PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
ESCUELA DE INGENIERIA

VALORIZACIÓN DE INSTRUMENTOS FINANCIEROS CON MODELOS DINÁMICOS DE NO ARBITRAJE Y ALGORITMOS DE MINERÍA DE DATOS.

ANDRÉS MATÍAS CASTRO ROJAS

Tesis para optar al grado de
Magister en Ciencias de la Ingeniería

Profesor Supervisor:
GONZALO CORTAZAR

Santiago de Chile, (Diciembre, 2010)

PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
ESCUELA DE INGENIERIA

VALORIZACIÓN DE INSTRUMENTOS FINANCIEROS CON MODELOS DINÁMICOS DE NO ARBITRAJE Y ALGORITMOS DE MINERÍA DE DATOS.

ANDRÉS MATÍAS CASTRO ROJAS

Tesis presentada a la Comisión integrada por los profesores:

GONZALO CORTÁZAR

JOSÉ PRINA

AUGUSTO CASTILLO

DOMINGO MERY

Para completar las exigencias del grado de
Magister en Ciencias de la Ingeniería

Santiago de Chile, (Diciembre, 2010)

Para toda mi familia y todos mis amigos.

AGRADECIMIENTOS

Me gustaría agradecer en primer lugar a mi profesor guía Gonzalo Cortázar por el tiempo y dedicación invertidos en esta investigación. Sus aportes fueron indispensables para que pudiera ser llevada a cabo exitosamente.

También me gustaría agradecer a toda mi familia no sólo por el apoyo incondicional que me dieron durante este período de tiempo, sin el cual todo hubiese sido sumamente más difícil, sino que, por todo lo que me han dado a lo largo de mi formación personal y académica.

También quisiera agradecer a todos mis compañeros y amigos del FinLab UC, los cuales siempre estuvieron dispuestos a ayudarme, escucharme y aportar cuando más lo necesité. Ellos fueron un aporte fundamental.

Finalmente, quisiera agradecer el apoyo financiero proporcionado por Fondecyt (Proyecto 1100597).

INDICE GENERAL

Pág.

INDICE DE TABLAS

Pág.

INDICE DE FIGURAS

Pág.

RESUMEN

A lo largo de los años, los mercados financieros han extendido su alcance para satisfacer nuevas necesidades cada vez más sofisticadas y complejas. Hoy en día no sólo cubren formas de deuda gubernamentales o privadas sino que permiten a las personas financiar proyectos personales, transar riesgos o invertir en mercados de *commodities* entre muchos otros. Esto ha motivado la creación de modelos cada vez más sofisticados que puedan satisfacer la necesidad de determinar precios justos para cada uno de estos instrumentos.

Esta investigación propone una metodología tanto para encontrar precios de instrumentos financieros como para obtener información del mercado en que éstos se transan. La valorización se realiza en dos etapas: La primera usa un modelo dinámico de tasas de interés para encontrar el comportamiento promedio del mercado, mientras que la segunda, un modelo estático de minería de datos para encontrar los *spreads* específicos de cada papel sobre el promedio. En particular, el algoritmo utilizado en esta segunda etapa es el de árboles de regresión, el cual no sólo ofrece una estimación de los *spreads* sino que, dada su estructura de ramas de decisión, permite extraer información del mercado en general.

El modelo es aplicado en el mercado Chileno de letras de crédito hipotecario, instrumentos que presentan opciones implícitas. Los resultados muestran que se pueden obtener buenas estimaciones de precios para estos instrumentos, y al mismo, tiempo

información de cómo el mercado castiga o premia estos precios dependiendo de las características propias de cada uno.

Palabras claves: Minería de datos, árboles de regresión, Modelos dinámicos, Filtro de Kalman, letras de crédito hipotecario, spreads.

ABSTRACT

Throughout the years, financial markets have extended their scope to satisfy new more sophisticated and complex needs. Today, not only they cover government or private debt, but also allow people to finance personal projects, trade risks or invest in commodities among many others. This has motivated the creation of new more sophisticated models that can satisfy the necessity of determining fair prices for each of these instruments.

This research proposes a methodology both for determining fair prices for financial instruments and obtaining information about the market they are traded in. valorization is done in two stages: the first one uses a dynamic interest rate model to find the average behavior of the market, while the second one uses a static data mining model to find the specific spread over that average. In particular, the algorithm used is Regression Trees, which not only offers spreads estimates but also, due to its branched decision structure, allows to extract information of the market in general.

The model is applied to the Chilean mortgage market, instruments which present implicit options. Results show that good price estimates can be obtained and, at the same time, information regarding how the market punishes or rewards these prices depending on each instrument's own characteristics.

Key words: Data mining, Regression Trees, dinamic models, Kalman filter, mortgage market, spreads.

1 INTRODUCCIÓN

Los mercados financieros hace mucho tiempo comenzaron a tomar fuerza hasta llegar a lo que son hoy en día. Básicamente, buscan introducir eficiencia en los mercados reales tratando de juntar agentes provistos de liquidez que quieran invertirla con agentes con necesidad de financiamiento para la realización de proyectos de creación de valor. Un claro ejemplo de esto es una empresa que emite deuda para poder financiar una cartera de proyectos que la puede hacer crecer. Luego, el mercado financiero toma esta emisión y la ofrece a algún segundo agente en cualquier parte del mundo, el cual recibe esa promesa de pagos a cambio de liquidez para la empresa.

Estos mercados se han extendido desde las deudas gubernamentales hasta nuevos y complicados instrumentos que buscan satisfacer las necesidades de financiamiento de inversionistas cada vez más específicos. Se extienden también a las deudas privadas, de financiamiento hipotecario y a la industria de la minería o de los commodities entre muchos otros.

En base a esto se puede afirmar que éstos mercados tienen una gran importancia en los mercados reales, por lo que es de gran importancia también encontrar los precios justos para cada uno de los instrumentos que están presentes en los mercados financieros. Esto es así debido a que un precio mal puesto tiene como consecuencia la transferencia no justificada de valor entre distintos agentes del mercado, o el desincentivo a la realización de proyectos que de otra forma sí se realizarían.

Otra razón por la que es importante encontrar precios correctos de los instrumentos dentro del mercado financiero es que no sólo las empresas forman parte de estos mercados sino que, indirectamente, las personas también por medio de instituciones de inversiones (Administradoras de fondos de pensiones, fondos mutuos, etc.) que toman sus ahorros y los usan para realizar inversiones en estos mercados y ofrecer así rentabilidades en base al movimiento de los precios de los instrumentos.

Esfuerzos para lograr el objetivo de encontrar precios justos para los instrumentos financieros pueden encontrarse, por ejemplo, en Nelson & Siegel (1987) o Svensson (1994) quienes asumen una estructura parsimoniosa de tasa de interés la cual depende de un número limitado de parámetros. Estos modelos son de carácter estático, vale decir, no usan información histórica sino que sólo la actual, lo que los hace obtener buenos resultados en mercados con alto número de transacciones, pero no así en con mercados de bajo número de transacciones.

También están los modelos dinámicos los cuales estudian el movimiento estocástico de la estructura de tasas de interés. Algunos de estos trabajos se pueden encontrar en Vasicek (1977) y Brennan & Schwartz (1979) quienes usan uno y dos factores estocásticos respectivamente, o Cortazar, Schwartz, & Naranjo (2007) quienes plantean un modelo Vasicek generalizado para la determinación de la estructura de tasas de interés aplicándolo al mercado financiero Chileno.

Por otro lado, si tomamos las emisiones de deuda privada podemos encontrar en la literatura a Merton (1974), Longstaff & Schwartz (1992) quienes asumen que para describir el riesgo crediticio hay que considerar a los bonos y acciones de una firma como

derechos contingentes sobre sus activos, o Cortazar, Schwartz, & Tapia (2010) quienes modelan el *spread* de una firma sobre la curva libre de riesgo realizando agrupaciones en base a características similares de ellas.

Los mercados financieros también se extienden a los mercados mundiales de *commodities*, los cuales reflejan la demanda y oferta mundial de estos productos en base a las características propias de cada uno de éstos. Estos mercados son usados por empresas insertas en esos rubros para poder, por ejemplo, vender producciones futuras, hacer estimaciones de sus ingresos, etc. Luego, esfuerzos para encontrar el precio correcto pueden encontrarse en Brennan & Schwartz (1985) quienes intentan especificar el precio en base a un factor estocástico o Cortazar & Naranjo (2006) que generalizan el modelo a uno N-factorial.

Por otro lado, otra potencialidad de los mercados financieros tiene que ver con la distribución del riesgo entre los agentes que participan del mercado mediante los instrumentos derivados. Estos instrumentos tienen flujos que van a depender de un subyacente, como puede ser por ejemplo el precio de un *commodity* o el tipo de cambio entre dos monedas. La idea básicamente es que un agente que no quiere estar expuesto a los movimientos de, por ejemplo, la tasa de interés, puede realizar un contrato con otro agente de forma que a cambio de un pago inicial, este último se comprometa a responder en los escenarios malos. En la literatura, uno de los estudios más importante para encontrar el precio correcto de este tipo de contratos, llamados opciones, es el de Black & Scholes (1973).

Es por esto que esta investigación se ubica en la línea de la especificación de los precios de instrumentos financieros, tomando como objeto de estudio uno en particular dentro del mercado Chileno: las Letras Hipotecarias, las cuales presentan algunas dificultades propias que hacen que este objetivo sea especialmente dificultoso y nuevas técnicas y alternativas sean abordadas. En particular, la más importante es la presencia de una opción de prepago, la cual significa que el emisor de uno de estos títulos de deuda tiene la opción de pagar por adelantado lo que falta ahorrándose los intereses que se devengarían.

Esta investigación pretende descomponer el precio de un instrumento en distintos aportes. Primero, el que hace la estructura libre de riesgo compuesta por los bonos emitidos por el gobierno, ya sea por el Banco Central o la Tesorería General de la República. Luego, el aporte que hace la familia de papeles como un conjunto por sobre el aporte libre de riesgo y finalmente, el aporte hecho por las características propias del instrumento que se está transando, por sobre o bajo el de la familia de papeles.

De esta forma esta investigación se enmarca en los trabajos de Cortazar, Schwartz, & Naranjo (2007) quienes estudian la especificación de la estructura de tasas de interés de los bonos libres de riesgo en un mercado de bajas transacciones como el caso de Chile, en conjunto con el de Cortazar, Schwartz, & Tapia (2010) quienes buscan una especificación de los *spreads* de papeles con riesgo por sobre la estructura libre de riesgo, agrupándolos en familias de riesgo similares.

Estos dos trabajos dejan un *spread* remanente sin especificar, el cual se explica principalmente por las características propias de cada instrumento. Este remanente es el

que se desea estudiar mediante algoritmos de minería de datos, de forma de poder encontrar una forma de estimarlos correctamente y al mismo tiempo aprender la forma en que estos se especifican. La forma en que se busca lograrlo es usando algoritmos de árboles de regresión, siguiendo la metodología introducida por Breiman et al. (1984), alimentados por características propias de cada papel. La motivación para proceder de esta manera está en que dichos algoritmos tienen una alta flexibilidad y poder de adaptación a cualquier set de datos. Por otro lado, esta técnica de minería de datos ofrece una intuición sobre las variables que son importantes y cómo deben ser tomadas en cuenta al momento de la especificación del remanente sin especificar.

Este último objetivo puede encontrarse parcialmente en la literatura en Collin-Dufresne, Goldstein, & Martin (2001) donde se asume que las variables son el riesgo de *default* y la tasa de recuperación en aquel caso, logrando no muy buen ajuste, Elton et al. (2001) quienes asumen que las variables son las expectativas de pérdida en caso de *default*, el premio por impuestos y el premio por riesgo, logrando un mejor ajuste. También Dungey, Martin, & Pagan (2000) en el cual se definen de factores latentes que afectan a todos los papeles por igual, a todos los papeles de distintas maneras y que afectan sólo a un papel independientemente.

No obstante lo anterior, para lo mejor de nuestro conocimiento, ningún trabajo intenta encontrar sin definir, a priori, los determinantes más importantes a la hora de encontrar la estructura de tasas de interés de algún mercado en particular, usando características propias de los papeles estudiados.

Este documento se organiza de la siguiente manera. El capítulo 2 se describe el estado del arte en la modelación de estructuras de tasas de interés. El capítulo 3 resume la evolución de la minería de datos, algunos de sus algoritmos y una descripción de aquel que se usará en esta investigación. El capítulo 4 describe las principales características del mercado de renta fija chileno. El capítulo 5 plantea el modelo propuesto para la estimación de los remanentes no explicados. El capítulo 6 muestra los resultados de la aplicación del modelo al mercado chileno. El capítulo 7 finalmente concluye.

2 MARCO TEÓRICO

El modelo propuesto en esta investigación intenta descomponer el precio de la transacción de una Letra Hipotecaria en tres partes. A partir de ahora, en vez de precio hablaremos de la TIR o tasa interna de retorno, la cual hace que la suma de los pagos descontados del instrumento sean igual al precio de este. La primera corresponde al aporte que hace la estructura de tasas de interés que se obtiene a partir de los bonos libres de riesgo, tanto del Banco Central como de la Tesorería de la República. Esta sirve como estructura base para comenzar el análisis. Luego, sobre esta estructura, se agrega un spread correspondiente a una familia de papeles. Este aporte extra representa el riesgo extra que tiene un papel que no está respaldado por el gobierno o que tiene otras características. En este caso, la familia que se desea analizar (Letras Hipotecarias) tienen asociadas opciones sobre sus pagos, las cuales también influyen en el spread de ésta.

