de la densidad de cotización futura, ajustando generalmente la segunda y tercera cifra decimal. Esta mejora en la aproximación se resume en las diferencias que se generan en la estimación de la jubilación que tendrían en el futuro los afiliados. La sensibilidad en el cálculo es tal que para una variación de 0.02 en la densidad de cotización, dejando las demás variables constantes, el afiliado podrá o no jubilarse. Por ejemplo para un salario de $10.000 y densidad de cotización de 0.92 el afiliado se jubilaría a los 61 años de edad mientras que si la densidad de cotización es 0.90 se jubilaría a los 62 años.
En forma análoga un afiliado con un salario de $40.000 con densidad de cotización 0.82 se jubila por AFAP a los 65 y por BPS a los 66. Si la densidad de cotización se reduce a 0.80 se jubilaría por BPS a los 67.

CUADRO 28: Comparativo del monto de jubilación detallada por densidad de cotización según edad para un salario de $40.000.

Salario=$40.000	DC=0.92			DC=0.90		
Edad	**Jubilación BPS**	**Jubilación AFAP**	**Jubilación Total**	**Jubilación BPS**	**Jubilación AFAP**	**Jubilación Total**
60	NSJ[17]	NSJ	NSJ	NSJ	NSJ	NSJ
61	11.157	26.920	38.077	NSJ	NSJ	NSJ
62	11.800	29.331	41.132	11.586	28.725	40.311
63	12.444	31.969	44.413	12.230	31.308	43.538
64	13.088	34.859	47.947	12.873	34.138	47.011
65	13.731	38.030	51.762	13.517	37.243	50.760
66	14.375	41.503	55.878	14.161	40.644	54.804
67	15.019	45.324	60.342	14.804	44.385	59.189
68	15.662	49.533	65.195	15.448	48.507	63.954
69	16.306	54.188	70.494	16.092	53.065	69.157

CUADRO 29: Comparativo del monto de jubilación detallada por densidad de cotización según edad para un salario de $20.000.

Salario=$20.000	DC=0.92			DC=0.90		
Edad	**Jubilación BPS**	**Jubilación AFAP**	**Jubilación Total**	**Jubilación BPS**	**Jubilación AFAP**	**Jubilación Total**
60	NSJ	NSJ	NSJ	NSJ	NSJ	NSJ
61	11.157	6.007	17.164	NSJ	NSJ	NSJ
62	11.800	6.561	18.362	11.586	6.433	18.019
63	12.444	7.168	19.612	12.230	7.028	19.257
64	13.088	7.832	20.920	12.873	7.679	20.552
65	13.731	8.562	22.293	13.517	8.394	21.911
66	14.375	9.361	23.736	14.161	9.177	23.338
67	15.019	10.240	25.259	14.804	10.039	24.843
68	15.662	11.209	26.871	15.448	10.989	26.436
69	16.306	12.281	28.587	16.092	12.039	28.131

[17] NSJ: No se jubila

CUADRO 30: Comparativo del monto de jubilación detallada por densidad de cotización según edad para un salario de $10.000.

Salario=$10.000	DC=0.92			DC=0.90		
Edad	**Jubilación BPS**	**Jubilación AFAP**	**Jubilación Total**	**Jubilación BPS**	**Jubilación AFAP**	**Jubilación Total**
60	NSJ	NSJ	NSJ	NSJ	NSJ	NSJ
61	6.463	5.511	11.974	NSJ	NSJ	NSJ
62	6.835	6.001	12.837	6.711	5.879	12.591
63	7.208	6.537	13.745	7.084	6.404	13.488
64	7.581	7.124	14.705	7.457	6.979	14.436
65	7.954	7.768	15.722	7.830	7.610	15.440
66	8.327	8.473	16.800	8.203	8.301	16.504
67	8.700	9.249	17.949	8.576	9.061	17.636
68	9.073	10.104	19.177	8.949	9.898	18.847
69	9.446	11.049	20.495	9.322	10.824	20.146

En este documento se selecciona como modelo de referencia el construido a partir de los datos del período 2007-2005 con agrupación según la variación de la densidad de cotización entre 2005 y 2003. El rezago de predicción en este caso es de dos años pero es posible trabajar con rezagos mayores. Para la predicción de rezagos más amplios se plantean dos opciones, utilizar un solo modelo con el rezago completo o dividir el rezago en tramos y aplicar más de un modelo. En esta segunda alternativa los períodos considerados para construir los modelos deben ser consecutivos y se utilizan como variables explicativas (densidad de cotización previa) las predicciones de los modelos precedentes. La variable Franja utilizada como predictora en todos los modelos es la correspondiente al inicio del período, la edad y antigüedad pueden calcularse para cada modelo y el giro principal se supone que no varía. Por último en ambos planteos se utiliza información de aportación previa para la agrupación.

Por ejemplo, si se pretende predecir la densidad de cotización de 2008 con información de 2004 las alternativas son; utilizar un modelo con rezago de 4 años y utilizar información de 2004 para predecir 2008 o utilizar dos modelos, uno con información de 2004 para predecir 2006 y otro con información de 2006 para predecir 2008. En el segundo modelo los datos de densidad de cotización de 2006 se estiman a partir de las predicciones del primer modelo. Respecto al resto de las variables explicativas, Franja de Renta se supone constante y el resto son estáticas o pueden calcularse. Consideramos razonable el supuesto de franja constante por 2 años ya que solo el … % de los casos en la muestra cambian de franja en el período 2004-2006. La información utilizada para agrupar en el primer procedimiento sería 2004-2003. En el caso de los dos modelos se utiliza información de 2003-2004 para el primero y 2004-2006 para el segundo, imputando nuevamente como información de 2006 las predicciones del primer modelo. A su vez los modelos aplicados pueden ser construidos a partir de un período diferente como por ejemplo 2003 – 2007 en el caso de uno solo y 2003 – 2005 “concatenado” con 2005 - 2007, con reglas de clasificación obtenidas a partir de la variación de la densidad de cotización entre 2002 y 2003.

Se evalúa el poder predictivo de cada uno de los procedimientos para este caso y se presentan los resultados para comparar la bondad de ajuste y encontrar la alternativa más apropiada.

CUADRO 31: Resultados de la predicción de los modelos concatenados

MODELOS	GRUPO	$\bar{e}_*$	VALOR DE REF.	I.P.	APORTE	$(\bar{e}/\mu_{DC,(t+p)})$	$\sigma_{DC(t+p)}/\bar{e}$
2007-2005/2005-2003 2003-2002	A	0.102	0.140	0.038	27%	20%	1.358
2007-2005/2005-2003 2003-2002	B	0.056	0.145	0.090	62%	9%	2.205
2007-2005/2005-2003 2003-2002	C	0.055	0.062	0.007	11%	8%	5.813

CUADRO 32: Resultados de la predicción del modelo "único"

MODELOS	GRUPO	$\bar{e}_*$	VALOR DE REF.	I.P.	APORTE	$(\bar{e}/\mu_{DC,(t+p)})$	$\sigma_{DC(t+p)}/\bar{e}$
2007-2003 2003-2002	A	0.102	0.127	0.025	20%	22%	1.625
2007-2003 2003-2002	B	0.076	0.125	0.049	39%	12%	2.104
2007-2003 2003-2002	C	0.047	0.050	0.003	6%	7%	7.082

Cuadro 33: Comparación de los resultados de predicción

GRUPO	$\hat{e}_{M2} - \hat{e}_{M1}$	IPM2 - IPM1	$APORTE_{M2}$-$APORTE_{M1}$
A	0.000	0.013	0.074
B	-0.020	0.040	0.224
C	0.008	0.003	0.044

En ambos casos la aplicación de la metodología mejora la precisión de la estimación de la densidad de cotización "futura" respecto a la imputación del último valor disponible como estimación de la misma, es decir el IP y el Aporte son superiores a 0. Si analizamos el poder predictivo a partir de la mejora en la precisión respecto a la información disponible (IP y Aporte) concluimos que los mejores resultados se obtienen utilizando modelos encadenados en todos los grupos, especial en el grupo B. las medidas de error indican que para el grupo B se obtienen mejores resultados con los modelos concatenados y en el resto de los grupos los errores son similares. De acuerdo a esta información en este ejemplo particular es preferible utilizar dos modelos "concatenados" dividiendo el rezago en dos tramos que uno que considere el rezago completo pero es necesario realizar más pruebas para sugerir una metodología de forma definitiva.

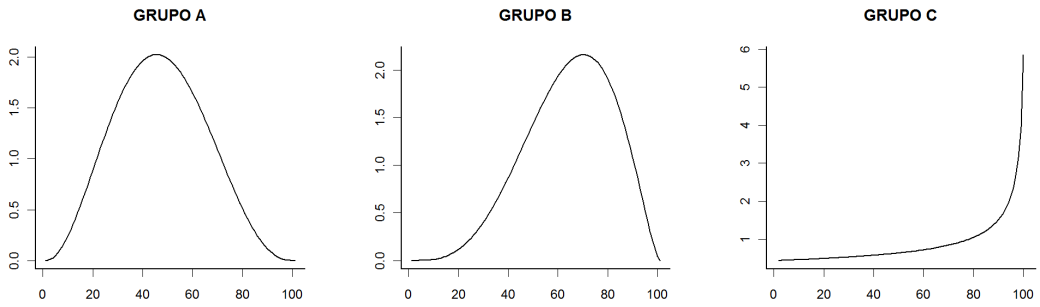
Una vez obtenidas las estimaciones de la media para cada afiliado resta estimar el parámetro de dispersión para cada grupo. Para calcular la máxima verosimilitud en

función de ϕ, se utiliza como estimación de la media de la densidad de cotización futura del grupo el promedio de las estimaciones obtenidas para los afiliados. Dadas la densidad de cotización de los grupos no fue necesario aplicar aproximaciones para obtener el $\hat{\phi}$ óptimo.

CUADRO 34: Parámetros de la distribución Beta para cada grupo.

GRUPO	$\hat{\mu}$	$\hat{\phi}$	$\hat{\sigma}$	p	Q
A	0.461	8.118	0.165	3.745	4.373
B	0.637	6.861	0.171	4.369	2.493
C	0.696	1.435	0.295	0.999	0.436

GRÁFICO 9: Distribución Beta asociada a cada grupo



Para el grupo C la distribución es estrictamente creciente concentra más probabilidad en los valores cercanos al 1 dado que lo integran los afiliados con mayor frecuencia de aportación.

Para los grupos A y B se obtienen distribuciones unimodales, en el primero la media es más baja consecuencia de que ambos parámetros son mayores que la unidad. La dispersión para ambos es superior a la del grupo C lo que se asocia las densidades de cotizaciones inferiores de los afiliados que lo integran.

CAPÍTULO 6: CONCLUSIONES

La frecuencia de aportación de los afiliados además de depender de sus antecedentes, se asocia a características propias de los individuos como el salario, la edad y el giro de actividad en el que se desempeña. Los jóvenes tienen mayor posibilidad de mejorar su nivel salarial y su estabilidad laboral con el transcurso del tiempo. El giro de actividad del afiliado y la franja salarial a la que pertenece están asociados a la estabilidad laboral y repercuten por lo tanto en el valor y la evolución de la densidad de cotización. Por ejemplo, personas con salarios elevados del sector público presentan comportamiento de aportación más regular que otras de bajo salario del sector construcción.

La transformación logit de la variable de respuesta realizada para adaptar los datos y trabajar con un modelo de regresión lineal provoca que la interpretación de los coeficientes del modelo se vuelva compleja. Adicionalmente dificulta la verificación de la coherencia de la expresión obtenida con los resultados de la etapa exploratoria. Dicha etapa es la que proporciona la información sobre la relación entre la variable dependiente y las variables predictoras del modelo.

El mayor alcance de la metodología se logra cuando se aplica a los afiliados cuyo comportamiento en el tiempo de la densidad de cotización es fluctuante. Esta propiedad permite focalizar el análisis en un conjunto de afiliados del cual se tiene mayor incertidumbre respecto al valor futuro. Para estos casos se obtienen estimaciones de un año dado, que superan ampliamente en precisión el dato de la densidad de cotización en años inmediatos anteriores.

La aplicación de la metodología para el período 2003 – 2007 permite la reducción del error de predicción, en el grupo de variación negativa es aproximadamente del 25% respecto al valor de la densidad de cotización en un momento previo (en este caso 2005); en el conjunto de variación positiva es del 50%.
Existe un tercer grupo en el que la variación de la densidad de cotización previa es despreciable, sobre el mismo se decide no aplicar el modelo y utilizar como estimación puntual el último valor disponible. Los cambios en la densidad de cotización en este caso no ameritan cálculos de predicción.

La información proporcionada por las distribuciones de probabilidad Beta estimada para cada grupo resume algunas características de los mismos. Una de ellas es que el desvío dentro del grupo con variación despreciable es pequeño y los valores más frecuentes se concentran en el rango 0.90 – 1. Por otra parte, los grupos con variación considerable presentan mayor desvío que el anterior. En el caso del grupo con decrecimiento la moda se encuentra cerca del 0.50 mientras que el conjunto con crecimiento los valores más frecuente están entre 0.70 y 0.80.

Los modelos construidos a partir de un intervalo de tiempo determinado son aplicables a otros períodos. No obstante, se considera necesario actualizar periódicamente los modelos para mantener sus bondades explicativas y predictivas. Para la actualización

además debe efectuarse el proceso de selección de coeficientes de agrupación así como la construcción de los modelos. La selección de los coeficientes se realiza a partir de una técnica exploratoria y criterios previstos, por lo que no es un proceso automático y debe evaluarse cada situación en particular.

Una limitación de la metodología es que no contempla la información macroeconómica en forma explícita, restringiendo su alcance a contextos de estabilidad económica. No obstante, alternativas como diferentes escenarios según contexto económico o cálculo de un deflactor de la estimación en función de variables como empleo e índice medio de salario pueden aplicarse para introducir este tipo de efecto. El método más sencillo es el primero, en el que se considera un escenario de crisis (por ejemplo el año 2002) otro de estabilidad (por ejemplo 2005) y otro de prosperidad (por ejemplo 2008). Para cada escenario se construye el modelo planteado en esta investigación y se estima la densidad de cotización futura. De esta manera se obtiene un rango de valores (en lugar de una estimación puntual) cuyos extremos serían las estimaciones para el contexto de crisis y el de prosperidad con un valor de referencia que es el del escenario de estabilidad económica.
La segunda alternativa es la consideración de un deflactor de la densidad de cotización futura (es decir, en t+k) que depende de variables macroeconómicas como la tasa de empleo y el Índice Medio de Salarios. El deflactor puede definirse en función de las expectativas de comportamiento de estas variables macroeconómicas para el momento t+k. El efecto del contexto en el momento t ya está incluido en una de las variables explicativas del modelo predictivo, densidad de cotización al momento t.

La información disponible a 12 años del inicio del régimen mixto y la influencia del contexto permite realizar únicamente estimaciones a corto plazo. De todas maneras se sienta precedente para aplicaciones similares en las que se obtengan predicciones a mediano plazo, por ejemplo rezagos de entre 5 y 8 años. No se tiene certeza de que sea posible una estimación precisa de la densidad de cotización a largo plazo dada la multiplicidad y dinámica de los factores que intervienen en su determinación

La estimación de la densidad de cotización futura será un insumo para obtener cálculos jubilatorios más precisos y para estudios de cobertura del sistema previsional. El programa de cálculo jubilatorio empleado por las AFAP en la actualidad toma como parámetros de entrada el salario al momento del cálculo, el fondo ahorrado en la AFAP, los años aportados, la edad y la densidad de cotización. El cálculo estima la jubilación que obtendría el afiliado por el tramo de la AFAP y por BPS así como la edad de configuración de causal jubilatoria, basándose en la reglamentación vigente y la lógica de funcionamiento de la capitalización individual. Respecto a las variables dinámicas, considera la evolución salarial pero asume densidad de cotización constante desde el momento del cálculo hasta el retiro. Lo ideal en este caso sería imputar como valor de densidad de cotización la acumulada al momento de configurar causal. La estimación de valores de densidad de cotización futura es una alternativa para reducir el error. El alcance del estudio está sujeto a la identificación de evolución de la densidad de cotización y la contribución respecto a la información disponible.

Actualmente los valores de densidad de cotización introducidos en el cálculo son el 1 o el último valor disponible. Si el cálculo se realiza con el valor 1, se obtiene el Monto jubilatorio y la edad de causal jubilatoria correspondiente a una aportación del cien por ciento en el período que va desde el momento del cálculo hasta la jubilación. Este procedimiento ocasiona una sobreestimación del monto jubilatorio y una subestimación de la edad de causal.
La alternativa de utilizar el valor actual de la densidad de cotización debe analizarse contemplando la operativa del cálculo que supone que éste dato se mantiene constante. En este contexto, entendemos que el valor actual de la Densidad de cotización es una buena aproximación para una gran proporción de la cartera, pero no del todo apropiada para determinado grupo de afiliados con una evolución de la frecuencia de aportación diferencial. La metodología desarrollada en esta investigación permite identificar este último segmento de afiliados y calcular en ese caso mediante el modelo predictivo planteado una estimación más precisa. Es importante destacar que los rezagos para los que la aplicación es válida son aquellos menores a 3 años.

Un aspecto importante de la metodología así como de los resultados aquí presentados es que se basan en una muestra representativa de la cartera de afiliados a RAFAP al año 2000. Dicha cartera no es un conjunto estático por lo que su conformación depende del contexto y por consiguiente del momento del tiempo. En este sentido es prudente que en el caso de que desee replicar la metodología en el futuro sed genere una nueva muestra de afiliados para actualizar las características de ese conjunto. Es de suponer que la tipología o relaciones identificadas no se modifiquen de manera sustancial, dado que se evalúa el comportamiento de aportación de los afiliados. De todas formas las estimaciones generadas se mueven en un rango de valores muy pequeño y la actualización del modelo asegura mayor solidez en las conclusiones.

El estudio de una proporción continua es frecuente, y este tipo de variables aparece en diversas áreas como economía, medicina, física, etc. La metodología seleccionada para modelar la densidad de cotización es una entre un conjunto de las posibilidades, como modelos de quasi-verosilimilitud, modelos gamma del logaritmo de la variable dependiente, distribución beta, regresión lineal de transformaciones de la variable de respuesta, etc.
La opción seleccionada para estimar la densidad de cotización de cada afiliado y para modelar la distribución de grupos de afiliados es una combinación de algunas de estas opciones. En particular se considera una transformación logit de la densidad de cotización y una regresión lineal para la estimación de su valor esperado y una distribución beta para el comportamiento grupal.
Existen otras alternativas válidas para identificar un modelo predictivo de la densidad de cotización futura que no han sido abordados en profundidad en este estudio y se recomienda considerar esas opciones para estudios posteriores.
Si bien este trabajo no es exhaustivo respecto a las metodologías para resolver el problema planteado se compararon resultados con una transformación alternativa al logit como la logarítmica. Con esta opción no se obtienen cambios relevantes en los resultados ni en la complejidad del trabajo.

Una de estas alternativas válidas es el modelo gama que se ajusta tanto al objetivo de la investigación como a la estructura de los datos. Se considera este modelo para aquellos casos con variable de respuesta no negativa sobre la cual se asume una distribución Gamma.

En cuanto al modelo de quasi-verosimilitud, se caracteriza por no imponer distribuciones sobre la variable y considerar la relación entre la media y la varianza, la que obtiene a partir de los datos observados. La desventaja de este método es que no permite el estudio de los afiliados de manera individual sino para grupos lo que implica pérdida de precisión respecto a la información disponible.

Esta técnica permite incorporar al estudio longitudinal componentes transversales que enriquecen el análisis. El componente transversal permite complementar el estudio del factor dinámico considerado normalmente en los modelos de autorregresión, con datos socioeconómicos y demográficos de los afiliados.

Un enfoque diferente es el bayesiano que se caracteriza por asociar distribuciones de probabilidad a parámetros de distribuciones de interés. Por ejemplo, una enfoque bayesiano puede considerar la densidad de cotización como una proporción poblacional que es resultado de una secuencia de intentos Bernoulli, en este caso aportar o no un mes determinado serían el éxito y el fracaso respectivamente. Dado un período de tiempo de *n* meses, en cada uno de ellos se evalúa si aporta o no aporta y la densidad de cotización sencillamente es la proporción de éxitos en el total de intentos.
Dentro de este enfoque se identifica una distribución a priori que por la naturaleza de los datos puede tomarse la probabilidad de aportar (*y*) dada la densidad de cotización (θ):

$$p(y/\theta) = Binomial(y/n,\theta) = C_y^n \theta^y (1-\theta)^{n-y}$$

En el caso más sencillo, asumiendo una distribución Uniforme(0,1) para θ la distribución a posteriori sería de la forma:

$$p(\theta/y) \propto \theta^y (1-\theta)^{n-y}$$
$$\theta/y \sim Beta(y+1, n-y+1)$$

La predicción para este tipo de enfoque en lugar de dar una estimación puntual de la densidad de cotización, como trabaja con distribuciones de los parámetros, permite calcular la probabilidad de que la densidad de cotización alcance determinado valor. Es útil para reproducir el comportamiento de la densidad de cotización de la cartera de afiliados o de grupos de interés. En el ejemplo presentado la formulación sería de la forma:

$$p(\tilde{y} = k/y) = E(\theta/y) = \frac{y+1}{n+2}$$

Otro elemento en el que se puede considerar el enfoque bayesiano es en los modelos de regresión normal en los que se asocia una distribución de probabilidad a los coeficientes y la varianza de la variable dependiente. Los modelos de regresión bajo el enfoque bayesiano asignan una distribución de probabilidad a las variables explicativas y a la variable dependiente. Un modelo bayesiano completo incluye una distribución

para X, $p(X/\Psi)$, la verosimilitud, $p(X, y/\Psi, \theta)$ y una distribución a priori para los parámetros, $p(\Psi, \theta)$.
De esta manera se pueden obtener estimaciones para los individuos de la densidad de cotización futura, asumiendo que la distribución a posteriori no depende del momento del tiempo.