A pesar de que el modelo que se propone incorpora también un tercer aporte, el cual va a depender de las características propias de los papeles dentro de la familia, en esta parte de la tesis se va a profundizar en los dos primeros, ya que la parte final se discutirá en el capítulo 3 de minería de datos.

El modelo se puede resumir según la ecuación (2-1), donde TIR_{LH} es la TIR de la transacción, TIR_{Gob} la TIR que aporta la estructura libre de riesgo, $S_{Familia\ LH}$ el aporte de la familia de instrumentos de letras de crédito hipotecario, $S_{específico}$ el aporte específico que depende de cada transacción y ε la parte de la tasa que no queda explicada por el modelo.

$$TIR_{LH} = TIR_{Gob} + S_{Familia\ LH} + S_{específico} + \varepsilon \quad (2\text{-}1)$$

A continuación se discuten las diferentes alternativas que existen para definir la estructura libre de riesgo, seguido por la descripción de los trabajos más recientes relacionados con familias de papeles.

2.1 Estructura libre de riesgo

Para encontrar una estructura libre de riesgo existen dos corrientes principales, los modelos estáticos y los modelos dinámicos.

2.1.1 Modelos estáticos

Los principales modelos estáticos que existen en la literatura son Nelson & Siegel (1987) y Svensson (1994) en los cuales se asume una estructura parsimoniosa de tasa de interés que depende de un número limitado de parámetros, los cuales son encontrados período a período con las transacciones existentes. Este modelo logra un resultado exitoso en mercados profundos ya que en éstos se puede encontrar información de toda la curva, característica que no se encuentra en mercados de baja liquidez como lo es el mercado Chileno, tanto de bonos del Banco Central, Empresariales, Letras hipotecarias, etc.

En particular, Nelson & Siegel (1987) define la tasa *forward* para un tiempo T según la ecuación (2-2).

$$f(T) = \beta_0 + \beta_1 e^{-\frac{T}{\tau}} + \beta_2 \frac{T}{\tau} e^{-\frac{T}{\tau}}$$

(2-2)

En ésta, $f(T)$ es tasa forward, T el plazo, β_0, β_1, β_2 y τ son parámetros que deben encontrarse cada día.

Luego, esta tasa *forward* es integrada en el tiempo al vencimiento para obtener la estructura completa.

$$R(T) = \frac{1}{T} \int_0^T f(s)\, ds$$

(2-3)

El resultado de la integración de la tasa *forward* según la relación (2-3) está dado por la ecuación

$$R(T) = \beta_0 + \beta_1 \left(1 - e^{-\frac{T}{\tau}}\right) \frac{\tau}{T} + \beta_2 \left(\left(1 - e^{-\frac{T}{\tau}}\right) \frac{\tau}{T} - e^{-\frac{T}{\tau}}\right)$$

(2-4)

En esta relación, los parámetros β_0, β_1 y β_2 representan el corto, mediano y largo plazo, dándole así la posibilidad de ajustar diversas formas de estructuras.

Por otra parte, Svensson (1994) introduce un nuevo término en la relación (2-2), cuyo resultado se muestra en la ecuación (2-5).

$$f(T) = \beta_0 + \beta_1 e^{-\frac{T}{\tau_1}} + \beta_2 \frac{T}{\tau_1} e^{-\frac{T}{\tau_1}} + \beta_3 \frac{T}{\tau_2} e^{-\frac{T}{\tau_2}}$$

(2-5)

En ésta se puede observar que se agregan dos parámetros más, β_3 y τ_2, por lo que la estructura que se puede obtener a partir de esta especificación tiene más grados de libertad,

por ende tiene mayor facilidad de adaptarse a estructuras más complejas, pero tiene el potencial problema de agregar volatilidad no justificada al modelo.

Si se realiza el mismo ejercicio de integrar la tasa *forward* según la ecuación (2-3), se obtiene un nuevo resultado para la estructura completa, el cual está representado en la siguiente ecuación.

$$R(T) = \beta_0 + \beta_1\left(1 - e^{-\frac{T}{\tau_1}}\right)\frac{\tau_1}{T} + \beta_2\left(\left(1 - e^{-\frac{T}{\tau_1}}\right)\frac{\tau_1}{T} - e^{-\frac{T}{\tau_1}}\right) + \beta_3\left(\left(1 - e^{-\frac{T}{\tau_2}}\right)\frac{\tau_2}{T} - e^{-\frac{T}{\tau_2}}\right) \quad (2\text{-}6)$$

Dado que estos modelos sólo funcionan bien en mercados donde hay información a lo largo de toda la estructura, caso diferente al Chileno, nacen los modelos dinámicos, los cuales para resolver el problema usan la información de las transacciones del día en conjunto con la información histórica.

A continuación se discute el modelo utilizado en Cortazar, Schwartz, & Naranjo (2007).

2.1.2 Modelos dinámicos

Los modelos dinámicos son una respuesta al problema enunciado anteriormente. Éstos permiten estimar de forma conjunta tanto la estructura de tasas como su dinámica, utilizando toda la información histórica y actual que esté disponible. Algunos de ejemplos de este tipo de modelos se pueden encontrar en Vasicek (1977) y Brennan & Schwartz

(1979) quienes buscan encontrar la estructura de tasas de interés usando uno y dos factores estocásticos respectivamente, o Langetieg (1980) que extiende estos trabajos a uno multifactorial.

Por otro lado, esta investigación se enmarca en el trabajo de Cortazar, Schwartz, & Naranjo (2007) quienes proponen un modelo Vasicek generalizado para la tasa de interés que depende de N factores estocásticos x que siguen un proceso de reversión a una media de largo plazo.

$$r = 1^{r\prime} x^r + \delta \tag{2-7}$$

La dinámica de cada uno de los factores estocásticos x queda determinada por la expresión (2-8)

$$dx_t = -Kx_t dt + \Sigma dw_t \tag{2-8}$$

donde K y Σ son matrices diagonales, en las que cada elemento k_{ii} y σ_{ii} representa la velocidad de reversión a la media y varianza del factor x_i respectivamente. Por otro lado, el término dw_t representa un vector de movimientos Brownianos correlacionados tal que:

$$(dw)'(dw) = \Omega dt \tag{2-9}$$

donde cada elemento de la matriz Ω, ρ_{ij} representa la correlación entre los factores i y j.

Para lograr que la tasa resultante sea de tipo estacionario se pide como condición que cada uno de los elementos κ_i de la matriz K sean distintos de cero con el fin de evitar

raíces unitarias. De esta forma, la tasa r tiende a una media de largo plazo δ, ya que cada uno de los factores tiende a una media en el largo plazo igual a 0.

Finalmente asumiendo un precio de mercado del riesgo λ, el proceso ajustado por riesgo para los factores x_i queda expresado en la siguiente ecuación:

$$dx_t = -(\lambda + Kx_t)dt + \Sigma dw_t$$

(2-10)

Aplicando el Lema de Îto sobre la estructura planteada en (2-10), es posible encontrar la expresión cerrada para el valor de un bono de descuento (u otro instrumento cero cupón) en función de los factores que determinan la tasa, cuyo valor esta dado por las ecuaciones (2-11) y (2-12). La derivación de los términos $u(t)$ y $v(t)$ se muestra en el ANEXO A:.

$$P(x_t, \tau) = e^{(u(\tau)' x_t + v(\tau))}$$

(2-11)

$$u_i(\tau) = -\left(\frac{1 - e^{-k_i \tau}}{k_i}\right)$$

$$v(\tau) = \sum_{i=1}^{n} \frac{\lambda_i}{k_i}\left(\tau - \frac{1 - e^{-k_i \tau}}{k_i}\right) - \delta\tau + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\sigma_i \sigma_j \rho_{ij}}{k_i k_j}\left(\tau - \frac{1 - e^{-k_i \tau}}{k_i} - \frac{1 - e^{-k_j \tau}}{k_j} + \frac{1 - e^{-(k_i + k_j)\tau}}{k_i + k_j}\right)$$

(2-12)

Esta modelación es una representación canónica en el sentido de Dai & Singleton (2002), dado que posee el mínimo número de factores necesarios identificar el modelo.

Esto se logra luego de aplicar una transformación lineal $T(x)$ a las variables de estado, lo cual es factible dado que estas son variables no observables, luego, cualquier conjunto de ellas puede ser usado para explicar a la variable observable.

Otros modelos de tasas de interés necesitan forzar un valor positivo para ésta, debido a que se modelan estructuras nominales, usando para la dinámica de los factores un proceso CIR [1]. Para la creación de una estructura real de tasas de interés, es decir, ajustada por inflación, esta restricción es relajada dado que éstas sí presentan valores negativos, por ejemplo en períodos de alta inflación.

El número N de factores que afectan la dinámica de la tasa libre de riesgo es una elección que implica un compromiso entre dificultad de estimación y flexibilidad para ajustar de mejor forma estructuras temporales complejas. Cortazar, Schwartz, & Naranjo (2007) aplican este planteamiento al mercado chileno usando tres factores para modelar la tasa libre de riesgo en UF, logrando un buen ajuste a los datos y explicando un 99% de su varianza.

2.1.2.1 Metodología de estimación: El filtro de Kalman extendido.

El modelo dinámico requiere determinar por un lado los parámetros que lo rigen y por otro las variables de estado que representan la variable que se desea modelar. El filtro de Kalman (1960), permite realizar estas dos tareas de forma simultánea, usando tanto la

[1] Un proceso CIR (Cox, Ingersoll y Ross) es un proceso de difusión de raíz cuadrada donde la dinámica de los factores está dada por la ecuación:

$$dx = (\alpha - \beta x)dt + \sigma\sqrt{x}dw$$

El segundo término de la derecha fuerza a que el factor x sólo puede tomar valores mayores que cero para conservar las soluciones dentro del espacio de números reales.

información observada en la fecha que se quiere estimar como la de fechas pasadas. También permite un cierto nivel de error en la medición de la variable dependiente observada, aspecto importante dado que en un mercado poco desarrollado la observación podría estar determinada por aspectos específicos de la transacción que no necesariamente representan información de mercado.

Para estimar los parámetros, el filtro calcula la distribución de las innovaciones de observaciones en el proceso, donde una innovacion es el cambio de la curva generado por el filtro entre una fecha y la fehca siguiente. Esta distribución permite obtener una función de verosimilitud que al ser maximizada da lugar a estimaciones consistentes de los parámetros que rigen la dinámica de las variables de estado. La metodología propuesta por Kalman se define para procesos gaussianos, por lo tanto es necesario suponer que los errores de observación del modelo siguen una distribución normal, aunque este supuesto puede ser relajado.

El filtro de Kalman se define en el espacio de estados mediante ecuaciones de medida, transición y predicción. La ecuación de medida es la que relaciona las variables observadas y no observadas.

$$z_t = H_t x_t + d_t + v_t \qquad v_t : N(0, R_t)$$

(2-13)

El vector z_t de longitud $(m_t \; x \; 1)$ reúne las m_t variables observadas en el período t. H_t es una matriz de $(m_t \; x \; n)$, donde n es el número de factores del modelo o de variables de estado x_t que explican las observaciones, d_t es un vector de constantes que representa la distancia entre las variables no observables a las observadas y v_t es una variable

estocástica no correlacionada de media cero y varianza R_t que representa los errores de medición, los cuales siguen una distribución normal.

Por otro lado, la dinámica de las variables de estado x_t se representa en el filtro mediante la ecuación de transición.

$$x_t = A_t x_{t-1} + c_t + \varepsilon_t \qquad \varepsilon_t : N(0, Q_t)$$

(2-14)

La ecuación anterior describe tanto la relación lineal que existe de un estado a otro como la distribución normal multivariada de las perturbaciones ε_t, la cual permite modelar la incertidumbre en la estimación de las variables de estado del período siguiente. Definidas las ecuaciones (2-13) y (2-14), Kalman determina la forma de estimación del espacio de estados mediante la minimización del error cuadrático en la estimación, es decir si se define $\hat{x}_t$ como la estimación de la variable de estado x_t dada la información disponible hasta t, el error cuadrático de estimación queda definido por la siguiente ecuación:

$$P_t = E[(x_t - \hat{x}_t)(x_t - \hat{x}_t)']$$

(2-15)

Dado que $\hat{x}_t$ y P_t contienen toda la información disponible hasta el período t, el filtro de Kalman es un método recursivo, lo que implica que de haber nueva información observada en $t + \Delta t$, se puede estimar los nuevos valores $\hat{x}_{t+\Delta t}$ y $P_{t+\Delta t}$. A continuación se detalla el procedimiento para realizar esta estimación.