Si bien existen varias aproximaciones al tema de predicción de la densidad de cotización se considera que la metodología planteada en este trabajo cumple con el objetivo específico, adaptándose a la estructura de datos disponible.

Vale destacar que en la valoración de los resultados debe considerarse que el cometido de la investigación es en cierta forma ambicioso dado que se pretende mejorar la predicción respecto a información disponible que en general es una aproximación certera. Los resultados generados constituyen una mejora en la estimación en especial en aquellos casos de comportamiento atípico en la cartera respecto a la aportación. Además se concluye que las mejoras obtenidas en la estimación repercuten considerablemente en los resultados de cálculo jubilatorio en particular en la edad de configuración de causal jubilatoria.

Como propuestas para trabajos posteriores pueden considerarse algunas de las técnicas mencionadas para comparar resultados predictivos o de diagnóstico y profundizar en el estudio de la densidad de cotización futura. Consideramos particularmente interesante la aplicación del modelo Gamma manteniendo el criterio de agrupación de este trabajo, generando estimaciones de la densidad de cotización para cada afiliado.

BIBLIOGRAFÍA:

1. Agresti,A.,(1996). *An Introduction to Categorical Data Analysis.* U.S.A.: John Wiley & Sons.

2. Benedetti,E. *Empleo informal en el Uruguay,* Instituto Nacional de Estadística, Encuesta continua de hogares 2006.

3. Bertranou, F. M., Sánchez, A. P., Oficina Internacional del Trabajo,(2003) *Características y determinantes de la densidad de aportes a al Seguridad Social en la Argentina 1994-2001. (pág 37)*Argentina, Ministerio de Trabajo, Empleo y Seguridad Social. Series de publicaciones de la Secretaría de la seguridad social. Año 1 no 1.

4. Breiman, L., Friedman, J. H., Olshen R. A., Stone,C. J. ,(1993) *Classification and Regression Trees.*U.S.A.:Chapman & Hall.

5. Bucheli,M. , Ferreira-Coimbra, N. ,Forteza, A., Rossi, I. (2006).*El acceso a la jubilación o* pensión en Uruguay: ¿cuántos y quiénes lo lograrían? Publicación de las Naciones Unidas.

6. Bucheli M., Forteza A. Rossi Ianina,(2006) *Seguridad Social y género en Uruguay: un análisis de las diferencias de acceso a la jubilación.* Proyecto seleccionado por un proyecto realizado por el Ministerio de Economía y Finanzas(Uruguay).

7. Caristo, A., Forteza, A.(2003).*El déficit del Banco de Previsión Social y su impacto en las finanzas del gobierno.* Trabajo realizado en el Departamento de Economía de Facultad de Ciencias Sociales(Uruguay)

8. Casella,G., Berger, R. L. ,(2002) *Statistical Inference.*(2º ed.) U.S.A.:Duxbury

9. Flajolet, P., Sedgewick, R.,(2008) *Analytic Combinatorics, U.S.A:*

10. Forteza, A., (2004) *Efectos Distributivos de la reforma de la seguridad social, el caso uruguayo.* Trabajo realizado a solicitud de la Conferencia Interamericana de la seguridad social(CISS)

11. Gelman, A., Carlin, J., Stern, H., Rubin,D., (2004).Bayesian Data Analysis.U.S.A: Chapman & Hall/CRCv

12. Kahaner, D.,Moler, C. ,Nash, S.(1989).*Numerical Methods and Software.* U.S.A.:Prentice-Hall.

13. Lagomarsino, G., Lanzilotta, B., *Densidad de aportes a la seguridad social en Uruguay. Análisis de su evolución y determinantes a partir de los datos registrales de historia laboral (1997-2003).*, Equipo de representación de los trabajadores en el BPS

14. Pardo, J.(2005-2006), *El mercado de las AFAP en Uruguay: caracterización y proyección de resultados por estratos representativos de la empresa líder del mercado.* Master en dirección y gestión de planes y fondos de pensión. Fundación CIFF, Organización Iberoamericanaza de la Seguridad Social, Universidad de Alcalá de Henarres(España).

15. Rencher, A. C. (2000). *Linear Models in Statistics.* U.S.A.: John Wiley & Sons.

16. Salinas-Rodriguez,A., Pérez-Nuñez R., Ávila-Burgos, L.,(2006) *Modelos de regresión para variables expresadas como una proporción continua.* Salud Pública de México/ vol. 48 no. 5, setiembre-octubre 2006

17. Särndal, C. E., Swensson, B., Wretman, J.(1992) *Model Assisted Survey Sampling.*U.S.A.:Springer

ANEXO 1: GLOSARIO

- Antigüedad: Tiempo transcurrido entre el momento de la afiliación a RAFAP y la actualidad.
- Densidad de cotización: Es la cantidad de aportes sobre la cantidad de meses en un período determinado.
- Franja salarial: Rango dentro del cual se ubica el salario por el que aporta el individuo.
- Giro: Rama de actividad en la que se desempeña el individuo. Las categorías son Civil, Construcción, Domestico, Industria y comercio y Rural.
- Origen: Refiere a la forma mediante la cuál el individuo se afilia a RAFAP. Las opciones son:
    - Voluntario: Eligen la AFAP a la que quieren ser afiliados y permanecen en la misma.
    - Oficio: La AFAP es asignada por BPS. Cuando se supera el tope salarial de $19.805 es obligatoria la afiliación a una AFAP, en este caso el individuo puede elegir su administradora y ser afiliado voluntario o el BPS le asigna una de oficio.
    - Traspaso: Afiliados a otra AFAP anteriormente
    - Reingreso de traspaso: refiere a individuos cuya afiliación a RAFAP fue interrumpida durante un período en el cual el individuo se traspasa a otra administradora.

TABLA DE VARIABLES EXPLICATIVAS.

TIPO	NOMBRE	CATEGORÍAS
NUMÉRICAS	ANTIGÜEDAD (AÑOS)	0 A 24
		25 A 49
		50 A 86
		87 A 123
		124 Y MÁS
	EDAD (AÑOS)	18 A 25
		26 A 35
		36 A 45
		46 A 55
		56 Y MÁS
	FRANJA ($)	0 a 2449
		2450 a 4544
		4545 a 7057
		7058 a 10586
		10587 a 15186
		15187 a 22780
		22781 a 34179
		34180 a 45461
		45462 y más
CATEGÓRICAS	GIRO	CIVIL
		CONSTRUCCIÓN
		DOMÉSTICO
		INDUSTRIA Y COMERCIO
		OTROS
		RURAL
	SEXO	FEMENINO
		MASCULINO
	ORIGEN	OFICIO
		REINGRESO DE TRASPASO
		TRASPASO
		VOLUNTARIO

ANEXO 2:
RESULTADOS DEL MODELO DE REGRESIÓN SIN AGRUPACIÓN

Los datos a continuación corresponden a el modelo en el que la variable explicada es del año 2007, las explicativas del año 2005 sin agrupación según tasa de crecimiento de la Densidad de Cotización.Los resultados son lo que argumentan el rechazo del modelo como una herramienta de predicción

	Estimación	Error Std.	Valor t	Pr(>|t|)	
Constante	0.071	0.040	1.766	0.077	.
LDC_{2005}	0.987	0.004	2.304	$<2.00x10^{-16}$	***
$Franja_{2005}.L$	0.069	0.019	3.685	$2.3\ x10^{-4}$	***
$Franja_{2005}.Q$	-0.052	0.013	-3.876	$1.1\ x10^{-4}$	***
Giro	-0.033	0.024	-1.366	0.172	
Edad $_{2005}$	$4.9\ x10^{-4}$	0.001	0.477	0.633	

- significación del modelo:

  Estadístico F: 16180, g.l. (2444)
  p-valor: $2.2\ x\ 10^{-16}$
  R^2 ajustado: 0.9706

	ERROR MEDIO($\bar{e}$)	VALOR DE REFERENCIA	I.P.	APORTE	$(\bar{e}/\mu_{DC,(t+p)})\times 100$	$\sigma_{DC,(t+p)}/\bar{e}$
PREDICCIÓN	0.036	0.039	0.002	6%	6%	8.5

Los datos a continuación corresponden a el modelo en el que la variable explicada es del año 2007, las explicativas del año 2006 sin agrupación según tasa de crecimiento de la Densidad de Cotización

	Estimación	Error Std.	Valor t	Pr(>|t|)	
Constante	0.0723884	0.0237514	3.048	0.00233	**
LDC_{2006}	0.995031	0.0024512	405.939	$2x10^{-16}$	***
$Franja_{2006}.L$	0.0581826	0.011058	5.262	$1.55x10^{-7}$	***
$Franja_{2006}.Q$	-0.0083457	0.0077228	-1.081	0.27996	
Giro	-0.0098911	0.0137392	-0.72	0.47164	
Edad $_{2006}$	-0.0006108	0.0005918	-1.032	0.30212	

- significación del modelo:

  Estadístico F: 4973, g.l. (2444)
  p-valor: 2.2 x 10^{-16}
  R^2 ajustado: 0.9902

	ERROR MEDIO($\bar{e}$)	VALOR DE REFERENCIA	I.P.	APORTE	$(\bar{e}/\mu_{DC,(t+p)})\times 100$	$\sigma_{DC,(t+p)}/\bar{e}$
PREDICCIÓN	0.019	0.020	0.001	6%	3%	16.6

Los datos a continuación corresponden a el modelo en el que la variable explicada es del año 2007, las explicativas del año 2003 sin agrupación según tasa de crecimiento de la Densidad de Cotización

	Estimación	Error Std.	Valor t	Pr(>|t|)	
Constante	-0.04911	0.069385	-0.708	0.47914	**
LDC_{2003}	0.979803	0.007963	123.05	$2x10^{-16}$	***
$Franja_{2003}$.L	0.09847	0.036598	2.691	0.00718	**
$Franja_{2003}$.Q	-0.052662	0.025097	-2.098	0.03598	*
Giro	-0.127809	0.043111	-2.965	0.00306	**
Edad $_{2003}$	0.005372	0.001856	2.894	0.00384	**

- significación del modelo:

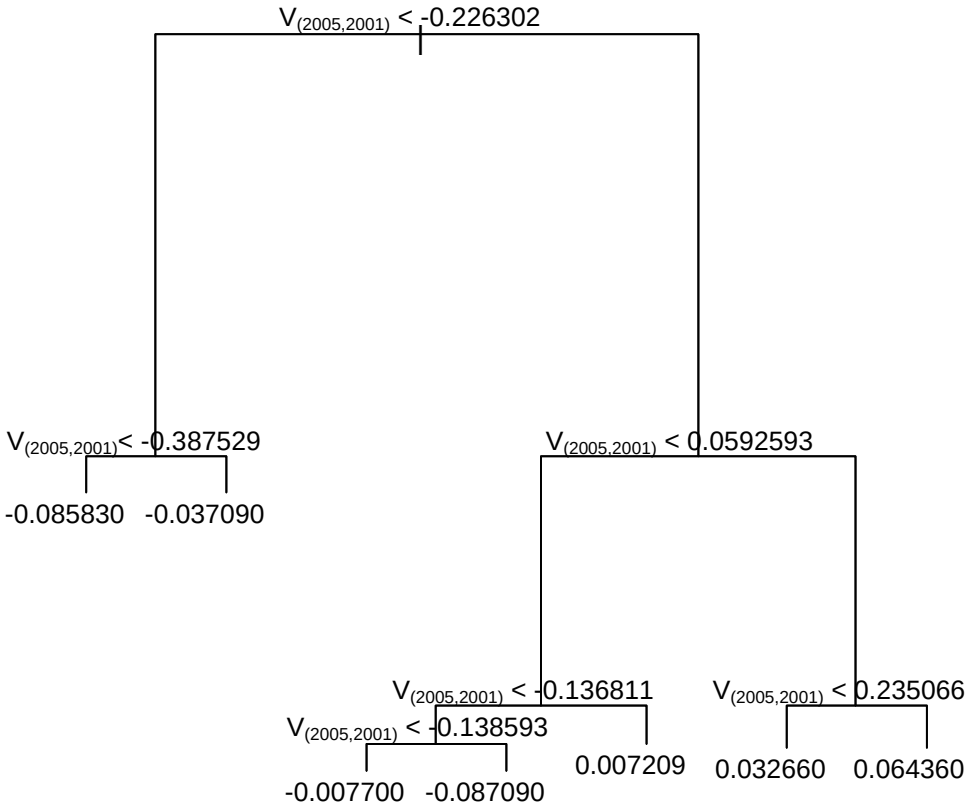
  Estadístico F: 4578, g.l. (2444)
  p-valor: 2.2 x 10^{-16}
  R^2 ajustado: 0.9033

	ERROR MEDIO($\bar{e}$)	VALOR DE REFERENCIA	I.P.	APORTE	$(\bar{e}/\mu_{DC,(t+p)})\times 100$	$\sigma_{DC,(t+p)}/\bar{e}$
PREDICCIÓN	0.072	0.075	0.003	4%	11%	4.3

ANEXO 3:
ÁRBOLES DE REGRESIÓN CONSIDERADOS PARA GENERAR GRUPOS DE AFILIADOS

Este anexo se presenta algunos ejemplos de árboles de regresión a partir de los cuales se puede evaluar el efecto del rezago y el momento del tiempo considerado en las variables. Los árboles son una herramienta para identificar los coeficientes que determinan los grupos.

Variable de respuesta: $V_{(2007,2005)}$ Variable explicativa $V_{(2005,2001)}$



Medidas descriptivas

- Número de nodos terminales: 7
- Deviance media de los residuos: 0.002453 = 5.992 / 2443
- Distribución de los residuos:

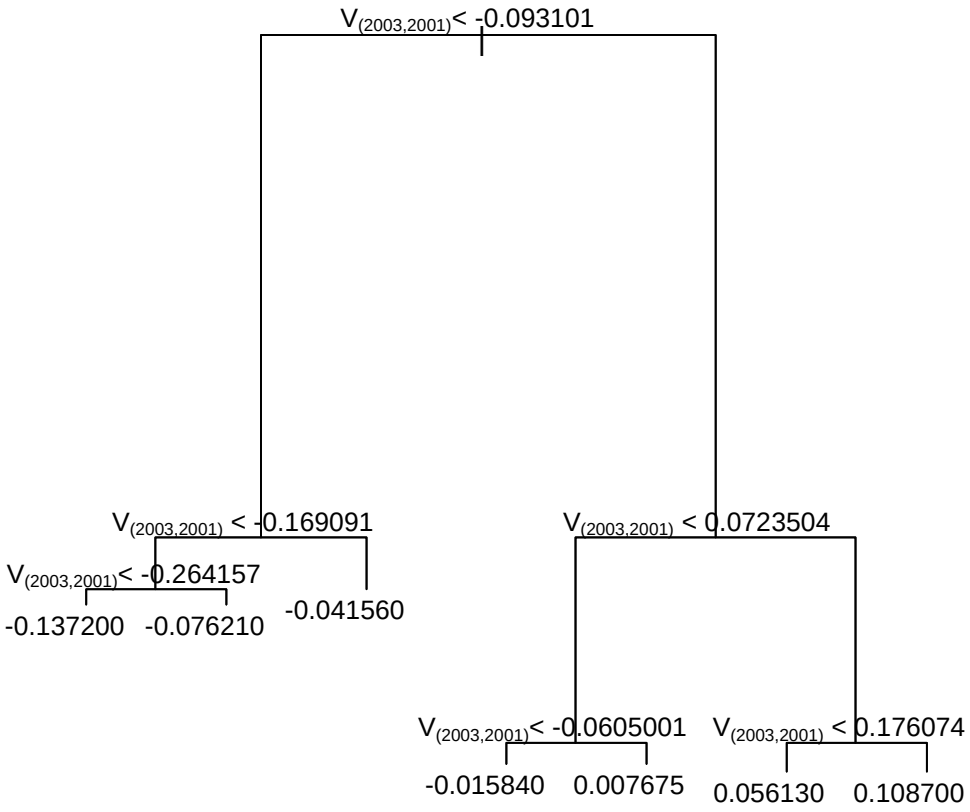
Mínimo	1er Cuartil	Mediana	Media	3er Cuartil	Máximo
-0.207	-0.0179	-0.00216	-2.98×10^{-19}	0.0154	0.222

Resumen de la variable $V_{(2005,2001)}$

Media	Desvío	0%	10%	20%	30%	40%
-0.027	0.138	-0.5317	-0.226	-0.123	-0.055	-0.021

50%	60%	70%	80%	90%	100%
-0.0026	0.011	0.025	0.050	0.115	0.560

Variable de respuesta $V_{(2005,2003)}$ Variable explicativa $V_{(2003,2001)}$



Medidas descriptivas consideradas para el árbol

- Número de nodos terminales: 7
- Deviance media de los residuos: 0.003258 = 7.96 / 2443
- Distribución de los residuos:

Mínimo	1er Cuartil	Mediana	Media	3er Cuartil	Máximo
-0.268	-0.023	-0.001	0.000	0.020	0.282

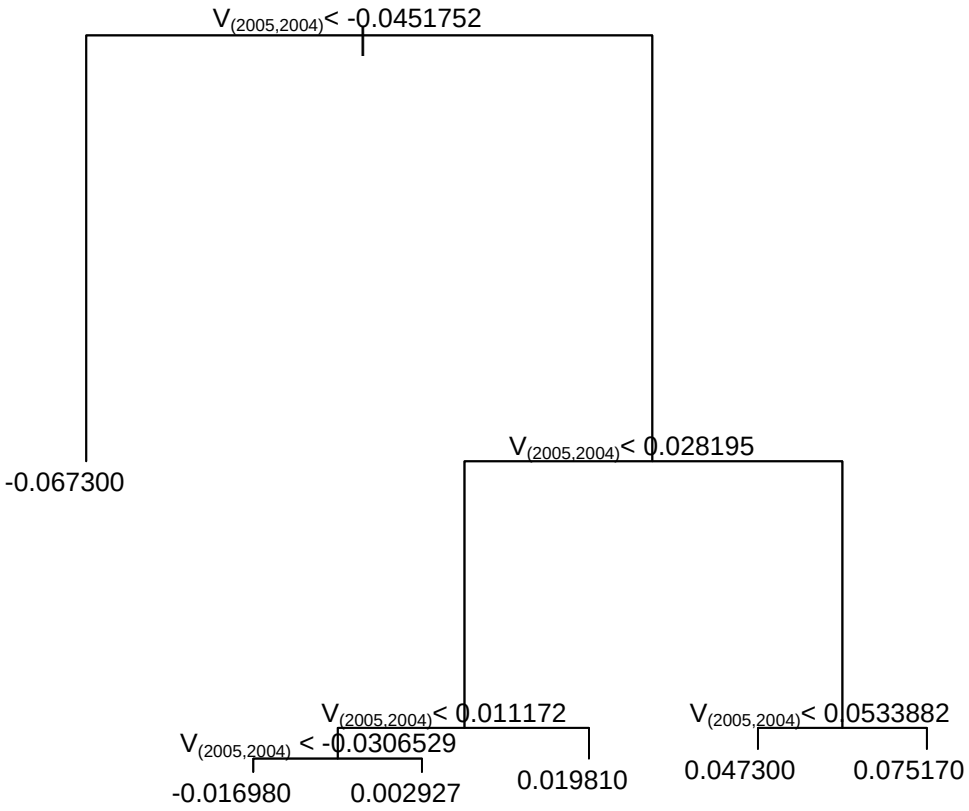
- Resumen de la variable $V_{(2003,2001)}$

Media	Desvío	0%	10%	20%	30%	40%

-0.025	0.091	-0.372	-0.152	-0.095	-0.052	-0.024

50%	60%	70%	80%	90%	100%
-0.007	0.004	0.011	0.026	0.063	0.405

Variable de respuesta $V_{(2007,2005)}$ Variable explicativa $V_{(2005,2004)}$



Medidas descriptivas consideradas para el árbol

- Número de nodos terminales: 6
- Deviance media de los residuos: 0.001876 = 4.586 / 2444
- Distribución de los residuos:

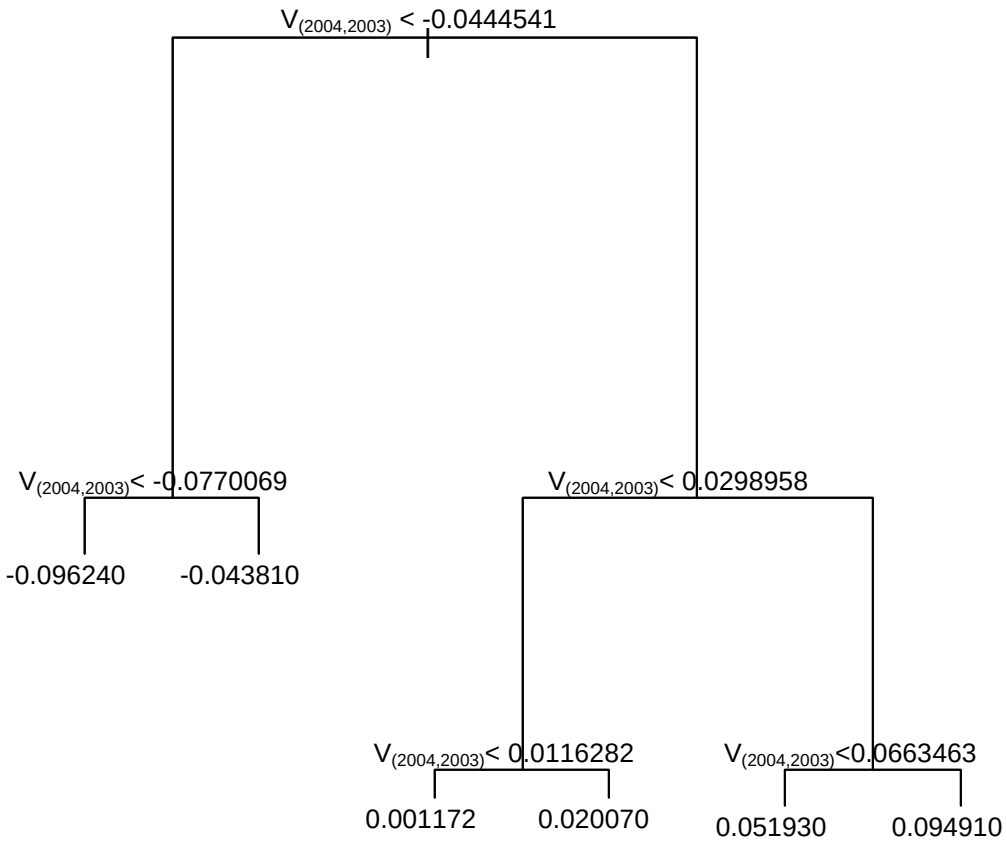
Mínimo	1er Cuartil	Mediana	Media	3er Cuartil	Máximo
-0.202	-0.013	0.002	0.000	0.011	0.195

- Resumen de la variable $V_{(2005,2004)}$

Media	desvío	0%	10%	20%	30%	40%
0.006	0.054	-0.199	-0.070	-0.026	-0.005	0.003

50%	60%	70%	80%	90%	100%
0.007	0.012	0.022	0.041	0.073	0.227

$V_{(2006, 2004)}$ en función de $V_{(2004,2003)}$



Medidas descriptivas consideradas para el árbol

- Número de nodos terminales: 6
- Deviance media de los residuos: 0.002227 = 5.444 / 2444
- Distribución de los residuos:

Mínimo	1er Cuartil	Mediana	Media	3er Cuartil	Máximo
-0.240	-0.018	0.002	0.000	0.013	0.200

- Resumen de la variable $V_{(2003,2001)}$

Media	Desvío	0%	10%	20%	30%	40%
-0.003	0.039	-0.149	-0.058	-0.030	-0.011	-0.002

50%	60%	70%	80%	90%	100%
0.003	0.006	0.011	0.020	0.041	0.126

ANEXO 4:
RESULTADOS DEL MODELO DE REGRESIÓN DE LA DENSIDAD DE COTIZACIÓN CON AGRUPACIÓN SEGÚN LA EVOLUCIÓN PREVIA DE LA DENSIDAD DE COTIZACIÓN (DATOS DEL PERÍODO 2001-2007)

Los datos a continuación corresponden a el modelo en el que la variable explicada es del año 2007, las explicativas del año 2005 y el árbol de construye para en función de la variación 2005-2001.

GRUPO A

SIGNIFICACIÓN DE LAS VARIABLES	Estimación	Error Std.	Valor t	Pr(>|t|)	
Constante	-0.116	0.077	-1.504	0.133	
LDC_{2005}	0.947	0.019	47.550	2e-16	***
$Franja_{2005}.L$	-0.001	0.039	-0.021	0.983	
$Franja_{2005}.Q$	0.086	0.116	0.741	0.459	
giro	0.075	0.050	1.490	0.137	
edad $_{2005}$	-0.001	0.002	-0.342	0.732	

Códigos de Significación: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- significación del modelo:

  Estadístico F: 477.3 g.l. (5, 443)
  p-valor: 2.2×10^{-16}
  R^2 ajustado: 0.842

	ERROR MEDIO($\bar{e}$)	VALOR DE REFERENCIA	I.P.	APORTE	$(\bar{e}/\mu_{DC(t+p)})\times 10($	$\sigma_{DC(t+p)}/\bar{e}$
PREDICCIÓN	0.061	0.069	0.008	11%	14%	2.9
VALIDACIÓN CRUZADA	0.062	0.068	0.006	9%	14%	2.9
ROBUSTEZ RESPECTO AL MOMENTO 06-04	0.063	0.069	0.006	9%	15%	3

GRUPO B

SIGNIFICACIÓN DE LAS VARIABLES	Estimación	Error Std.	Valor t	Pr(>|t|)	
Constante	0.203	0.054	3.741	2.1×10^{-4}	***

LDC_{2005}	0.956	0.010	89.777	2×10^{-16}	***
$Franja_{2005}$.L	0.073	0.024	3.063	2.3×10^{-4}	**
$Franja_{2005}$.Q	0.019	0.041	0.487	0.626	
giro	-0.111	0.040	-2.746	0.006	**
edad $_{2005}$	3.8×10^{-4}	0.001	0.256	0.798	

Códigos de Significación: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- significación del modelo:

  Estadístico F: 2071 g.l. (5,430)
  p-valor: $<2.2 \times 10^{-16}$
  R^2 ajustado: 0.959

	ERROR MEDIO($\bar{e}$)	VALOR DE REFERENCIA	I.P.	APORTE	$(\bar{e}/\mu_{DC(t+p)})\times 10($	$\sigma_{DC(t+p)}/\bar{e}$
PREDICCIÓN	0.028	0.051	0.023	45%	4%	7.2
VALIDACIÓN CRUZADA	0.029	0.055	0.026	47%	4%	7.1
ROBUSTEZ RESPECTO AL MOMENTO 06-04	0.034	0.054	0.020	37%	5%	5.8

- GRUPO C

SIGNIFICACIÓN DE LAS VARIABLES	Estimación	Error Std.	Valor t	Pr(>|t|)	
Constante	0.004	0.051	0.086	0.932	
LDC_{2005}	0.982	0.005	194.517	2×10^{-16}	***
$Franja_{2005}$.L	0.088	0.023	3.803	1.5×10^{-4}	***
$Franja_{2005}$.Q	0.069	0.032	2.132	0.033	*
giro	-0.072	0.031	-2.344	0.019	*
edad $_{2005}$	0.001	0.001	0.985	0.325	

Códigos de Significación: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- significación del modelo:

  Estadístico F: 12. 190 g.l. (5,1559)
  p-valor: $< 2.2 \times 10^{-16}$
  R^2 ajustado: 0.975

	ERROR MEDIO($\bar{e}$)	VALOR DE REFERENCIA	I.P.	APORTE	$(\bar{e}/\mu_{DC(t+p)})\times 10($	$\sigma_{DC(t+p)}/\bar{e}$
PREDICCIÓN	0.024	0.026	0.002	8%	3%	13.9
VALIDACIÓN CRUZADA	0.025	0.027	0.002	7%	4%	13.4
ROBUSTEZ RESPECTO AL MOMENTO 06-04	0.026	0.028	0.002	7%	4%	12.8

Los datos a continuación corresponden a el modelo en el que la variable explicada es del año 2007, las explicativas del año 2004 y el árbol de construye para en función de la variación 2004-2001.

GRUPO A

SIGNIFICACIÓN DE LAS VARIABLES	Estimación	Error Std.	Valor t	Pr(>|t|)	
Constante	-0.262	0.109	-2.392	0.0172	*
LDC_{2004}	0.886	0.032	27.842	2×10^{-16}	***
$Franja_{2004}.L$	-0.054	0.053	-1.024	0.307	
$Franja_{2004}.Q$	-0.017	0.144	-0.118	0.906	
giro	0.157	0.076	2.080	0.038	*
edad $_{2004}$	0.001	0.003	0.260	0.795	

Códigos de Significación: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- significación del modelo:

  Estadístico F: 159 g.l. (5,393)
  p-valor: < 2.2 x 10^{-16}
  R^2 ajustado: 0.665

	ERROR MEDIO($\bar{e}$)	VALOR DE REFERENCIA	I.P.	APORTE	$(\bar{e}/\mu_{DC(t+p)})\times 10($	$\sigma_{DC(t+p)}/\bar{e}$
PREDICCIÓN	0.086	0.102	0.016	16%	19%	2
VALIDACIÓN CRUZADA	0.091	0.108	0.017	16%	20%	1.8
ROBUSTEZ RESPECTO AL MOMENTO 06-03	0.088	0.101	0.013	13%	20%	2.1

GRUPO B

SIGNIFICACIÓN DE LAS VARIABLES	Estimación	Error Std.	Valor t	Pr(>|t|)	
Constante	0.246	0.085	2.906	0.004	**
LDC_{2004}	0.921	0.017	53.261	2×10^{-16}	***
$Franja_{2004}.L$	-0.003	0.038	-0.074	0.941	
$Franja_{2004}.Q$	-0.046	0.067	-0.0681	0.496	
giro	-0.210	0.064	-3.284	0.001	**
edad $_{2004}$	0.004	0.002	1.805	0.072	.

Códigos de Significación: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- significación del modelo:

  Estadístico F: 700 g.l. (5,389)
  p-valor: < 2.2 x 10^{-16}
  R^2 ajustado: 0.899

	ERROR MEDIO($\bar{e}$)	VALOR DE REFERENCIA	I.P.	APORTE	$(\bar{e}/\mu_{DC(t+p)})\times 10($	$\sigma_{DC(t+p)}/\bar{e}$
PREDICCIÓN	0.044	0.079	0.035	44%	6%	4.5
VALIDACIÓN CRUZADA	0.039	0.075	0.036	48%	6%	5
ROBUSTEZ RESPECTO AL MOMENTO 06-03	0.055	0.081	0.026	32%	8%	3.7

GRUPO C

SIGNIFICACIÓN DE LAS VARIABLES	Estimación	Error Std.	Valor t	Pr(>|t|)	
Constante	-0.011	0.065	-0.166	0.868	
LDC_{2004}	0.986	0.007	144.931	2.2×10^{-16}	***
$Franja_{2004}.L$	0.019	0.030	0.663	0.508	
$Franja_{2004}.Q$	0.041	0.044	0.937	0.349	
giro	-0.139	0.040	-3.484	5×10^{-4}	***
edad $_{2004}$	0.004	0.002	2.053	0.004	*

Códigos de Significación: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- significación del modelo:

  Estadístico F: 6.866 g.l. (5,1650)
  p-valor: $< 2.2 \times 10^{-16}$
  R^2 ajustado: 0.954

	ERROR MEDIO($\bar{e}$)	VALOR DE REFERENCIA	I.P.	APORTE	$(\bar{e}/\mu_{DC(t+p)})\times 10($	$\sigma_{DC(t+p)}/\bar{e}$
PREDICCIÓN	0.038	0.041	0.003	7%	5%	8.9
VALIDACIÓN CRUZADA	0.037	0.040	0.003	7%	5%	9
ROBUSTEZ RESPECTO AL MOMENTO 06-03	0.041	0.043	0.002	5%	6%	8.1

ANEXO 5: MODELO DE REGRESIÓN LINEAL: TRANSFORMACIÓN LOGARÍTMICA DE LA VARIABLE DE RESPUESTA

La Densidad de cotización es una proporción y la metodología seleccionada aplica un modelo de regresión lineal para explicar en una transformación logit de dicha proporción. Otra alternativa de transformación es la logarítmica, a partir de la cual se genera una variable con dominio en el intervalo $(0,+\infty)$. La expresión del modelo con esta transformación es del tipo:
$-ln(y)=x'\beta+e$, donde *y* es la densidad de cotización.

En este anexo se presentan los resultados obtenidos al reemplazar la transformación logit por una logarítmica en la metodología planteada en la investigación. Para comparar ambas alternativas se emplean los datos correspondientes al período 2007-2005-2003 (referencia) y se evalúa su desempeño predictivo en diferentes intervalos de tiempo previos a 2007 a partir de sample testing.