Dada toda la información disponible hasta el período t, es decir conocidos $\hat{x}_t$ y P_t, es posible estimar sus valores en $t + \Delta t$ gracias a las ecuaciones de predicción:

$$\hat{x}_{t+\Delta t|t} = A_{t+\Delta t}\hat{x}_t + c_{t+\Delta t}$$

(2-16)

$$P_{t+\Delta t|t} = A_{t+\Delta t}P_t A'_{t+\Delta t} + Q_{t+\Delta t}$$

(2-17)

La forma de la descripción de estas ecuaciones resalta la naturaleza bayesiana del filtro en el sentido de que la estimación de los valores en $t + \Delta t$ esta condicionada a los valores estimados en el tiempo t. Con estos estimadores, es posible realizar una predicción de las observaciones en el próximo período gracias a la ecuación de medida:

$$\hat{Z}_{t+\Delta t|t} = H_{t+\Delta t}\hat{x}_{t+\Delta t} + d_{t+\Delta t}$$

(2-18)

Luego al recibir la nueva información, ésta se puede contrastar con la predicción que hace el filtro, haciendo posible entonces encontrar la nueva predicción del filtro que es un promedio ponderado entre la información contemporánea y la histórica.

$$\hat{x}_{t+\Delta t} = \hat{x}_{t+\Delta t|t} + K_{t+\Delta t}(z_{t+\Delta t} - \hat{z}_{t+\Delta t})$$

(2-19)

$$\hat{P}_{t+\Delta t} = (I - K_{t+\Delta t}H_{t+\Delta t})P_{t+\Delta t|t}$$

(2-20)

En las ecuaciones anteriores (que son conocidas como paso de actualización), I corresponde a la matriz identidad de dimensión $(n\ x\ n)$, y K es la denominada ganancia de Kalman que se define por las siguientes ecuaciones.

$$K_{t+\Delta t} = P_{t+\Delta t}\mathrm{H}'_{t+\Delta t}\Omega^{-1}_{t+\Delta t}$$

(2-21)

$$\Omega_{t+\Delta t} = \mathrm{H}_{t+\Delta t} P_{t+\Delta t|t} \mathrm{H}'_{t+\Delta t} + \mathrm{R}_{t+\Delta t}$$

(2-22)

En éstas, $\Omega_{t+\Delta t}$ es la varianza de la predicción de las innovaciones, la cual, en conjunto con la esperanza de la predicción permite caracterizar la distribución de las innovaciones y por lo que son la base para realizar la estimación de los parámetros del modelo mediante el método de máxima verosimilitud enunciado anteriormente.

Este método busca encontrar el set de parámetros que maximiza la probabilidad de que los datos provengan de la distribución supuesta inicialmente. Bajo ciertas condiciones de regularidad, este método entrega estimaciones consistentes y asintóticamente normales. La distribución inicial supuesta, proviene de las distribuciones de los errores descritos en las ecuaciones (2-13) y (2-14). Por otro lado, las variables x_t y v_t distribuyen normal, lo que implica que z_t y por lo tanto $\hat{z}_{t+\Delta t}$ distribuyen normal (como la suma de variables normales y una constante). De esta forma, dado que la media y la varianza de las innovaciones esta definida en las ecuaciones (2-18) y (2-22), se puede describir su función de distribución de probabilidad según la siguiente ecuación:

$$f_{z_{t+\Delta t|t}}(z_{t+\Delta t}) = \frac{1}{2\pi|\Omega_{t+\Delta t}|^{1/2}} exp\left[-\frac{1}{2}\left(Z_{t+\Delta t} - \hat{Z}_{t+\Delta t|t}\right)' \Omega^{-1}\left(Z_{t+\Delta t} - \hat{Z}_{t+\Delta t|t}\right)\right]$$

(2-23)

Dado que la ecuación (2-23) describe la densidad de probabilidad de las innovaciones $z_{t+\Delta t}$ y la independencia de las variables $x_{t+\Delta t}$ y $v_{t+\Delta t}$, es posible obtener la densidad de probabilidad conjunta de estas innovaciones como la multiplicación de las funciones expresadas en (2-23), las cuales son representadas en su forma logarítmica en la siguiente ecuación:

$$\log\ L(\psi) = \sum_{t}\left(log\frac{1}{2\pi} - \frac{1}{2}\log|\Omega_t| - \frac{1}{2}\left(Z_{t+\Delta t} - \hat{Z}_{t+\Delta t|t}\right)'\Omega^{-1}\left(Z_{t+\Delta t} - \hat{Z}_{t+\Delta t|t}\right)\right)$$

(2-24)

La ecuación (2-24) resalta la dependencia de la función de verosimilitud del vector ψ, el cual reúne los parámetros que rigen la dinámica del modelo. De esta forma, al maximizar la ecuación (2-24) es posible obtener una estimación de los parámetros que sea consistente con las observaciones diarias. Como esta maximización no varía frente a constantes, la función que se maximiza está representada por:

$$\log\ \bar{L}(\psi) = -\frac{1}{2}\sum_{t}(\log|\Omega_t|) - \frac{1}{2}\sum_{t}\left(\left(Z_{t+\Delta t} - \hat{Z}_{t+\Delta t|t}\right)'\Omega^{-1}\left(Z_{t+\Delta t} - \hat{Z}_{t+\Delta t|t}\right)\right)$$

(2-25)

Hasta ahora se ha descrito la forma en la que se estiman de manera conjunta los parámetros y variables de estado del modelo, sin embargo aún quedan dos temas para poder aplicar el método de estimación. El primero son los valores iniciales que tomará el filtro y el segundo es cómo aplicar el filtro cuando la relación que hay entre las variables de estado y las observables no es lineal.

Para encontrar los valores iniciales del filtro es necesario definir los valores primitivos del vector de estados y la matriz de varianza-covarianza de los errores de estimación. Dado que estamos frente a un modelo estacionario (ya que se han supuesto que todos los componentes de la matriz K de reversión son distintos de cero) es posible utilizar los momentos incondicionales de las variables de estado, ya que no se tiene mayor información.

En Øksendal (2003) se demuestra que los momentos condicionales del vector de estados están dados por:

$$E_t\big(x_i(T)\big) = e^{-k_i(T-t)}x_i(t) \qquad i = 1, \dots, N \tag{2-26}$$

$$Cov_t\left(x_i(T), x_j(T)\right) = \sigma_i\sigma_j\rho_{ij}\frac{1 - e^{-(k_i+k_j)(T-t)}}{k_i + k_j} \qquad i, j = 1, \dots, N \tag{2-27}$$

donde las ecuaciones (2-26) y (2-27) representan los elementos de las matrices de esperanza y varianza-covarianza que describen los dos primeros momentos condicionales del vector de estados. Luego, los momentos incondicionales se desprenden del valor de largo plazo de las ecuaciones anteriores, el cual está dado por:

$$\lim_{T\to\infty} E_t\big(x_i(T)\big) = 0 \tag{2-28}$$

$$\lim_{T\to\infty} Cov_t\left(x_i(T), x_j(T)\right) = \frac{\sigma_i\sigma_j\rho_{ij}}{k_i + k_j} \tag{2-29}$$

Por lo tanto, el vector de variables de estado se inicializa en cero, mientras que la matriz de varianza-covarianza de los errores de estimación tiene como componentes iniciales cada uno de los términos descritos en la ecuación (2-29).

2.1.2.2 Extensión del Filtro de Kalman para una ecuación de medida no lineal.

El segundo tema que debe ser analizado para poder usar el método de estimación de Kalman es cómo aplicarlo cuando la relación entre las variables observables y no observables no es lineal.

La existencia de cupones en la mayoría de los papeles de largo plazo hace necesario la aplicación de la extensión del filtro de Kalman que define la ecuación de medida cuando no se cumple dicho supuesto, la cual fue propuesta por Harvey (1990) quién aplica un desarrollo de Taylor de primer orden alrededor de las mismas estimaciones que éste hace.

La ecuación de medida entonces supone que existe alguna función $P(x_t)$ que relaciona las variables observables y no observables según la siguiente ecuación:

$$z_t = P(x_t) + v_t \qquad v_t : N(0, R_t)$$

(2-30)

La ecuación anterior es linealizada utilizando una aproximación de primer orden mediante el desarrollo de Taylor donde el punto en torno al cual se linealiza está dado por la predicción del filtro $\hat{x}_{t|t-\Delta t}$. Esta aproximación se muestra en la siguiente ecuación:

$$P(x_t) = P\left(\hat{x}_{t|t-\Delta t}\right) + \left.\frac{\partial P(x_t)}{\partial x_t}\right|_{x_t = \hat{x}_{t|t-\Delta t}} (x_t - \hat{x}_{t|t-\Delta t})$$

(2-31)

De esta forma se define una nueva ecuación de medida linealizada dada por la ecuación (2-32).

$$z_t = \bar{H}_t x_t + \bar{d}_t + v_t \qquad v_t : N(0, R_t)$$

(2-32)

donde

$$\bar{H}_t = \left.\frac{\partial P(x_t)}{\partial x_t}\right|_{x_t = \hat{x}_{t|t-\Delta t}}$$

(2-33)

$$\bar{d}_t = P\left(\hat{x}_{t|t-\Delta t}\right) - \left.\frac{\partial P(x_t)}{\partial x_t}\right|_{x_t = \hat{x}_{t|t-\Delta t}} \left(\hat{x}_{t|t-\Delta t}\right)$$

(2-34)

El ANEXO B: muestra el desarrollo que permite comparar la matriz H para un bono de descuento y la matriz linealizada mediante el desarrollo de Taylor para un bono con cupones. De ahi se obtiene que los términos de la matriz $\bar{H}$ en (2-33) correspondientes a la observación i y el factor j y de la observación i del vector $\bar{d}$ están dados por las siguientes ecuaciones:

$$\bar{H}_t^{i,j} = \frac{u^{i,j} \sum_{k=1}^{Ncup} c_k^i e^{u'_k + v_k}}{\sum_{k=1}^{Ncup} c_k^i t_k e^{u'_k + v_k}}$$

(2-35)

$$\bar{d}_t^i = TIR^i - \bar{H}_t^i \hat{x}$$

(2-36)

donde TIR^i es la predicción de la tasa interna de retorno hecha por el filtro para el papel i, dado un vector de variables de estado $\hat{x}$ predichas y está dada por el valor que se obtiene al resolver numéricamente la ecuación (2-37).

$$TIR^i = y \qquad | \qquad \sum_{k=1}^{Ncup} C_k e^{yt_k} = \sum_{k=1}^{Ncup} C_k e^{u'_k + v_k}$$

(2-37)

Aplicando la linealización anterior, es decir reemplazando la matriz H original en las ecuaciones (2-20) a (2-22) por la nueva matriz $\overline{H}$ linealizada, se cumplen todas las condiciones necesarias para estimar un modelo dinámico usando papeles con cupones mediante el filtro de Kalman Extendido.

2.2 *Spread* de familias de papeles

Hasta este punto se ha discutido el problema de encontrar el aporte que hace a la TIR de una transacción la estructura libre de riesgo de los papeles de gobierno. Luego, faltan aquellos hechos por el *spread* de la familia de papeles y por el *spread* específico de cada transacción.

En esta parte de la tesis se discutirá el modelo planteado por Cortazar, Schwartz, & Tapia (2010), en el cual se modelan conjuntamente los papeles libres de riesgos con los riesgosos.

El modelo usa la estructura que propone Liu, Longstaff, & Mandell (2006) el cual usa las variables que modelan la estructura libre de riesgo en conjunto con las que modelan el *spread* de un papel riesgoso. Dado que el número de transacciones en el mercado Chileno son bajas, estos papeles riesgosos son agrupados en familias que tienen características similares que reflejan su riesgo crediticio.

Así, el *spread* de una transacción se modela con n^r factores que representan la estructura libre de riesgo, n^c factores comunes de los papeles riesgosos y n^g factores específicos. La representación de esta estructura esta dada por la ecuación (2-38).

$$s_j = \gamma_j 1^{r\prime} x^r + \delta_j^{c\prime} x^c + \delta_j^{g\prime} x^g + \delta_j^0$$

(2-38)

En esta ecuación s_j representa el spread de la familia j, x^r son las n^r variables de estado que modelan la tasa sin riesgo, γ_j es un factor que permite ajustar la correlación entre una clase j y la tasa libre de riesgo, δ_j^c y δ_j^g son los vectores de coeficientes para las variables de estado comunes y específicas, y finalmente, δ_j^0 representa la tendencia en el largo plazo que toma el *spread* de cada una de las familias.

Las variables de estado comunes y particulares de cada familia siguen un proceso Vasicek de la siguiente manera:

$$dx_t = -(\lambda + Kx_t)dt + \Sigma dw_t$$

(2-39)

donde K y Σ son matrices diagonales que representan la velocidad de reversión de las variables de estado x_t a la media y su varianza respectivamente. λ es un vector de premios por riesgo del mercado y dw_t es un vector de movimientos brownianos, donde la correlación entre ellos se da por la ecuación (2-40), donde Ω está formada por elementos ρ_{ij} que representan las correlaciones que existen entre los movimientos de la variable i y la variable j.

$$dw_i dw_j = \Omega dt$$

(2-40)

Luego, es necesario obtener una relación no para el *spread* de una familia sobre la curva libre de riesgo, sino que para la tasa de ésta. Así, usando la modelación de la tasa sin riesgo del capítulo anterior, se puede entonces obtener la relación que se busca.

$$\mathrm{R_j} = (1 + \gamma_\mathrm{j})1^{\mathrm{r}\prime}\mathrm{x^r} + \delta_\mathrm{j}^{\mathrm{c}\prime}\mathrm{x^c} + \delta_\mathrm{j}^{\mathrm{g}\prime}\mathrm{x^g} + \delta_\mathrm{j}^{0}$$

(2-41)

La ecuación (2-41) se obtiene directamente de la especificación del *spread* de una familia al sumar el término $1^{\mathrm{r}\prime}\mathrm{x^r}$ a ambos lados de la ecuación (2-38).