En el diagnóstico del modelo, al igual que en el modelo inicial, tampoco se cumple el supuesto de normalidad de los residuos. Además no se detectan observaciones influyentes y no se rechaza la hipótesis de homoscedasticidad.
El no cumplimiento de este afecta el análisis de la prueba de significación del modelo y de los parámetros.
En el análisis se identificaron observaciones atípicas en los tres grupos, las que pueden ser las posibles causantes de no normalidad. Sin embargo dado al excluir estas observaciones no se resuelve la no normalidad de los residuos y dado que se pretende estudiar grupos con comportamientos particulares se decide trabajar con la totalidad de los datos. Como se puede observar en el capítulo 4 el diagnóstico es similar para el modelo con la transformación logit.

Respecto a la expresión del modelo, la única variable explicativa significativa es la densidad de cotización previa. Cuando se utiliza la transformación logit las variables explicativas significativas son: la densidad de cotización y la franja de renta. En ambos casos la transformación dificulta la interpretación de los parámetros y la capacidad analítica del modelo es escasa. Es necesario recurrir a un análisis exploratorio para concluir respecto al comportamiento de la densidad de cotización y su relación con las variables predictoras. De todas maneras, consideramos conveniente un modelo que incluya otra variable explicativa significativa además de la densidad de cotización previa para que la predicción no dependa únicamente de los antecedentes de aportación y evitar así que se convierta en una herramienta similar a una serie de tiempo sencilla.
Por otra parte, la significación del modelo es similar aunque levemente superior para el modelo construido a partir de la transformación logit.

Cuadro Nº : Coeficientes y Significación de las variables del modelo con transformación logística período 2007-2003

	SIGNIFICACIÓN DE LAS VARIABLES	Estimación	Error Std.	Valor t	Pr(>|t|)	
GRUPO A	Constante	0.140601	0.053423	2.632	0.00886	**
	LDC_{2005}	0.994024	0.023775	41.81	<2e-16	***
	$Franja_{2005}.L$	0.048189	0.063084	0.764	0.44545	
	$Franja_{2005}.Q$	0.022826	0.039319	0.581	0.56193	
	Giro	-0.036845	0.026945	-1.367	0.17236	
	Edad $_{2005}$	-0.000217	0.001137	-0.191	0.84876	
GRUPO B	Constante	0.0648751	0.0369084	1.758	0.0796	.
	LDC_{2005}	0.7679159	0.0159138	48.255	<2e-16	***
	$Franja_{2005}.L$	-0.0143978	0.0211706	-0.68	0.4968	
	$Franja_{2005}.Q$	0.0271294	0.0141687	1.915	0.0562	.
	Giro	0.0276235	0.0218097	1.267	0.2061	
	Edad $_{2005}$	-0.0008575	0.0009069	-0.945	0.345	
GRUPO C	Constante	-0.0107026	0.0258682	-0.414	0.679	
	LDC_{2005}	0.9914892	0.0051499	192.526	<2e-16	***
	$Franja_{2005}.L$	0.0030233	0.0102432	0.295	0.768	
	$Franja_{2005}.Q$	-0.0035152	0.0078142	-0.45	0.653	
	Giro	-0.0106469	0.0145842	-0.73	0.465	
	Edad $_{2005}$	0.0001039	0.0006388	0.163	0.871	

Cuadro Nº : Coeficientes y Significación de las variables del modelo con transformación logística período 2007-2003

	SIGNIFICACIÓN DE LAS VARIABLES	Estimación	Error Std.	Valor t	Pr(>|t|)	
GRUPO A	Constante	-0.187	0.078	-2.01	0.0169	*
	LDC_{2005}	0.942	0.021	43.678	<2e-16	***
	$Franja_{2005}.L$	0.0173	0.039	0.440	0.6601	
	$Franja_{2005}.Q$	-0.165	0.168	-0.985	0.325	
	Giro	0.0365	0.051	0.711	0.477	
	edad $_{2005}$	-3.0×10^{-3}	0.002	-0.162	0.8713	
GRUPO B	Constante	0.210	0.059	3.559	4.2e-4	***
	LDC_{2005}	0.916	0.014	65.577	2e-16	***
	$Franja_{2005}.L$	0.100	0.025	4.085	5.3e-5	***
	$Franja_{2005}.Q$	0.026	0.056	0.460	0.645	
	Giro	-0.060	0.039	-1.528	0.127	
	edad $_{2005}$	0.0008	0.002	0.524	0.600	

Cuadro Nº :

GRUPO C	Constante	0.0039	0.047	0.081	0.935	
	LDC_{2005}	0.989	0.004	207.017	2e-16	***
	$Franja_{2005}$.L	0.0713	0.022	3.292	0.001	**
	$Franja_{2005}$.Q	0.0608	0.029	2.041	0.041	*
	Giro	-0.0546	0.029	-1.867	0.062	.
	edad $_{2005}$	0.0011	0.001	0.944	0.345	

Significación del modelo con transformación logística para el período 2007-2003

	ESTADÍSTICO F	P-VALOR	R^2 ajustado
GRUPO A	358.5, g.l. (5, 353)	2.2×10^{-16}	0.8331
GRUPO B	559.3, g.l.(5,396)	2.2×10^{-16}	0.8744
GRUPO C	9747 g.l. (5,1710)	2.2×10^{-16}	0.966

Cuadro Nº : Significación del modelo con transformación logit para el período 2007-2003

	ESTADÍSTICO F	P-VALOR	R^2 ajustado
GRUPO A	394, g.l. (5, 348)	2.2×10^{-16}	0,8477
GRUPO B	1026, g.l.(5,389)	2.2×10^{-16}	0.9286
GRUPO C	13.710 g.l. (5,1.695)	2.2×10^{-16}	0.9759

Cuadro N°: Diagnóstico del modelo con transformación logarítmica

GRUPO A	PRUEBA KOLMOGOROV-SMIRNOV	D: 0.2214
		p-value: < 9.992 e-16
	PRUEBA SHAPIRO-WILK	W: 0.824
		p-value: <2.2 e-16
	PRUEBA BONFERRONI (OUTLIERS)	Detecta outliers
	PRUEBA BREUSCH-PAGAN(VARIANZA DE LOS RESIDUOS)	Heteroscedasticidad
		p-value: 9.2 e-7
	PRUEBA (INFLUYENTES)	No detecta
GRUPO B	PRUEBA KOLMOGOROV-SMIRNOV	D: 0.2321
		p-value: < 2.2 e-16
	PRUEBA SHAPIRO-WILK	W: 0.7765
		p-value: <2.2 e-16
	PRUEBA BONFERRONI (OUTLIERS)	Detecta outliers
	PRUEBA BREUSCH-PAGAN(VARIANZA DE LOS RESIDUOS)	Heteroscedasticidad
		p-value< 2.2 e-16
	PRUEBA(INFLUYENTES)	No detecta
GRUPO C	PRUEBA KOLMOGOROV-SMIRNOV	D: 0.2736
		p-value: < 2.2 e-16
	PRUEBA SHAPIRO-WILK	W: 0.6433
		p-value: <2.2 e-16
	PRUEBA BONFERRONI (OUTLIERS)	Detecta outliers
	PRUEBA BREUSCH-PAGAN(VARIANZA DE LOS RESIDUOS)	Heteroscedasticidad
		p-value: <2.2 e-16
	PRUEBA(INFLUYENTES)	No detecta

Cuadro N°: Diagnóstico del modelo con transformación logit

GRUPO A	PRUEBA KOLMOGOROV-SMIRNOV	D: 0.3379
		p-value: < 2.2 e-16
	PRUEBA SHAPIRO-WILK	W: 0.863
		p-value: <2.2 e-16
	PRUEBA BONFERRONI (OUTLIERS)	Detecta outliers
	PRUEBA BREUSCH-PAGAN(VARIANZA DE LOS RESIDUOS)	Homoscedasticidad
		p-value: 0.096
	PRUEBA (INFLUYENTES)	No detecta
GRUPO B	PRUEBA KOLMOGOROV-SMIRNOV	D: 0.3663
		p-value: < 2.2 e-16
	PRUEBA SHAPIRO-WILK	W: 0.8755
		p-value: <2.2 e-16
	PRUEBA BONFERRONI (OUTLIERS)	Detecta outliers
	PRUEBA BREUSCH-PAGAN(VARIANZA DE LOS RESIDUOS)	Heteroscedasticidad
		p-value: 0.00
	PRUEBA(INFLUYENTES)	No detecta
GRUPO C	PRUEBA KOLMOGOROV-SMIRNOV	D: 0.3237
		p-value: < 2.2 e-16
	PRUEBA SHAPIRO-WILK	W: 0.8077
		p-value: <2.2 e-16
	PRUEBA BONFERRONI (OUTLIERS)	Detecta outliers
	PRUEBA BREUSCH-PAGAN(VARIANZA DE LOS RESIDUOS)	Homoscedasticidad
		p-value: 0.017
	PRUEBA(INFLUYENTES)	No detecta

El poder predictivo del modelo se evalúa tanto en los datos con los que fue construido el modelo como en conjuntos correspondientes a períodos diferentes e individuos diferentes. Al comparar estos resultados con los obtenidos a partir de la transformación logit se concluye que se consiguen mejores resultados de predicción para este último caso, en especial en el grupo B. Esto no significa que deba descartarse esta alternativa, por el contrario, s considera conveniente profundizar en la comparación y evaluación de los resultados en futuras investigaciones.