El valor de un papel con esta tasa sigue siendo el descrito por la ecuación (2-11), con un pequeño cambio en sus términos $u(T)$ y $v(T)$ debido al cambio en los coeficientes que acompañan a las variables de estado. Los nuevos términos se presentan en las ecuaciones (2-42).

$$u_i(\tau) = -\delta_i^x \left(\frac{1 - e^{-k_i\tau}}{k_i}\right)$$

$$v(\tau) = \sum_{i=1}^{n} \delta_i^x \frac{\lambda_i}{k_i}\left(\tau - \frac{1 - e^{-k_i\tau}}{k_i}\right) - \delta_0^r\tau$$

$$+ \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\delta_i^x)^2 \frac{\sigma_i\sigma_j\rho_{ij}}{k_i k_j}\left(\tau - \frac{1 - e^{-k_i\tau}}{k_i} - \frac{1 - e^{-k_j\tau}}{k_j} + \frac{1 - e^{-(k_i+k_j)\tau}}{k_i + k_j}\right)$$

(2-42)

donde $\delta_i^x = (1 + \gamma^j)$ para las variables que representan la estructura sin riesgo en la tasa de la familia j, $\delta_i^x = \delta_i^c(i)$ para las n^c variables comunes y $\delta_i^x = \delta_i^g$ para las n^g variables específicas de la clase j.

2.2.1 Metodología de estimación

La metodología propuesta en Cortazar, Schwartz, & Tapia (2010) usa la misma metodología usada para la estimación de la estructura libre de riesgo, sólo que mediante un proceso de dos etapas, siendo la primera la estructura libre de riesgo y luego, tomando esas variables como conocidas en el modelo presentado anteriormente. Así, el vector de variables de estado puede ser dividido de la siguiente manera:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x^r} \\ \hline \mathbf{x^c} \\ \mathbf{x^g} \end{bmatrix} = \begin{bmatrix} \mathbf{x^r} \\ \hline \mathbf{x^s} \end{bmatrix}$$

(2-43)

Esta división implica que no va a ser necesario que el filtro de Kalman realice una predicción de las variables $\mathrm{x^r}$ dado que estas ya van a existir a priori. De esta forma, la ecuación de medida queda representada por la ecuación

$$z_t^s = \bar{H}_t^s X^s + \bar{d}_t^s + v_t$$

(2-44)

y la matriz que relaciona las variables observables z_t^s con las no observables X^s queda definida de la siguiente manera:

$$\bar{H}_t^S = \begin{bmatrix} \frac{\sum_{k=1}^{Ncup_1} u_{1,k}^c C_k^1 e^{u'_{1,k}x+v_{1,k}}}{\sum_{k=1}^{Ncup_1} C_k^1 t_k e^{TIR_0^1 t_k}} & \frac{\sum_{k=1}^{Ncup_1} u_{1,k}^g C_k^1 e^{u'_{1,k}x+v_{1,k}}}{\sum_{k=1}^{Ncup_1} C_k^1 t_k e^{TIR_0^1 t_k}} & 0 \\ \vdots & \ddots & \vdots \\ \frac{\sum_{k=1}^{Ncup_m} u_{m,k}^c C_k^m e^{u'_{m,k}x+v_{m,k}}}{\sum_{k=1}^{Ncup_m} C_k^m t_k e^{TIR_0^m t_k}} & 0 & \frac{\sum_{k=1}^{Ncup_m} u_{m,k}^g C_k^m e^{u'_{m,k}x+v_{m,k}}}{\sum_{k=1}^{Ncup_m} C_k^m t_k e^{TIR_0^m t_k}} \end{bmatrix} \quad (2\text{-}45)$$

De esta forma, con las discusiones de los capítulos 2.1 y 2.2 respecto a la modelación de la estructura libre de riesgo y al *spread* de una familia de papeles respectivamente, se han cubierto dos de los tres aportes que el modelo de esta investigación propone para la especificación de la TIR de una transacción.

Luego, falta discutir la forma en que se abordará el problema de encontrar el *spread* específico de cada transacción el cual es determinado por las características propias de ésta. El modelo propuesto en esta investigación busca lograrlo mediante un modelo estático de minería de datos que no sólo pueda estimarlo sino que explicar la forma en que se encuentra cada uno de los *spreads* de cada transacción.

A continuación se presenta la tercera parte del modelo, precedida por una pequeña introducción a la minería de datos y su evolución.

3 MINERÍA DE DATOS

En este capítulo se explican algunas de las principales características de la minería de datos, de sus algoritmos más típicos y se detalla aquel que será usado en esta investigación.

A pesar de que esta disciplina, tal como se conoce hoy, no existe hace demasiado tiempo, sí ha existido en otras formas. Si se intentara llegar a una definición, podríamos decir que es el intento de hacer explícita información que está implícita en un conjunto de datos. Luego, no es raro encontrar ejemplos de lo anterior como lo es el análisis estadístico, el cual obtiene distintos tipos de información (media, desviación estándar, etc.) a partir de un set de datos. Otro ejemplo podría ser el gerente de una empresa, quién a partir de sus resultados y de los de la competencia intenta entender el por qué de éstos.

Para poder entender con mayor claridad lo que es la minería de datos es necesario mostrar la evolución de otras variables que fueron fundamentales. Los autores Han & Kamber (2006) en su libro *Data Mining: Concepts and Techniques* desarrollan una línea de tiempo simple de la evolución de estas variables, la cual se resume a continuación.

A partir de los años 60 la tecnología de la información y de almacenamiento de datos evoluciona desde simples sistemas de organización de archivos y carpetas a poderosos sistemas de bases de datos. Luego, en la década de los 70, éstos evolucionan a sistemas más complejos llamados bases de datos relacionales, los cuales se caracterizan por no sólo guardar diferentes tipos de información sino que también las relaciones que presentan entre sí. Se introducen también herramientas de modelación y técnicas de indexación y acceso a los datos.

Hacia mediados de los 80, la adopción de estos sistemas relacionales se vuelve muy común, lo que impulsa el desarrollo de nuevas actividades de investigación y desarrollo, que van desde la creación de nuevos modelos de bases de datos, distribución y diversificación hasta sistemas de bases de datos heterogéneos y globales como lo es la red mundial interconectada WWW.

Junto con lo anterior, también hay una evolución importante de la tecnología de *hardware* la cual da acceso a computadores con capacidad de procesamiento cada vez mayor, mejor recolección de datos y almacenamiento de información. Esto último da paso a la creación de repositorios de información de entre los cuales los *datawarehouse* son los más populares. Estos tienen la característica de poder juntar diferentes fuentes de información unificándolas en un solo sistema, teniendo también herramientas de acceso y análisis.

Gracias a todo esto, la cantidad de información almacenada crece cada vez más pero no necesariamente la información que se puede obtener, por lo que nuevas herramientas y técnicas para realizar análisis más profundos son requeridas, dando paso así a las nuevas técnicas de minería de datos.

La minería de datos puede dividirse en dos corrientes principales, dependiendo del uso que se le quiera dar. La primera representa aquellos algoritmos que buscan realizar tareas **descriptivas** sobre los datos, vale decir, aprender de ellos. Por otro lado, la segunda reúne a aquellos algoritmos que buscan realizar tareas de **predicción** de alguna característica de estos.

Dentro de los algoritmos descriptivos más comunes se encuentra el de *clustering*[2], el cual apunta a encontrar grupos de datos que presentan algún tipo de similitud bajo algún criterio de medida. Algunos de estos criterios de medida pueden ser diferencias, absolutas o no, de atributos numéricos o cantidad de características diferentes para atributos no numéricos. A su vez, estos algoritmos se pueden dividir en dos tipos: Los aglomerativos, que toman cada dato como un grupo y fusionan grupos similares y los divisionales, que agrupan a todos los datos en un solo conjunto el cual es posteriormente dividido en subgrupos. Estos se ejemplifican gráficamente en la Ilustración 1 y en la Ilustración 2.

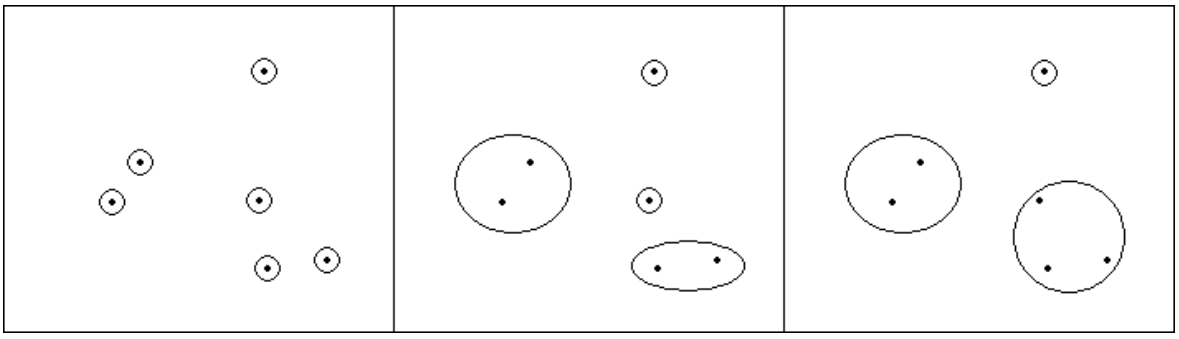

Ilustración 1: ejemplo *clustering* aglomerativo.

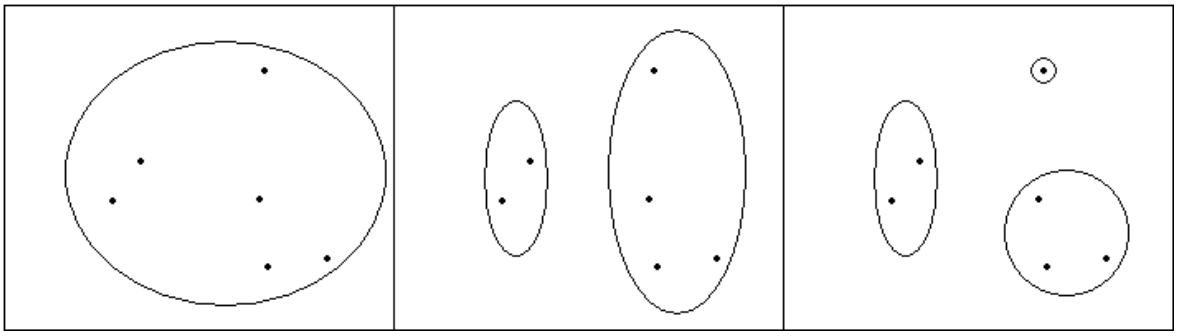

Ilustración 2: ejemplo clustering divisional.

[2] El *clustering* puede ser definido como el proceso de organizar objetos similares en grupos homogéneos, que a su vez sean heterogéneos con el resto de los grupos.

Otro algoritmo usado comúnmente es el de las reglas de asociación, el cual busca encontrar relaciones entre variables dentro de un set de datos, como por ejemplo compras de supermercado para ofrecer ventas cruzadas.

Por otro lado, algunos de los algoritmos predictivos más estudiados son los de redes neuronales o arboles de decisión.

Las redes neuronales son algoritmos que intentan emular el comportamiento del cerebro humano, relacionando un número de variables de entrada con otro número (diferente o igual) de variables mediante capas intermedias. Estas capas intermedias están formadas por elementos llamados *perceptrones*, los cuales a partir de un valor de entrada y un parámetro propio, obtenido mediante el entrenamiento de la red, dan respuestas binarias. Estas, compuestas a lo largo de todas las capas pueden predecir un resultado. La Ilustración 3 muestra un ejemplo simple de red neuronal, donde se puede observar que la capa de entrada está compuesta por tres valores, la capa de salida por dos y la capa oculta de los *perceptrones* formada por cuatro de estos.

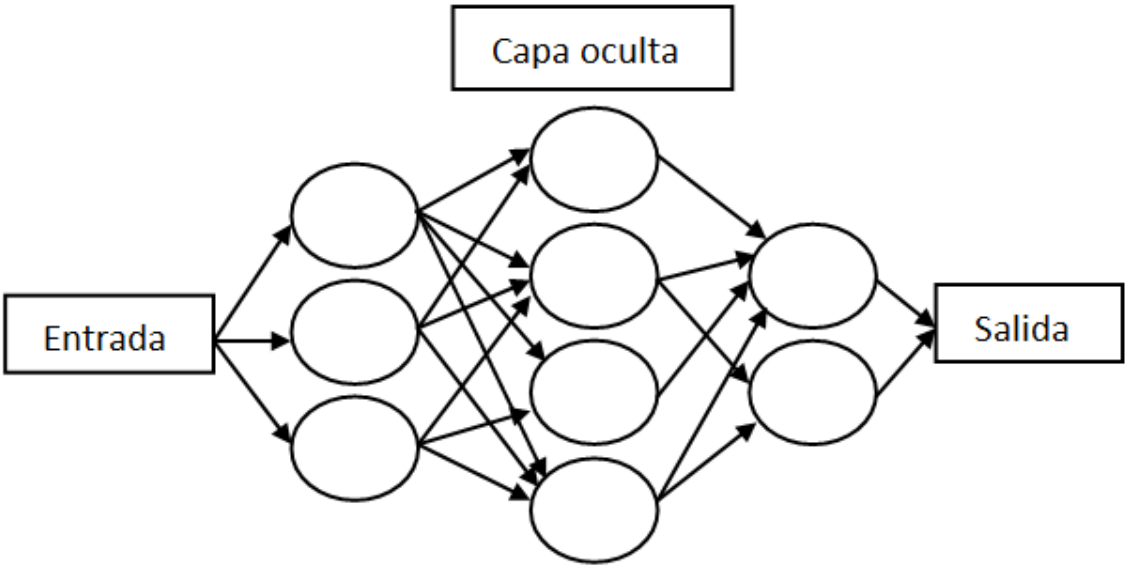

Ilustración 3: ejemplo simple de red neuronal

Ejemplos de aplicaciones de redes neuronales son reconocimientos de rostros o figuras, aprendizaje de tareas comunes como por ejemplo la conducción de un automóvil mediante un sistema robotizado.

Los árboles de decisión por su parte intentan clasificar elementos de un set de datos en diferentes categorías mediante el estudio de la relación entre éstas y las demás variables. Para esto, usan diferentes criterios como la entropía para evaluar cuáles aportan mayor información al proceso de clasificación y de esa forma encontrar una estrategia secuencial compuesta de preguntas que separan a los datos dependiendo de las respuestas a esas preguntas.

Este último algoritmo es particularmente útil cuando se requieren soluciones que pueden ser entendidas fácilmente (a diferencia de las redes neuronales) por personas para obtener información no sólo de la predicción de la clasificación sino que del proceso que llevó a ésta.

Hoy en día se pueden encontrar diversos ejemplos en la literatura como Adams, Hand, & Till (2001) quienes con técnicas de *clustering* y reconocimiento de patrones definen grupos de comportamiento homogéneo de clientes de un banco para encontrar aquellos que muestren desviaciones de éstos, y así prevenir incumplimientos de créditos. Por otro lado, O'Connor & Madden (2006) intentan predecir y aprovechar los movimientos del índice DJIA[3] utilizando redes neuronales alimentadas con indicadores externos como precios de *commodities* o tasas de cambio de monedas, logrando un retorno sobre la inversión sustancialmente mayores a los del índice. También Deconinck et al. (2005) intentan predecir la propiedad de absorción de distintas moléculas, relevante en el proceso de descubrimiento de nuevas drogas, usando la metodología CART[4].

El algoritmo utilizado en esta investigación es el de los árboles de regresión. Este es básicamente el mismo que el de árboles de clasificación, salvo que la variable que se intenta explicar es de carácter continuo.

A continuación se profundiza en el rol que este algoritmo cumple dentro de esta investigación y la forma en que funciona.

3.1 Arboles de Regresión.

La mayoría de la literatura en relación a algorítmos de árboles de regresión se basa en el trabajo de Breiman et al. (1984) publicado en su libro *Classification and Regression*

[3] *Dow Jones Industrial Average*
[4] CART se refiere a *Classification And Regression Trees*, una técnica introducida por (Breiman, Friedman, Olshen, & Stone, 1984).

Trees (CART). En éste, se muestra principalmente cómo debe ser abordado el problema de la construcción de dichos árboles.

El objetivo que se intenta completar es encontrar una relación entre un conjunto de variables explicativas y una variable dependiente de estas. Dado que no se conoce a priori la forma que esta relación debe tener, se desean encontrar cosas como las variables más influyentes, el órden en que deben considerarse, que tanto hay que profundizar, etc.

Luego, un árbol de regresión es una estructura que describe una variable de respuesta en función de un conjunto de variables explicativas. La metodología CART, que permite encontrar dicha estructura consiste de tres etapas. Estas etapas consisten básicamente en, primero, armar un árbol de gran tamaño que describa lo mejor que sea posible los datos con los que se está construyendo para luego, en una la segunda etapa, podarlo para encontrar diversos árboles que sean subconjuntos del anterior pero cada uno óptimo según un criterio de poder de predicción – complejidad (cada árbol pondera diferente ambos criterios). Finalmente, usando una técnica de validación cruzada, la tercera etapa encuentra el óptimo según el criterio del error cuadrático medio que se comete en la predicción.

A continuación se detalla el proceso de cada etapa.

La **primera etapa** se encarga de construir un árbol llamado *maximal tree* (de ahora en adelante MT) el cual tiene la característica de estar extremadamente sobre especificado. El nodo raiz o nodo padre (consistente en todos los objetos del conjunto de datos) es dividido en dos grupos mutuamente excluyentes mediante una regla simple sobre sólo una variable, obteniendose así dos nodos hijos. El proceso luego se repite tomando cada nodo

hijo como un nodo padre hasta que ya no se puedan realizar más divisiones, ya sea porque un nodo contiene sólo un dato o los datos dentro de éste son homogéneos.

La división en cada iteración del proceso es escogida óptimamente a partir de todas las divisiones posibles que puedan ser hechas sobre las variables que existan. Para ello se usa el criterio de impureza i que se define sobre un conjunto de datos t de la siguiente manera:

$$i(t) = \sum_{i=1}^{n} (y_i - \bar{y}(t))^2$$

(3-1)

donde $i(t)$ es la impureza del conjunto de datos t de tamaño n, y_i es el valor de respuesta del elemento i e $\bar{y}(t)$ es el promedio de los valores de respuesta de todos los elementos del conjunto t definido como:

$$\bar{y}(t) = \frac{1}{n} \sum_{i=1}^{n} y_i$$

(3-2)

Luego, la mejor división en cada iteración será aquella que minimice la impureza de los nodos hijos, lo que es análogo a maximizar la bondad de ésta, la cual está definida de la siguiente manera:

$$\Delta i(d, t) = i(t_p) - p_I i(t_I) - p_D i(t_D)$$

(3-3)

donde Δi es la bondad de la división d, t es el conjunto de datos que se está dividiendo, p_I es la proporción de elementos de t que luego de la división van al subconjunto t_I y p_D es la proporción de elementos de t que luego de la división van al subconjunto t_D.

El resultado de esta etapa es el MT. En general éste tiene las características de ser extremadamente grande, especificar completamente el conjunto original de datos y tener un bajo poder explicativo sobre datos nuevos.

Debido a lo anterior, la **segunda etapa** se encarga de podar el MT de forma de encontrar la relación óptima entre la complejidad del árbol y su poder predictivo.

Para realizar la poda se quitan ramas sucesivamente del árbol original, generando así una serie de árboles cada vez más pequeños que son subconjuntos del MT. Luego, para comparar los árboles en complejidad y poder predictivo, se define:

$$R_\alpha(T) = R(T) + \alpha\left|\tilde{T}\right| \tag{3-4}$$

donde $R_\alpha(T)$ es el criterio complejidad – poder predictivo para un árbol T, $R(T)$ es la suma de errores cuadráticos dentro de todos los nodos, $\left|\tilde{T}\right|$ la complejidad del árbol definida como el número de nodos terminales del árbol y α es el parámetro de complejidad que representa un castigo por cada nodo terminal extra. Este parámetro es incrementado gradualmente desde cero hasta uno, encontrando cada vez el árbol T que minimice el criterio $R_\alpha(T)$.

El resultado de esta etapa es un conjunto de árboles subconjuntos de MT que son óptimos en complejidad – poder predictivo.

La **tercera etapa** encuentra el óptimo valor de α mediante la técnica de validación cruzada. Este tipo de validación requiere de tomar m muestras M_i aleatorias del set de datos original S excluyentes entre sí, de forma que:

$$\bigcup_{i=1}^{m} M_i = S \tag{3-5}$$

$$\bigcap_{i=1}^{m} M_i = \emptyset \tag{3-6}$$

Posteriormente, se toma un conjunto de datos S, se le quita M_i, y se realizan las dos primeras etapas, encontrando así un árbol $T(\alpha)$ óptimo en complejidad – poder predictivo para cada valor de α. Estos árboles son evaluados finalmente según su poder predictivo contra la muestra M_i (que fue removida del conjunto de datos usados para obtener el árbol) según la medida de la raíz del error cuadrático medio definido de la siguiente forma:

$$RSME = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{3-7}$$

donde y_i es el valor de respuesta para un elemento i del set de datos S, $\hat{y}_i$ es el valor real de respuesta del elemento i y n es el número de elementos dentro del set de datos S.

Finalmente, la validación se repite eliminando del conjunto S siempre un subconjunto M_j diferente para el entrenamiento de los árboles óptimos en complejidad – poder predictivo.

De esta forma se obtienen m evaluaciones para los diferentes valores de α usados, pudiendo así encontrar el óptimo.

4 DESCRIPCIÓN DEL MERCADO DE RENTA FIJA CHILENO

Este capítulo pretende mostrar el contexto en el que se encuentran insertos los instrumentos hipotecarios en el mercado Chileno, explicando la finalidad que estos tienen, su presencia y otras de las características que influyen directamente en el modelo que se presenta, los resultados y conclusiones.

4.1 Instrumentos de renta fija Chilenos

El Mercado de Renta Fija Chileno se compone de todos aquellos instrumentos de deuda emitidos por el Banco Central de Chile, la Tesorería General de la República o las Instituciones Financieras y Empresas. Estos instrumentos son emitidos en 3 monedas principalmente: Pesos (CLP), Dólares (USD) y Unidades de Fomento (UF)[5], siendo esta última una moneda reajustable, indexada a la inflación vigente en el país. Esta se calcula de acuerdo a la tasa promedio geométrica de la variación del IPC del mes anterior[6], según la siguiente ecuación:

$$UF_t = UF_{t-1}(1 + IPC_{mes-1})^{1/días_{mes}} \tag{4-1}$$

Este mercado de renta fija puede dividirse en dos grupos: El primero, los Instrumentos de Intermediación Financiera (IIF), consta generalmente de papeles de corta

[5] Existe una cuarta moneda, el Indice Valor Promedio (IVP) que se calcula en base a variaciones de la UF, pero representa menos de un 1.5% de los saldos vigentes de renta fija al 20/07/2010.

[6] La inflación de un determinado mes es estimada mediante la variación del Índice de Precios al Consumidor (IPC) que publica mensualmente el Instituto Nacional de Estadísticas. Este se calcula en base al valor de mercado de una canasta de bienes representativa del consumo de una familia.

duración (menor a un año) que se emiten para satisfacer necesidades de financiamiento o como herramienta de regulación monetaria por el Banco Central. El segundo, los Instrumentos de Renta Fija (IRF), incluye las obligaciones de mayores plazos y constituye una de las principales fuentes de financiamiento de las empresas e instituciones financieras.

Tabla 1: número de papeles y saldos vigentes por mercado y moneda al 20/07/2010

Mercado	Moneda	Numero de Papeles	Saldo total (billones de CLP)	Porcentaje del saldo
IIF	**PESOS**	2006	17,68	58,29%
IIF	**UF**	1143	9,21	30,38%
IIF	**Dolar**	415	3,44	11,33%
Total IIF		**3564**	**30,32**	**38,83%**
IRF	**PESOS**	26309	5,02	10,50%
IRF	**UF**	12823	42,47	88,91%
IRF	**Dolar**	5	0,28	0,59%
Total IRF		**39137**	**47,77**	**61,17%**
Total General		**42701**	**78,09**	**100%**

La Tabla 1 muestra una descripción del número de papeles y saldos agrupados por mercado y moneda. Se puede observar que si bien el mercado IRF tiene un mayor número de emisiones vigentes, los saldos presentan una diferencia considerablemente menor. Por otro lado, la Intermediación Financiera está compuesta principalmente por depósitos a plazo fijo emitidos por instituciones financieras, observándose un mayor número de emisiones en UF, pero saldos vigentes mayores en pesos. Los saldos de papeles de largo plazo, por su parte se concentran casi en su totalidad en UF.

El mercado IRF puede clasificarse tanto según sus emisores como según su finalidad. A continuación se describen los principales tipos de papeles de este mercado:

Renta Fija Gubernamental: Papeles emitidos por el Banco Central o por la Tesorería General de la República ya sea en pesos o en UF.

- **Bonos del Banco Central**
    - **BCP**: Papeles del tipo *bullet* emitidos por el Banco Central en pesos que pagan intereses semestrales y la totalidad del principal al vencimiento. Se emiten en plazos de 2, 5 y 10 años con tasas de emisión de 6% y 8%.
    - **BCU**: Papeles del tipo *bullet* emitidos por el Banco Central en UF que pagan intereses semestrales y la totalidad del principal al vencimiento. Se emiten a plazos de 5, 10 y 20 años con tasas de emisión de 3% y 5%.
    - **PRC**: Papeles amortizables y reajustables emitidos por el Banco Central, pagan cupones semestrales e iguales. Las tasas de emisión varían entre 5% y 6,5% y sus plazos entre 4 y 30 años. Estos papeles dejaron de emitirse en agosto de 2002, pero continúan transándose activamente en el mercado.
    - **CERO**: Bono cero cupón emitido por el Banco Central como opción para sustitución o canje de cupones de pagarés reajustables (PRC).
    - **Bonos en USD**: Papeles cero cupón, *bullet* o amortizables emitidos en dólares de Estados Unidos.
- **Bonos de Reconocimiento**
    - **BR**: Bono emitido por el Instituto de Normalización Previsional para reconocer cotizaciones de trabajadores previas al cambio de sistema previsional en 1980.
- **Bonos de Tesorería**

- **BTP**: Bono igual a los BCP en estructura de pagos pero emitido por la Tesorería General de la República en pesos.