Cuadro Nº : Predicción del modelo con transformación logarítmica, sample testing y robustez respecto al momento del tiempo en el que se construye

GRUPO A	Error medio($\bar{e}$)	$\mu_{DC,(t+p)}$	I.P.	APORTE	$(\bar{e}/\mu_{DC,(t+p)})$	$\sigma_{DC,(t+p)}$	$\sigma_{DC,(t+p)}/\bar{e}$
PREDICCIÓN (2007-2005-2003)	0.054	0.072	0.017	24%	12%	3.1	0.054
SAMPLE TESTING	0.050	0.071	0.021	30%	11%	3.4	0.050
2006-2004-2002	0.057	0.076	0.019	24%	11%	2.7	0.057
2005-2003-2001	0.066	0.078	0.012	16%	14%	3.2	0.066
2004-2002-2000	0.178	0.151	-0.027	-18%	29%	1.6	0.178
GRUPO B	Error medio($\bar{e}$)	$\mu_{DC,(t+p)}$	I.P.	APORTE	$(\bar{e}/\mu_{DC,(t+p)})$	$\sigma_{DC,(t+p)}$	$\sigma_{DC,(t+p)}/\bar{e}$
PREDICCIÓN	0.034	0.063	0.029	45%	5%	5.2	0.034

(2007-2005-2003)							
SAMPLE TESTING	0.040	0.062	0.021	34%	7%	4.7	0.040
2006-2004-2002	0.040	0.045	0.005	12%	5%	5.5	0.040
2005-2003-2001	0.047	0.072	0.025	34%	7%	3.9	0.047
2004-2002-2000	0.091	0.119	0.028	24%	15%	2.1	0.091
GRUPO C	Error medio($\bar{e}$)	$\mu_{DC,(t+p)}$	I.P.	APORTE	$(\bar{e}/\mu_{DC,(t+p)})$	$\sigma_{DC,(t+p)}$	$\sigma_{DC,(t+p)}/\bar{e}$
PREDICCIÓN (2007-2005-2003)	0.025	0.026	0.001	5%	4%	13.5	0.025
SAMPLE TESTING	0.025	0.026	0.001	4%	4%	13.3	0.025
2006-2004-2002	0.033	0.033	0.000	1%	6%	11.0	0.033
2005-2003-2001	0.029	0.030	0.001	5%	4%	11.8	0.029
2004-2002-2000	0.046	0.048	0.002	5%	7%	7.5	0.046

Cuadro Nº: Predicción del modelo con transformación logarítmica, sample testing y robustez respecto al momento del tiempo en el que se construye

GRUPO A	Error medio($\bar{e}$)	$\mu_{DC,(t+p)}$	I.P.	APORTE	$(\bar{e}/\mu_{DC,(t+p)})$	$\sigma_{DC,(t+p)}$	$\sigma_{DC,(t+p)}/\bar{e}$
PREDICCIÓN (2007-2005-2003)	0.054	0.072	0.017	24%	12%	3.1	0.054
SAMPLE TESTING	0.052	0.068	0.016	23%	11%	3.3	0.052
2006-2004-2002	0.058	0.076	0.018	24%	11%	2.7	0.058
2005-2003-2001	0.065	0.078	0.013	17%	14%	3.2	0.065
2004-2002-2000	0.170	0.151	-0.019	-13%	27%	1.7	0.170
GRUPO B	Error medio($\bar{e}$)	$\mu_{DC,(t+p)}$	I.P.	APORTE	$(\bar{e}/\mu_{DC,(t+p)})$	$\sigma_{DC,(t+p)}$	$\sigma_{DC,(t+p)}/\bar{e}$
PREDICCIÓN (2007-2005-2003)	0.032	0.063	0.032	50%	5%	5.6	0.032
SAMPLE TESTING	0.036	0.070	0.034	48%	6%	5.2	0.036
2006-2004-2002	0.034	0.045	0.018	25%	4%	6.4	0.034
2005-2003-2001	0.043	0.072	0.029	41%	6%	4.3	0.043
2004-2002-2000	0.089	0.119	0.030	25%	15%	2.2	0.089
GRUPO C	Error medio($\bar{e}$)	$\mu_{DC,(t+p)}$	I.P.	APORTE	$(\bar{e}/\mu_{DC,(t+p)})$	$\sigma_{DC,(t+p)}$	$\sigma_{DC,(t+p)}/\bar{e}$
PREDICCIÓN (2007-2005-2003)	0.024	0.026	0.002	8%	3%	13.9	0.024
SAMPLE TESTING	0.025	0.028	0.003	9%	4%	13.0	0.025

2006-2004-2002	0.033	0.033	0.000	0%	6%	10.9	0.033
2005-2003-2001	0.028	0.030	0.002	6%	4%	12.0	0.028
2004-2002-2000	0.045	0.048	0.003	6%	7%	7.6	0.045

Por último se presentan los resultados de predicción de modelos correspondientes a diferentes períodos de tiempo para transformaciones logísticas y logarítmicas. La predicción se realiza sobre los mismos datos con los que se construye el modelo, los resultados para sample testing y evaluación de estacionariedad fueron presentados en los cuadros ….. y ….. para el caso de referencia. El error de predicción es menor con la transformación logarítmica en el grupo A, en el resto se logra una mejor performance con la transformación logit.

Cuadro Nº : Predicción de modelos con transformación logarítmica

GRUPO A	Error medio($\overline{e}$)	$\mu_{DC,(t+p)}$	I.P.	APORTE	$(\overline{e}/\mu_{DC,(t+p)})$	$\sigma_{DC,(t+p)}$	$\sigma_{DC,(t+p)}/\overline{e}$
2007-2005-2003	0.054	0.072	0.017	24%	12%	3.1	0.054
2006-2004-2002	0.055	0.072	0.017	24%	11%	3.1	0.055
2005-2003-2001	0.019	0.137	0.118	86%	4%	7.3	0.019
2004-2002-2000	0.013	0.166	0.153	92%	2%	6.3	0.013
GRUPO B	Error medio($\overline{e}$)	$\mu_{DC,(t+p)}$	I.P.	APORTE	$(\overline{e}/\mu_{DC,(t+p)})$	$\sigma_{DC,(t+p)}$	$\sigma_{DC,(t+p)}/\overline{e}$
2007-2005-2003	0.034	0.063	0.029	45%	5%	5.2	0.034
2006-2004-2002	0.038	0.046	0.009	19%	5%	5.3	0.038
2005-2003-2001	0.040	0.081	0.041	50%	6%	4.1	0.040
2004-2002-2000	0.085	0.247	0.162	66%	15%	1.7	0.085
GRUPO C	Error medio($\overline{e}$)	$\mu_{DC,(t+p)}$	I.P.	APORTE	$(\overline{e}/\mu_{DC,(t+p)})$	$\sigma_{DC,(t+p)}$	$\sigma_{DC,(t+p)}/\overline{e}$
2007-2005-2003	0.025	0.026	0.001	5%	4%	13.5	0.025
2006-2004-2002	0.034	0.032	-0.002	-5%	6%	10.6	0.034
2005-2003-2001	0.028	0.70	0.001	3%	4%	0.34	11.9
2004-2002-2000	0.055	0.058	0.003	5%	8%	6.0	0.055

Cuadro Nº : Predicción para modelos con transformación logit

GRUPO A	Error medio($\bar{e}$)	$\mu_{DC,(t+p)}$	I.P.	APORTE	$(\bar{e}/\mu_{DC,(t+p)})$	$\sigma_{DC,(t+p)}$	$\sigma_{DC,(t+p)}/\bar{e}$
2007-2005-2003	0.054	0.072	0.017	24%	12%	3.1	0.054
2006-2004-2002	0.055	0.072	0.017	24%	11%	3.2	0.055
2005-2003-2001	0.026	0.137	0.111	81%	5%	5.3	0.026
2004-2002-2000	0.015	0.166	0.151	91%	2%	5.5	0.015
GRUPO B	Error medio($\bar{e}$)	$\mu_{DC,(t+p)}$	I.P.	APORTE	$(\bar{e}/\mu_{DC,(t+p)})$	$\sigma_{DC,(t+p)}$	$\sigma_{DC,(t+p)}/\bar{e}$
2007-2005-2003	0.032	0.063	0.032	50%	5%	5.6	0.032
2006-2004-2002	0.034	0.046	0.013	28%	4%	5.9	0.034
2005-2003-2001	0.036	0.081	0.045	55%	6%	4.6	0.036
2004-2002-2000	0.087	0.247	0.160	65%	15%	1.6	0.087
GRUPO C	Error medio($\bar{e}$)	$\mu_{DC,(t+p)}$	I.P.	APORTE	$(\bar{e}/\mu_{DC,(t+p)})$	$\sigma_{DC,(t+p)}$	$\sigma_{DC,(t+p)}/\bar{e}$
2007-2005-2003	0.024	0.026	0.002	8%	3%	13.9	0.024
2006-2004-2002	0.032	0.032	0.001	2%	5%	11.4	0.032
2005-2003-2001	0.030	0.032	0.002	7%	4%	11.1	0.030
2004-2002-2000	0.051	0.058	0.008	13%	8%	6.6	0.051

ANEXO 6: IMPLEMENTACIÓN EN R[18](PSEUDO CÓDIGO)

```
library(Rcmdr)
library(MASS)
library(stats4)
library(tree)

#-------------------------------------------------------------------------------------------------#
#CON EL ÁRBOL IDENTIFICO LOS PUNTOS DE CORTE "negativo" Y "positivo" PARA GENERA LOS GRUPOS
arbol=tree(dif_futura~dif_pasada,base, control= tree.control(dim(base)[1],mincut = 5, minsize = 20, mindev = 0.002))
plot(arbol)
text(arbol)

#-------------------------------------------------------------------------------------------------#
####LA FUNCIÓN CALCULA EL MODELO Y EVALÚA EL ERROR DE PREDICCIÓN PARA LOS DATOS "data"####
#LA SALIDA DE LA FUNCIÓN ES: EL MODELO, EL ERROR DE PREDICCIÓN SI IMPUTARA COMO OBSERVACIÓN FUTURA EL ÚLIMO
#DATO DISPONIBLE (E1), EL ERROR DE PREDICCIÓN CON LOS RESULTADOS DEL MODELO (E2),EL COCIENTE ENTRE EL DESVÍO DE LA
#VARIABLE EXPLICADA Y E2, Y EL COCIENTE ENTRE E2 Y LA MEDIA DE LA VARIABLE EXPLICADA

MODELO=function(data)
{
m=lm(Ldc_fut~Ldc_pas+franjaCat_pas+giroCat+edad_pas,data)
print(summary(m))

pre=predict(m,data)              #PREDICCIÓN
mu=exp(pre)/(1+exp(pre))         #DESPEJO LA MEDIA SABIENDO QUE MODELÉ EL LOGIT DE LA MEDIA

E1=mean(abs(data$dc_fut-data$dc_pas)) #CALCULO EL ERROR CUANDO SE USA COMO PREDICCIÓN FUTURA EL DATO MÁS RECIENTE DED MODELO
E2=mean(abs(data$dc_fut-mu))     #CALCULO EL ERROR MEDIO QUE TENGO CUANDO USO LA PREDICCIÓN

##COMPARO LOS ERROR CON EL DESVÍO Y LA MEDIA DE LA VARIABLE INDEPENDIENTE##
```