- **BTU**: Bono iguale a los BCU en estructura de pagos pero emitido por la Tesorería General de la República en UF. Se emiten en plazos entre 10 y 30 años, con tasa de emisión entre 2,1% y 4,5%. Comienzan a emitirse a partir de octubre de 2003.

Renta Fija Privada: Papeles emitidos por Instituciones Financieras y Empresas que operen en el país.

- **Bonos Bancarios (BB)**: Instrumentos de deuda de largo plazo emitidos por bancos o instituciones financieras como forma de financiamiento.

- **Bonos Corporativos (BC)**: Instrumentos de deuda de largo plazo emitidos por empresas como forma de financiamiento.

- **Bonos Securitizados (BS)**: Instrumentos de deuda de largo plazo emitidos por empresas securitizadoras formando patrimonios separados con objetivos específicos fijados por sus clientes.

- **Bonos Subordinados (BU)**: Bonos emitidos por bancos para cumplir con los requerimientos de capital impuestos por la superintendencia que los regula.

- **Letras Hipotecarias (LH)**: Instrumentos de deuda emitidos por bancos e instituciones financieras para el otorgamiento de créditos hipotecarios.

La Tabla 2 muestra la distribución de papeles y saldos de renta fija de largo plazo según los distintos tipos antes mencionados [7] y moneda de emisión.

Tabla 2: distribución por tipo de papel del mercado de renta fija de largo plazo el 20/07/2010.

Moneda	Tipo	Numero papeles	Saldo (Billones CLP)	Porcentaje Saldo
PESOS	BONO BANCO	6	0,36	0,74%
	BONO BANCO CENTRAL	15	2,16	4,51%
	BONO TESORERIA	4	1,16	2,41%
	BONO CORPORATIVO	23	0,66	1,39%
	BONO SECURITIZADO	38	0,63	1,33%
	LETRA HIPOTECARIA	7	0,00	0,00%
UF	BONO BANCO	105	7,59	15,86%
	BONO BANCO CENTRAL	654	7,00	14,63%
	BONO TESORERIA	15	4,66	9,74%
	BONO CORPORATIVO	277	13,83	28,90%
	BONO SECURITIZADO	165	0,68	1,43%
	LETRA HIPOTECARIA	11398	5,39	11,28%
	BONO SUBORDINADO	108	3,53	7,37%
	Otros	101	0,05	0,11%
DOLAR	BONO BANCO CENTRAL	0	0,00	0,00%
	BONO CORPORATIVO	1	0,10	0,22%
	BONO SECURITIZADO	2	0,03	0,07%
TOTAL				100%

[7] Se muestran los saldos los nominales vigentes de los instrumentos presentes en la custodia del Depósito Central de Valores de Chile (DCV).

De ella se puede extraer que la suma de los saldos (en sus diferentes monedas) de los bonos corporativos es de 14,59 billones de pesos en saldos vigentes, constituyendo así el principal activo de largo plazo de la economía chilena. De hecho si se consideran además los bonos bancarios, subordinados y securitizados, en un grupo llamado bonos empresariales (BE) este monto aumenta a 19,58 billones de pesos, representando más del 40% del total de saldos de renta fija de largo plazo. El segundo grupo en importancia es del Banco Central con 9,16 billones de pesos. Estos papeles son los que permiten construir la estructura libre de riesgo en el mercado nacional. Estos papeles junto a los de la Tesorería General de la Republica forman el grupo de los bonos de gobierno (BG) y juntos suman 16,75 billones de pesos, representando aproximadamente un 35% del total de saldos de renta fija.

4.2 Las letras de crédito hipotecarias

Las letras de crédito hipotecario (LCH) tienen una participación importante en el mercado financiero Chileno. Éstas son emitidas por bancos y sociedades financieras cuando un tercero tiene la necesidad de obtener financiamiento hipotecario. Cada letra tiene asociado un contrato mutuo hipotecario que se pacta directamente con el deudor el cual representa la obligación que éste contrae con el emisor de la LCH.

A continuación se presentan algunas de las principales características de estos instrumentos.

- **Distintos tipos de letras de crédito hipotecario**

Éstas pueden ser emitidas tanto por motivos de construcción de viviendas como para financiar actividades productivas de distintos tipos.

La amortización de las letras de crédito puede realizarse en forma ordinaria, ya sea directa o indirectamente, o en forma extraordinaria.

- **Amortización Ordinaria Directa**

Es aquélla en que periódicamente el emisor de la letra paga la parte del capital y de los intereses convenidos según la tabla de desarrollo.

Es el tipo más común con cerca del 100% del mercado.

- **Amortización Ordinaria Indirecta**

Se efectúa mediante compra o rescate de letras o por sorteo a la par, hasta por un valor nominal igual al fondo de amortización[8] correspondiente al periodo respectivo.

- **Amortización Extraordinaria (prepagos)**

Consiste en el retiro de una LCH del mercado ya sea por compra, rescate o sorteo a la par de un monto igual al prepagado por el deudor. La amortización extraordinaria se produce también cuando el deudor paga anticipadamente todo o parte de su deuda mediante la entrega de letras de crédito.

- **Emisión de letras**

Cualquier institución calificada para emitir letras de crédito hipotecario debe hacerlo presentando en la Superintendencia de Bancos e Instituciones Financieras un prospecto de emisión que contenga el monto total de la emisión, moneda o unidad de valor en que se

[8] El fondo que se menciona es una cuenta especial que forma la institución emisora con los dineros correspondientes a los pagos efectuados por el mutuario.

expresa, plazo de los préstamos (mayor a 1 año), tasa de interés (fija o flotante), amortización, lugar de impresión (Casa de Moneda o Imprenta), cortes y el valor del cupón

Una vez efectuada la emisión de la LCH, el emisor debe proceder a liquidarlas para lo cual tiene las siguientes alternativas:

- Colocar las letras en la Bolsa de Comercio para permitir que sean transadas libremente.
- Comprar sus propias emisiones, por ejemplo, si existiese una contracción del mercado normal o cuando las necesidades financieras así lo exigen.
- Vender las letras de crédito directamente una institución financiera diferente.
- Entregar las letras de crédito a su dueño (vendedor de la casa) si este no requiere de liquidez inmediata.

- **Tablas de Desarrollo**

Cada crédito hipotecario posee 2 tablas de desarrollo que se encuentran relacionadas. La primera es la tabla de las letras hipotecarias y la segunda es la tabla del mutuo hipotecario asociado.

El dividendo de los deudores está compuesto de: amortización, interés, comisión y seguros, y los cupones de las letras emitidas están compuestos por amortización e interés.

El ANEXO E: detalla la creación de las tablas de desarrollo asociadas a estos instrumentos.

- **Prepago**

La principal característica de una letra de crédito hipotecario es la posibilidad de efectuar un prepago (o amortización extraordinaria) de parte o todo el capital no amortizado.

Cuando se trata de amortizaciones parciales, el pago se aplica proporcionalmente a los dividendos restantes de la deuda (en vez de concentrarlos al comienzo o al final), de modo que no se altere el plazo pactado en ella.

El deudor que amortiza en forma extraordinaria, total o parcialmente el saldo de su deuda, debe pagar a la entidad emisora, adicionalmente, una suma equivalente al interés y comisión correspondiente a un período de amortización de las letras de su préstamo, factor que dificulta la realización de estas amortizaciones.

El rescate de las LCH se efectúa mediante un sorteo, sobre las LCH pertenecientes a la misma serie, que se realiza dentro de los primeros diez días de cada mes.

4.3 Principales inversionistas del mercado Chileno

A continuación se describen los principales inversionistas del mercado financiero chileno. Estos pueden separarse en inversionistas institucionales, privados o extranjeros. Los primeros corresponden a bancos, sociedades financieras, compañías de seguros, entidades nacionales re aseguradoras y administradoras de fondos autorizados por la ley. Por otro lado los inversionistas privados son todos aquellos intermediarios de valores y personas naturales o jurídicas que declaren y acrediten contar con inversiones financieras no inferiores a 2.000 UF. Los principales inversionistas institucionales son:

- **Administradoras de Fondos de Pensiones**: son los inversionistas institucionales más importantes en términos de volúmenes de inversión.

- **Compañías de Seguros:** Se dividen en compañías de seguros de vida (CSV) o generales (CSG), las primeras invierten prioritariamente en instrumentos de deuda de largo plazo, mientras que las generales son inversionistas de más corto plazo.

- **Fondos Mutuos (FM):** son patrimonios integrados por aportes comunes de personas naturales y jurídicas para su inversión en valores de oferta pública. El patrimonio de cada fondo se divide en cuotas rescatables iguales.

- **Fondos de Inversión (FI)**: son patrimonios integrados por aportes comunes de personas naturales y jurídicas para su inversión en valores y bienes que autorice la Ley de Fondos de Inversión. Los aportes quedan expresados en cuotas que no pueden ser rescatadas antes de la liquidación del fondo.

- **Fondos de Inversión de Capital Extranjero (FICE)**: su patrimonio está formado por aportes comunes realizados fuera de Chile por personas naturales o jurídicas.

- **Fondos para la Vivienda**: patrimonio constituido con los recursos depositados en las cuentas de ahorro para arrendamiento de viviendas con promesa de compraventa.

5 MODELO

A continuación se describe el modelo propuesto en esta tesis para la estimación de precios de instrumentos hipotecarios mediante modelos dinámicos y herramientas de minería de datos.

El modelo se puede resumir según la ecuación (5-1), donde TIR_{LH} representa la TIR de la transacción de una Letra Hipotecaria. Por otro lado, TIR_{Gob} es la TIR que aporta la estructura libre de riesgo que se encuentra a partir de los bonos libres de riesgo que emite el gobierno. Luego, $S_{Familia\ LH}$ representa el aporte de la familia de papeles de las letras, $S_{específico}$ el aporte específico que depende de las características de cada transacción y ε la parte de la tasa que no queda explicada por el modelo.

$$TIR_{LH} = TIR_{Gob} + S_{Familia\ LH} + S_{específico} + \varepsilon \tag{5-1}$$

Para efectos de calibración del modelo, se tomará en cuenta una variación de la ecuación (5-1), en la que el aporte de la TIR de gobierno y el *Spread* aportado por la familia (letras de crédito hipotecario) se toman como uno según la siguiente ecuación:

$$TIR_{Curva\ Promedio} = TIR_{Gob} + S_{Familia} \tag{5-2}$$

La especificación de esta curva promedio, o primera etapa del modelo, será discutida en el siguiente capítulo, mientras que el $S_{específico}$ se obtiene en la segunda etapa del modelo (discutida también en el capítulo siguiente) mediante el algoritmo de minería de datos Árboles de Regresión.

Para efectos de análisis posteriores, la $TIR_{Curva\ Promedio}$ será descompuesta en el aporte de la TIR de gobierno y el *Spread* de la familia de papeles. El primer aporte TIR_{Gob} es obtenido a partir del trabajo hecho por Cortazar, Schwartz, & Naranjo (2007) discutido en el punto 2.1.2. Por otro lado, el $S_{Familia\ LH}$ se obtiene a partir de la diferencia entre la curva promedio ($TIR_{Curva\ Promedio}$), y el aporte hecho por la TIR_{Gob}. Finalmente, el modelo concluye con una corrección diaria del resultado obtenido hasta este punto, la cual será discutida más adelante en este capítulo.

5.1 Primera etapa: La curva promedio

El modelo propuesto para esta etapa guarda una alta similitud con el modelo propuesto en Cortazar, Schwartz, & Naranjo (2007) para la modelación de la tasa de interés libre de riesgo, el cual es explicado en mayor detalle en el punto 2.1.2 de esta tesis.

El modelo pretende encontrar en una primera instancia el comportamiento promedio de las LCH de la misma forma en que se modelan los precios de instrumentos libres de riesgo. Este mismo objetivo puede encontrarse en el trabajo hecho por Cortazar, Schwartz, & Tapia (2010) cuyo modelo se discute en el punto 2.2 de esta tesis, en el cual en vez de modelar los instrumentos como si fueran libres de riesgo, se modela el *spread* que estos tienen sobre una curva libre de riesgo.

La curva promedio de las LCH se modela usando una especificación Vasicek N factorial de la siguiente manera:

$$r = 1'x + \delta_0$$

(5-3)

estando la dinámica de cada uno de los factores x determinada por la siguiente expresión:

$$dx_t = -Kx_t dt + \Sigma dw_t$$

(5-4)

En ésta, al igual que en el modelo libre de riesgo, K y Σ son matrices diagonales donde cada elemento k_{ii} y σ_{ii} representa la velocidad de reversión a la media y varianza del factor x_i respectivamente. Luego de ajustarlo por riesgo la dinámica queda especificada por la siguiente ecuación:

$$dx_t = -(\lambda + Kx_t)dt + \Sigma dw_t$$

(5-5)

donde λ_i representa el precio del riesgo para cada factor i y el término dw_t representa un vector de movimientos Brownianos correlacionados tal que:

$$(dw)'(dw) = \Omega dt$$

(5-6)

Finalmente, al igual que en el modelo libre de riesgo, aplicando el lema de Îto sobre la estructura planteada anteriormente obtenemos el precio de la LCH según:

$$P(x_t, \tau) = \sum_{i}^{Cupones} e^{(u(\tau_i)'x_t + v(\tau_i))}$$

(5-7)

$$u_i(\tau) = -\left(\frac{1 - e^{-k_i\tau}}{k_i}\right)$$

(5-8)

$$v(\tau) = \sum_{i=1}^{n} \frac{\lambda_i}{k_i}\left(\tau - \frac{1 - e^{-k_i\tau}}{k_i}\right) - \delta\tau$$
$$+ \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\sigma_i\sigma_j\rho_{ij}}{k_i k_j}\left(\tau - \frac{1 - e^{-k_i\tau}}{k_i} - \frac{1 - e^{-k_j\tau}}{k_j} + \frac{1 - e^{-(k_i+k_j)\tau}}{k_i + k_j}\right)$$

(5-9)

La estimación de los parámetros K, Σ, λ y ρ se realiza mediante la metodología del filtro de Kalman descrita en el punto 2.1.2.2.

El resultado que se obtiene de esta etapa es una curva dinámica que N factores que describe el comportamiento general que presentan las LCH, el cual es actualizado cada día mediante el ingreso de nuevas transacciones. De la mano de este resultado vienen errores de estimación de las tasas de dichos instrumentos, los cuales serán tratados de explicar en la segunda etapa.

5.2 Segunda etapa: Estimación del Spread Específico de la letra

Esta etapa del modelo intenta encontrar el aporte hacen a la TIR de transacción las características propias de cada instrumento. Esta etapa deja un remanente sin poder ser explicado o error.

A su vez, usa como *input* el resultado de la etapa anterior, el cual consta principalmente de la componente promedio de la tasa interna de retorno de una LCH y, en la etapa de calibración del modelo, los *spreads* específicos que tiene cada transacción de LCH sobre la TIR de la curva promedio, los cuales dependen de las características propias del instrumento al momento de la ocurrencia de ésta.

Este *spread* específico se mide como la diferencia que existe entre la TIR de transacción u observada de la LCH y la estimada por la curva promedio de la primera etapa del modelo. Para encontrar dicho valor usamos las ecuaciones (5-10) y (5-11):

$$P = \sum_{i}^{Cupones} C_i e^{-TIR_{modelo} * T_i} \tag{5-10}$$

$$P = \sum_{i}^{Cupones} C_i e^{(u(T_i)' X + v(T_i))} \tag{5-11}$$

A partir de las ecuaciones anteriores se puede obtener un valor para TIR_{modelo} de tal forma que al descontar todos los cupones se obtenga el mismo precio que se encuentra usando la expresión (5-11).

Definiendo finalmente el *spread* específico o diferencia entre la TIR del modelo y la TIR observada como SE_i para cada LCH i:

$$SE_i = TIR_{Modelo} - TIR_{Observada} \tag{5-12}$$

La forma en que se crea el modelo de árbol de regresión para estimar los futuros *spreads* de las LCH a partir de los SE, es aquella descrita en el punto 3.1 de esta tesis, donde se muestra el algoritmo para crear dichos árboles. En ese punto se explica que el árbol de regresión debe ser alimentado con características propias de cada transacción de forma de poder así encontrar diversos patrones que cumplan el objetivo de la estimación. Las características propias de la letra usadas para este fin en esta investigación se presentan a continuación.

En un primer lugar, Longstaff, Mithal, & Neis (2005), usando información del mercado de los *credit default swap*, muestran que la mayoría del *spread* corporativo se explica por el riesgo de *default*, mientras que Bedendo, Cathcart, & El-Jahel (2007) muestran que existen factores idiosincráticos de la firma entre otros. Luego se puede justificar el uso de las siguientes características.

- **Clasificación:** la clasificación de riesgo que recibe la LCH, ligada directamente con la clasificación que su emisor recibe en el mercado financiero.

- **Emisor:** el banco o institución financiera que emite la LCH.

En segundo lugar, Chen, Lesmond, & Wei (2007) muestran el efecto significativo de la liquidez en los precios de bonos privados, luego, se puede justificar el uso de las siguientes características

- **Presencia:** cantidad de transacciones de una LCH dentro de un período.

- **Rotación:** Monto transado de una LCH como porcentaje del total existente para el mismo papel.

En tercer lugar, Wang (2008) muestra como los efectos de *default* y liquidez aumentan junto con el plazo al vencimiento de un papel, usando precios de bonos de la tesorería de Estados Unidos, mientras que Rodriguez (1988) muestra que en general los *spreads* de los bonos riesgosos son una función de la madurez de estos. Luego, se puede justificar el uso de la siguiente característica.

- **Plazo:** tiempo restante para llegar a la madurez de la LCH.

Finalmente, Kau (2005) muestra como el *spread* de las LCH disminuye a medida que disminuye el valor de la opción de prepago, por lo que se puede justificar el uso de las siguientes características.

- **Índice probabilidad de prepago:** un índice en base al comportamiento histórico del prepago que da una estimación de la probabilidad que tiene una LCH de ser prepagada, calculado en base a N (1, 5 y 10) años de información histórica del comportamiento de prepago de dicho instrumento de la siguiente manera:

$$Pr_{prepago}(N, i) = 1 - \prod\nolimits_{m \in N} (1 - p_m) \quad (5\text{-}13)$$

donde, $Pr_{prepago}(N, i)$ es la probabilidad de prepago de una letra i tomando información con N años de antigüedad, m es la cantidad de períodos de amortización extraordinaria (prepago) dentro de los N años, y p_m es el porcentaje de la letra que se prepaga en cada período m.

- **Tasa de emisión:** Aquella tasa con la cual se emite el instrumento LCH por el banco o institución financiera correspondiente, la cual representa los intereses que se pagaran en el tiempo.

El resultado de esta etapa es un árbol de regresión construido a partir de una variable de respuesta SE_i y un set de variables explicativas propias de cada SE_i, el cual intenta representar y estimar el *spread* específico de cada LCH.

Este resultado parcial, en conjunto con la curva obtenida en la primera etapa del modelo, ofrece una estimación de la tasa interna de retorno de cada LCH. Esta tasa tiene

una componente general y dinámica, obtenida a partir del comportamiento promedio de los instrumentos, y otra componente particular y estática obtenida a partir de las desviaciones y características particulares a cada LCH.

5.3 Tercera etapa: Corrección mediante el sesgo.

Esta etapa del modelo pretende corregir errores sistemáticos cometidos por el árbol de regresión. Estos errores sistemáticos pueden surgir debido a falta de generalidad o poder de adaptación del árbol de regresión al intentar predecir las desviaciones. Otra razón por la que estos errores pueden ocurrir es que las transacciones dentro de un día no sean representativas de todo el mercado, luego, el modelo dinámico que se estima según estas va a quedar sesgado al tener sólo estas transacciones como fuente de nueva información. Luego, como el árbol se ha entrenado usando todos los datos dentro de la muestra (por ende sí es representativo) el resultado de la corrección va a tener un error que va a ser sistemático a lo largo de toda la curva del día.

Luego, el procedimiento mediante el cual se pretende corregir estos problemas consta de tomar el error promedio que el modelo ha cometido hasta este punto (diariamente) para luego ajustar las estimaciones hechas por éste según ese sesgo calculado. Para ver esto matemáticamente, tenemos:

$$\varepsilon_{i,j} = TIR_{Curva\ Promedio}{}^{i,j} + S_{Específico}{}^{i,j} - TIR_{Observada}{}^{i,j} \quad (5\text{-}14)$$

donde $\varepsilon_{i,j}$ es el error cometido por el modelo para la transacción j ocurrida el día i, $TIR_{Curva\ Promedio}{}^{i,j}$ la TIR estimada por la primera etapa del modelo para la transacción j

del día i, ${S_{Específico}}^{i,j}$ el *spread* estimado por el árbol de regresión para la transacción j del día i y ${TIR_{Observada}}^{i,j}$ es la TIR efectiva de esa transacción.

Luego, podemos encontrar el error promedio dentro de un día i cualquiera de la siguiente manera:

$$\bar{\varepsilon}_i = \frac{1}{N_i}\sum_{k=1}^{N_i} \varepsilon_{i,j} \tag{5-15}$$

donde N_i es la cantidad de transacciones que hay en el día i.

Así finalmente podemos encontrar la TIR del modelo para una transacción j ocurrida el día i según la siguiente ecuación:

$${TIR_{Modelo}}^{i,j} = {TIR_{Curva\ Promedio}}^{i,j} + {S_{Específico}}^{i,j} - \bar{\varepsilon}_i \tag{5-16}$$

5.4 Metodologías alternativas

Dado que la metodología usada para estimar la curva promedio (primera etapa del modelo) ha sido discutida ampliamente en la literatura, la principal motivación de usar una metodología alternativa es lograr evaluar que tan bien el modelo de árbol de regresión (segunda etapa del modelo) logra encontrar los *spreads* específicos de las transacciones de LCH. Es por esta razón que las metodologías alternativas que se utilizarán serán básicamente lo mismo que el modelo original salvo en su segunda etapa.

5.4.1 Mantención del último *spread* específico

Para reemplazar el árbol de regresión en esta metodología alternativa se utilizará la información histórica del último error cometido por el modelo dinámico (primera etapa), para usarlo como si fuera la predicción de la segunda etapa del modelo (predicción del árbol). De esta forma, para realizar una estimación de TIR para algún día i, definimos:

$$\varepsilon_{t<i,j} = TIR_{CP}^{t<i,j} - TIR_{Observada}^{t<i,j} \tag{5-17}$$

donde $\varepsilon_{t<i,j}$ es la diferencia entre la TIR, para la transacción j, de la curva promedio en el día t (donde ocurre la última transacción del mismo instrumento) y la TIR observada ese mismo día, $TIR_{CP}^{t<i,j}$ es la TIR estimada por la curva promedio para el día t para la transacción j y $TIR_{Observada}^{t<i,j}$ es la TIR a la que se transó aquel instrumento el día t.

Luego, la TIR de este modelo para una transacción j ocurrida el día i se define de la siguiente manera:

$$TIR_{Modelo}{}^{i,j} = TIR_{Curva\ Promedio}{}^{i,j} - \varepsilon_{t<i,j} \tag{5-18}$$

donde $TIR_{Curva\ Promedio}{}^{i,j}$ es la tasa interna de retorno estimada por el modelo dinámico (primera etapa) para la transacción j el día i y $\varepsilon_{t<i,j}$ es la diferencia entre la TIR del modelo dinámico y la observada el día t (cuando se transó el mismo instrumento j).

Metodologías similares han sido ampliamente utilizadas como solución a este tipo de problemas, debido a la inexistencia de modelos que permitan mejores estimaciones o simplemente a modo de aproximaciones.

Por otro lado es razonable usar el último *spread* observado del papel ya que si asumimos que el comportamiento día a día de este *spread* sigue una caminata aleatoria (al estar expuesto a todas las variables del mercado), entonces el valor esperado de los cambios de éste va tender a cero.

5.4.2 Regresión

Para reemplazar al árbol de regresión en esta metodología alternativa se intentará estimar las desviaciones mediante un análisis de regresión.

La forma en que esto se llevará a cabo será regresionar los errores cometidos por la primera etapa (desviaciones) usando las mismas variables que se usan para entrenar el árbol de regresión. Luego, secuencialmente se eliminaran aquellas no significativas y se repetirá el procedimiento hasta que todas las variables sí lo sean.

Finalmente, a partir del resultado que se obtenga, se estimarán las desviaciones de cada transacción con respecto de la curva promedio (*spreads* específicos) y se evaluará el desempeño de la misma forma en que se evalúa para el modelo de árbol de regresión para facilitar su comparación.

6 IMPLEMENTACIÓN Y RESULTADOS

En este capítulo primero se realiza una discusión de los datos usados, mostrando sus principales características y su evolución dentro del período de estudio. Luego se muestran los resultados obtenidos de la implementación del modelo de tres etapas descrito en el capítulo anterior para su posterior análisis.

6.1 Datos

Se testea el modelo usando datos del mercado Chileno de Letras Hipotecarias. La decisión de utilizar estos instrumentos se debe principalmente a la existencia de opciones de prepago y características muy variadas propias de cada papel. La primera razón es importante debido a que la opción de prepago hace que cada instrumento independientemente del resto presente desviaciones del promedio de las transacciones. Estas desviaciones se deben a que esta opción presenta un riesgo para quién posee la letra y por ende la castigue al momento de comprarla según lo que crea de la potencialidad de la ejecución de dicha opción.

La segunda razón cobra importancia cuando se desea implementar el árbol de regresión para explicar las desviaciones del promedio. Para lograr este objetivo es necesario que las características de cada LCH (aquellas discutidas en la descripción de la segunda etapa del modelo como por ejemplo la clasificación de riesgo, emisor, probabilidad de prepago, etc.) ofrezcan algún grado de información de la forma en que éstas son transadas, de forma que el algoritmo logre extraerla y usarla tanto para corregir las estimaciones como para entender la forma en que son castigadas por el mercado.