[18] R version 2.7.1 (2008-06-23) Copyright© 2008 The R Foundation for Statistical Computing

```
  ##LOS USO COMO INDICADRES RELATIVOS PARA COMPARAR DISTINTOS MODELOS##
sd(data$dc_fut)
mean(data$dc_fut)
IND_DESVÍO=sd(data$dc_fut)/E2
IND_MEDIA=E2/mean(data$dc_fut)
return(c(m,E1,E2,IND_DESVÍO,IND_MEDIA,mean(mu)))
}

#-------------------------------------------------------------------------------------------------------#
                        ###SAMPLE TESTING###
#PARA VALIDACIÓN CRUZADA GENERO EL MODELO CON UN 70% DE LOS CASOS EN LA BASE, CON EL 30% EVALÚO PODER PREDICTIVO
##SORTEO UNA MUESTRA UNIFORME PARA PARTICIONAR AL BASE##

SAMPLE_TEST=function(datos)
{
tamaño=dim(datos)[1]#cantidad de afiliados en la base
posiciones=trunc(runif(1,1,tamaño))#inicializo el vector con un elemento
while(length(posiciones)<tamaño*.70)
{
u=trunc(runif(1,1,tamaño))
#chequeo que el valor sorteado no esté en el vactor, si está repetido no lo almaceno
if (!any (posiciones==u))
   posiciones=append(posiciones,u)
}

m=lm(Ldc_fut~Ldc_pas+franjaCat_pas+giroCat+edad_pas,datos[-pos,])

pre=predict(m,datos[-posiciones,]) #PREDICCIÓN
mu=exp(pre)/(1+exp(pre))        #DESPEJO LA MEDIA SABIENDO QUE MODELÉ EL LOGIT DE LA MEDIA

E1=mean(abs(datos[-posiciones,]$dc_fut-datos[-posiciones,]$dc_pas)) #CALCULO EL ERROR CUANDO SE USA COMO PREDICCIÓN FUTURA EL DATO MÁS RECIENTE DED MODELO
E2=mean(abs(datos[-posiciones,]$dc_fut-mu))        #CALCULO EL ERROR MEDIO QUE TENGO CUANDO USO LA PREDICCIÓN

##COMPARO LOS ERROR CON EL DESVÍO Y LA MEDIA DE LA VARIABLE INDEPENDIENTE##
  ##LOS USO COMO INDICADRES RELATIVOS PARA COMPARAR DISTINTOS MODELOS##
sd(datos$dc_fut)
mean(datos$dc_fut)
IND_DESVÍO=sd(datos$dc_fut)/E2
IND_MEDIA=E2/mean(datos$dc_fut)
```

```
return(c(E1,E2,IND_DESVÍO,IND_MEDIA))
}
#-----------------------------------------------------------------------------------------------------#
              #DADO UN MODELO, EVALÚO EL PODER PREDICTIVO CON DATOS DE OTRO PERÍODO
ESTACIONARIEDAD=function(modelo,datos)
{
pre=predict(modelo,datos) #PREDICCIÓN
mu=exp(pre)/(1+exp(pre))        #DESPEJO LA MEDIA SABIENDO QUE MODELÉ EL LOGIT DE LA MEDIA
E1=mean(abs(datos$dc_fut-datos$dc_pas)) #CALCULO EL ERROR CUANDO SE USA COMO PREDICCIÓN FUTURA EL DATO MÁS RECIENTE DED MODELO
E2=mean(abs(datos$dc_fut-mu))            #CALCULO EL ERROR MEDIO QUE TENGO CUANDO USO LA PREDICCIÓN

##COMPARO LOS ERROR CON EL DESVÍO Y LA MEDIA DE LA VARIABLE INDEPENDIENTE##
  ##LOS USO COMO INDICADRES RELATIVOS PARA COMPARAR DISTINTOS MODELOS##
sd(datos$dc_fut)
mean(datos$dc_fut)
IND_DESVÍO=sd(datos$dc_fut)/E2
IND_MEDIA=E2/mean(datos$dc_fut)
return(c(E1,E2,IND_DESVÍO,IND_MEDIA))
}
#-----------------------------------------------------------------------------------------------------#

                              # OPTIMIZACIÓN #
 ###LOG-VERSOIMILITUD###
  db=function(phi=20)
   sum(log((gamma(phi)*y^(m*phi-1)*(1-y)^((1-m)*phi-1))/(gamma(m*phi)*gamma((1-m)*phi))))

#aproximación de striling
# n!~(n^n)*exp(-n)*sqrt(2*pi*n)

#LOGARITMO NATURAL DE LA FUNCIÓN DE STIRLING
log_sti=function(n)
n*log(n)-n+(1/2)*log(2*pi*n)

#APROXIMACIÓN NUMÉRICA DE LA LOG VEROSIMILITUD
dbnum=function(phi=20)
sum(log_sti(phi-1)-log_sti(m*phi-1)-log_sti((1-m)*phi-1)+(m*phi-1)*log(y)+((1-m)*phi-1)*log(1-y)
tope=171;#límite superior para optimizar usando función gamma-(LIMITACIÓN: función gamma calcula hasta n=171 (i.e.=(170)!)
```

```
#ALGORITMO PARA OBTENER EL MÁXIMO DE LA VEROSIMILITUD
#LOS PARÁMETROS DE ENTRADA SON LA MUESTRA Y LA MEDIA DE LA DISTRIBBUCIÓN
BETA

MAXIMO=function(observaciones,media)
{m=media
 y=observaciones
Opt=optimize(db ,lower =0.0000001, upper=tope,maximum=TRUE)
if ((as.numeric(Opt)[1]>(tope-1))|(is.na(as.numeric(Opt)[2])) )
Opt=optimize(dbnum ,lower =tope, upper=1000,maximum=TRUE)
return (as.numeric(Opt)[1])
 }

#-------------------------------------------------------------------------------------------------------#
##ejemplo para 2003-2005-2007##
base$dif_pasada=base$tasa53    #base$dc05-base$dc03
base$dif_futura=base$tasa75    #base$dc07-base$dc05


#LOS COEFICIENTES OBTENIDOS CON EL ÁRBOL
                        negativo=-0.072
                        positivo=0.047


#LOS DATOS
data=data.frame(matrix(1,dim(base)[1],7))
names(data)=c('Ldc_fut','Ldc_pas','franjaCat_pas','giroCat','edad_pas','dc_fut','dc_pas'
)
        data$Ldc_fut=base[,"Ldc07"]
        data$Ldc_pas=base[,"Ldc05"]
        data$dc_fut=base[,"dc07"]
        data$dc_pas=base[,"dc05"]
        data$franjaCat_pas=base[,"franja05cat"]
        data$giroCat=base[,"GIROCAT"]
        data$edad_pas=base[,"edad05"]
#DATOS PARA ESTACIONARIEDAD
dataEst=data.frame(matrix(1,dim(base)[1],7))
names(dataEst)=c('Ldc_fut','Ldc_pas','franjaCat_pas','giroCat','edad_pas','dc_fut','dc_p
as')
        dataEst$Ldc_fut=base[,"Ldc05"]
        dataEst$Ldc_pas=base[,"Ldc03"]
        dataEst$dc_fut=base[,"dc05"]
        dataEst$dc_pas=base[,"dc03"]
        dataEst$franjaCat_pas=base[,"franja03cat"]
        dataEst$giroCat=base[,"GIROCAT"]
        dataEst$edad_pas=base[,"edad03"]
        base$dif_pasadaEst=base$dc03-base$dc01
```

```
p_neg=which(base$dif_pasada<=negativo)      #POSICIONES DE AQUELLOS AFILIADOS QUE TIENEN EVOLUCIÓN NEGATIVA
p_pos=which(base$dif_pasada>=positivo)      #POSICIONES DE AQUELLOS AFILIADOS QUE TIENEN EVOLUCIÓN POSITIVA
aux=base[which(base$dif_pasada<positivo),]
p_estable=which(aux$dif_pasada>negativo)    #POSICIONES DE AQUELLOS AFILIADOS SIN CAMBIO SIGNIFICATIVO

##MATRIZ PARA GUARDAR ERRORES E INDICADORES##
PERFORMANCE=data.frame(matrix(1,12,3))
row.names(PERFORMANCE)=c('e1','e2','id','im','e1_st','e2_st','id_st','im_st','e1_est','e2_est','id_est','im_est')
names(PERFORMANCE)=c('decrece','crece','estable')

##MATRIZ PARA GUARDAR PARÁMETROS##
PARAMETROS=data.frame(matrix(1,2,3))
row.names(PARAMETROS)=c('MU','PHI')
names(PARAMETROS)=c('decrece','crece','estable')

            ############APLICACIÓN##############

#########################CAMBIO NEGATIVO#####################################
#----------------------#modelo
M=NULL
M=MODELO(data[p_neg,])
datamodelo=M[1:13]
class(datamodelo)="lm"
PERFORMANCE[1,1]=M[14]
PERFORMANCE[2,1]= M[15]
PERFORMANCE[3,1]=M[16]
PERFORMANCE[4,1]=M[17]
media=M[18]
#----------------------#estimación por máxima verosimilitud
y=data[p_estable,"dc_fut"]
opt=MAXIMO(y,media)
PARAMETROS[,1]=as.numeric(c(media,opt))
#----------------------#sample test
ST=SAMPLE_TEST(data[p_neg,])
PERFORMANCE[5,1]=ST[1]
PERFORMANCE[6,1]= ST[2]
PERFORMANCE[7,1]=ST[3]
PERFORMANCE[8,1]=ST[4]
#----------------------#estacionariedad
p=which(base$dif_pasadaEst<=negativo)
EST=ESTACIONARIEDAD(datamodelo,dataEst[p,])
```

```
PERFORMANCE[9,1]=EST[1]
PERFORMANCE[10,1]= EST[2]
PERFORMANCE[11,1]=EST[3]
PERFORMANCE[12,1]=EST[4]
########################CAMBIO
POSITIVO###################################
#----------------------#modelo
M=NULL
M=MODELO(data[p_pos,])
datamodelo=M[1:13]
class(datamodelo)="lm"
PERFORMANCE[1,2]=M[14]
PERFORMANCE[2,2]=m[15]
PERFORMANCE[3,2]=M[16]
PERFORMANCE[4,2]=M[17]
media=M[18]
#----------------------#estimación por máxima verosimilitud
y=data[p_estable,"dc_fut"]
opt=MAXIMO(y,media)
PARAMETROS[,2]=as.numeric(c(media,opt))
#----------------------#sample test
ST=SAMPLE_TEST(data[p_pos,])
PERFORMANCE[5,2]=ST[1]
PERFORMANCE[6,2]= ST[2]
PERFORMANCE[7,2]=ST[3]
PERFORMANCE[8,2]=ST[4]
#----------------------#estacionariedad
p=which(base$dif_pasadaEst>=positivo)
EST=ESTACIONARIEDAD(datamodelo,dataEst[p,])
PERFORMANCE[9,2]=EST[1]
PERFORMANCE[10,2]= EST[2]
PERFORMANCE[11,2]=EST[3]
PERFORMANCE[12,2]=EST[4]

############################## SIN
CAMBIO###########################################
#----------------------#modelo
M=NULL
M=MODELO(data[p_estable,])
datamodelo=M[1:13]
class(datamodelo)="lm"
PERFORMANCE[1,3]=M[14]
PERFORMANCE[2,3]=m[15]
PERFORMANCE[3,3]=M[16]
PERFORMANCE[4,3]=M[17]
media=M[18]
#----------------------#estimación por máxima verosimilitud
```

```
y=data[p_estable,"dc_fut"]
opt=MAXIMO(y,media)
PARAMETROS[,3]=as.numeric(c(media,opt))
#----------------------#sample test
ST=SAMPLE_TEST(data[p_estable,])
PERFORMANCE[5,3]=ST[1]
PERFORMANCE[6,3]= ST[2]
PERFORMANCE[7,3]=ST[3]
PERFORMANCE[8,3]=ST[4]
#----------------------#estacionariedad
aux=which(base$dif_pasadaEst<positivo)
p=which(aux$dif_pasadaEst>negativo)
EST=ESTACIONARIEDAD(datamodelo,dataEst[p,])
PERFORMANCE[9,3]=EST[1]
PERFORMANCE[10,3]= EST[2]
PERFORMANCE[11,3]=EST[3]
PERFORMANCE[12,3]=EST[4]
#-------------------------------------------------------------------------------------------------#
```