Los datos usados corresponden a todas las transacciones de la bolsa de comercio de Santiago ocurridas entre las fechas 01-01-2006 y 01-07-2010. En la Tabla 3 se muestra un resumen de las transacciones del mercado Chileno de letras hipotecarias. En ella se puede apreciar que existe una disminución tanto en la cantidad de transacciones como en los montos que se han transado. Esto se explica principalmente por la aparición de nuevas formas de financiamiento hipotecario llamadas Mutuos Hipotecarios. Estos, a diferencia de las Letras de Crédito Hipotecario, son arreglos privados entre un deudor y una institución en los cuales el monto prestado proviene del capital propio de esta institución en vez del resultado de la venta del instrumento hipotecario en el mercado financiero.

Tabla 3: resumen de transacciones del mercado de letras hipotecarias entre el 01-01-2006 y el 30-06-2010

Año	Semestre	Transacciones	Montos (Billones de pesos)
2006	**1**	9597	1,46
	2	7782	1,17
2007	**1**	6449	1,00
	2	4123	0,45
2008	**1**	3259	0,51
	2	3192	0,48
2009	**1**	2908	0,45
	2	3202	0,41
2010	**1**	1901	0,18

Otro aspecto importante a analizar es el total de los montos transacciones como porcentaje de los saldos vigentes. La Tabla 4 nos muestra cómo en general, este porcentaje es siempre menor al 10% y en promedio tiende a disminuir.

Tabla 4: porcentaje de montos transados sobre el total vigente y total vigente.

Año	Período	Porcentaje transado	Total vigente (Billones de pesos)
2006	**Trim.1**	9,41%	10,19
	Trim.2	6,35%	10,14
	Trim.3	6,03%	10,23
	Trim.4	6,27%	10,27
2007	**Trim.1**	5,45%	10,04
	Trim.2	5,00%	10,04
	Trim.3	2,34%	9,95
	Trim.4	2,53%	10,00
2008	**Trim.1**	2,43%	9,66
	Trim.2	3,12%	9,47
	Trim.3	2,69%	9,61
	Trim.4	3,01%	9,70
2009	**Trim.1**	2,80%	9,36
	Trim.2	2,97%	9,16
	Trim.3	2,72%	9,04
	Trim.4	2,56%	8,90
2010	**Trim.1**	1,55%	8,61
	Trim.2	1,02%	8,55

Esta disminución en la cantidad de transacciones hace más compleja la tarea de encontrar modelos de valorización debido a la falta de información. También, la necesidad de encontrar precios justos para estos instrumentos no desaparece debido a que el total vigente de instrumentos no disminuye como las transacciones, y debe ser valorizado día a día.

Por otro lado, esta característica no afecta mayormente a nuestro estudio debido a que éste puede ser extendido a cualquier otra familia de papeles sin mayores inconvenientes. Otra razón por la que no afecta es debido a que se estima que algunas modificaciones reglamentarias, como aumentar el límite de financiamiento, puedan incrementar en un futuro cercano los volúmenes de transacciones.

Las transacciones son utilizadas en su totalidad sin agrupaciones de ningún tipo, como por ejemplo plazos, montos, clasificaciones u otras, debido que al hacerlo se perdería mucha información por culpa de la consolidación en dichos grupos. Por ejemplo, si se agrupara por plazo, se perdería la información que aporta cada una de las otras características como la tasa de emisión, prepagos o emisor.

El modelo se implementa para el mediano y largo plazo de la curva de spreads, esto debido a que en el corto plazo existen factores distintos que determinan el comportamiento de la curva promedio, junto a una disminución en la liquidez de transacciones, lo que dificulta la aplicación de la metodología para papeles cercanos a su madurez. De esta forma se tomarán únicamente transacciones con plazo al vencimiento mayor a dos años, obteniendo una muestra final que incorpora un 97,6% de las observaciones factibles y un 99,8% de los montos transados durante el período.

Las características que serán usadas para la explicación del comportamiento desviado de la media mediante el árbol de regresión fueron presentadas en el capítulo 5.2 donde se discute la segunda etapa del modelo planteado en esta investigación. La Tabla 5 muestra la cantidad de transacciones (y el porcentaje del total que éstas representan) agrupadas por clasificación. En ésta se puede observar que para los años 2006 y 2007 la mayor cantidad de transacciones se concentra en las clasificaciones AA+ y AA- seguidas por la clasificación A. Luego, el año 2008 ya no presenta una cantidad significativa en la clasificación A pero sí en la clasificación AAA, y finalmente en los años 2009 y 2010 la mayoría de éstas se concentran en las clasificaciones AA- y AAA. Esta evolución puede explicarse por la última crisis financiera que se vivió a nivel mundial, denominada crisis

subprime, la cual no sólo afectó al mercado hipotecario sino que a todos los mercados tanto financieros como reales.

Tabla 5: cantidad y porcentaje del total de transacciones agrupadas por emisores para cada año

	2006	**2007**	**2008**	**2009**	**2010**
A	3096 (17,8%)	1720 (16,3%)	591 (9,2%)	157 (2,6%)	24 (1,3%)
A-	167 (1%)	45 (0,4%)	14 (0,2%)	87 (1,4%)	31 (1,6%)
A+	210 (1,2%)	940 (8,9%)	298 (4,6%)	0 (0%)	0 (0%)
AA	925 (5,3%)	463 (4,4%)	152 (2,4%)	6 (0,1%)	0 (0%)
AA-	4602 (26,5%)	2924 (27,7%)	2453 (38%)	2287 (37,4%)	832 (43,8%)
AA+	8379 (48,2%)	4479 (42,4%)	1157 (17,9%)	405 (6,6%)	103 (5,4%)
AAA	0 (0%)	1 (0%)	1786 (27,7%)	3168 (51,8%)	911 (47,9%)
Total	100%	100%	100%	100%	100%

Por otro lado, la Tabla 6 muestra la cantidad de transacciones (y el porcentaje del total que éstas representan) agrupadas por emisor. En ésta se puede observar que en todos los años los tres principales emisores presentes en las transacciones son el Banco del Estado, el Banco del Desarrollo y Banco Corpbanca con un promedio de 34,6%, 17,8% y 12,2% respectivamente.

Tabla 6: cantidad y porcentaje del total de transacciones agrupadas por emisores para cada año.

	2006	**2007**	**2008**	**2009**	**2010**	**Promedio**
ESTADO	4276 (24,6%)	3019 (28,6%)	2255 (35%)	2737 (44,8%)	765 (40,2%)	34,6%
DESARROLLO	2816 (16,2%)	1811 (17,1%)	1352 (21%)	1107 (18,1%)	316 (16,6%)	17,8%
CORPBANCA	2562 (14,7%)	1624 (15,4%)	672 (10,4%)	445 (7,3%)	252 (13,3%)	12,2%
CHILE	1209 (7%)	597 (5,6%)	327 (5,1%)	254 (4,2%)	67 (3,5%)	5,1%
SANT-CHILE	1218 (7%)	588 (5,6%)	225 (3,5%)	213 (3,5%)	50 (2,6%)	4,4%
BOSTON	1357 (7,8%)	300 (2,8%)	153 (2,4%)	108 (1,8%)	40 (2,1%)	3,4%
FALABELLA	388 (2,2%)	601 (5,7%)	349 (5,4%)	409 (6,7%)	110 (5,8%)	5,2%
SECURITY	760 (4,4%)	425 (4%)	151 (2,3%)	118 (1,9%)	74 (3,9%)	3,3%
CREDITO	694 (4%)	342 (3,2%)	145 (2,2%)	236 (3,9%)	84 (4,4%)	3,6%
RIPLEY	343 (2%)	384 (3,6%)	290 (4,5%)	106 (1,7%)	31 (1,6%)	2,7%
Otros	1756 (10,1%)	881 (8,3%)	532 (8,2%)	377 (6,2%)	112 (5,9%)	7,7%
Total	100%	100%	100%	100%	100%	

Finalmente, la Tabla 7 muestra el promedio y desviación estándar del resto de las variables explicativas. Se puede observar que la cantidad de prepagos en períodos de tiempo cortos son bajos en todos los años de la muestra, la tasa de emisión bordea los 4,6% con baja desviación dentro de cada año y a lo largo de éstos, y que los plazos transados más comúnmente son medianos y largos. También podemos observar que la Rotación de papeles es baja para todos los años menos el año 2010 y que la presencia tiene una alta desviación estándar, lo que nos indica que hay grupos de letras que se transan significativamente más que otros grupos de letras que se transan menos.

Tabla 7: promedio y desviación estándar de las probabilidades de prepago, plazo, tasas de emisión, presencia y rotación para cada año.

	2006	**2007**	**2008**	**2009**	**2010**
Probabilidad (N=10)	10 % (0,19)	9,5 % (0,19)	9,2 % (0,19)	12 % (0,21)	11,2 % (0,2)
Probabilidad (N=5)	10 % (0,19)	9,4 % (0,19)	9,1 % (0,19)	11,4 % (0,2)	10,2 % (0,18)
Probabilidad (N=1)	4,6 % (0,1)	3 % (0,07)	3,1 % (0,08)	3,2 % (0,07)	2,8 % (0,07)
Plazo (Años)	13,6 (5,88)	14,5 (6,88)	15,7 (7,17)	15,2 (7,23)	13,9 (6,98)
Tasa de Emisión (%)	4,8 (0,88)	4,6 (0,87)	4,5 (0,75)	4,7 (0,78)	4,4 (0,79)
Presencia (cantidad)	25,5 (44,65)	28,6 (49,41)	32,5 (49,38)	37,1 (57,77)	33,4 (55,2)
Rotacion (%)	0,6 (0,74)	0,5 (0,8)	0,4 (0,54)	0,3 (1,7)	1,6 (9,05)

6.2 La crisis financiera

El período que se analiza en esta investigación (2006-2010) incluye una de las principales crisis que han sufrido tanto los mercados financieros cómo reales. Es por esto que este capítulo realiza un análisis de cómo esta crisis afecta al mercado financiero Chileno y, en particular, al mercado de las letras de crédito hipotecario.

En agosto del año 2008 se llega al punto crítico de la crisis de hipotecas *subprime* que afectó tanto a los mercados financieros como a los reales a lo largo de todo el mundo y que había comenzado a influir en los primeros desde mediados del año 2007.

Esta crisis fue precedida por períodos de alzas sostenidas en las tasas de interés, seguidos de una abrupta baja de las mismas a partir del cuarto trimestre del año 2008. La Ilustración 4 muestra la evolución de la TPM o Tasa de Política Monetaria del Banco Central de Chile. En ésta se puede observar claramente como a finales del año 2008 hay un cambio importante en el nivel de ésta.

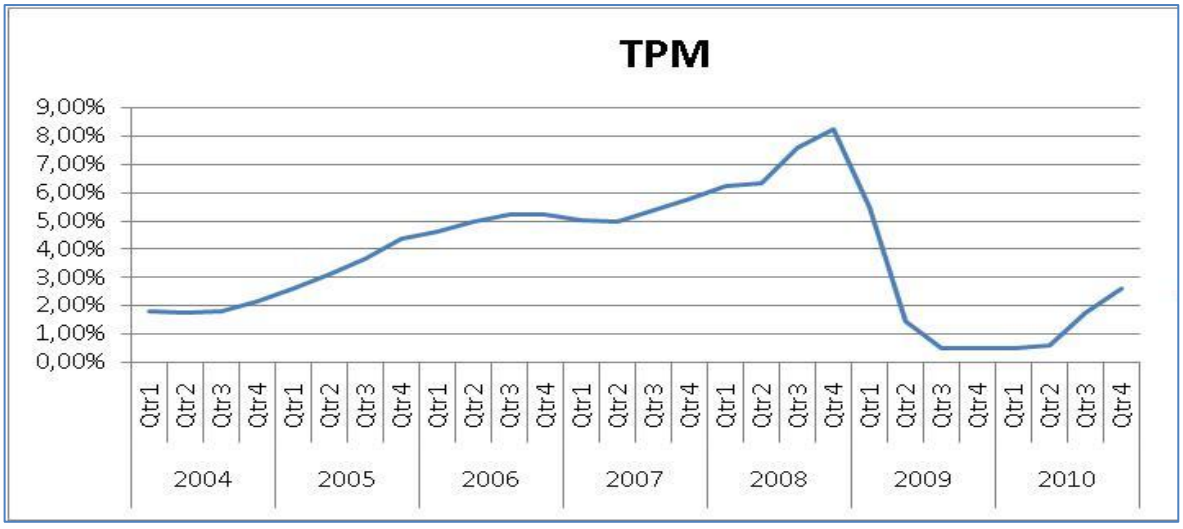

Ilustración 4: evolución de la Tasa de Política Monetaria del Banco Central de Chile.

Este fenómeno afectó primero a los mercados financieros y luego a los mercados reales, reflejándose en bajas de consumo de la gente, quiebras de empresas (por ejemplo *Lehman Brothers*), etc. En particular, el mercado de las letras hipotecarias se vio también afectado observándose cambios en los niveles de transacciones y en la volatilidad de éstas.

La Ilustración 5 nos muestra como evolucionó el promedio de las TIR de transacción para letras de plazos entre 2-12, 12-22 y 22-30 años. En esta podemos observar un marcado aumento en el nivel a finales del año 2008.

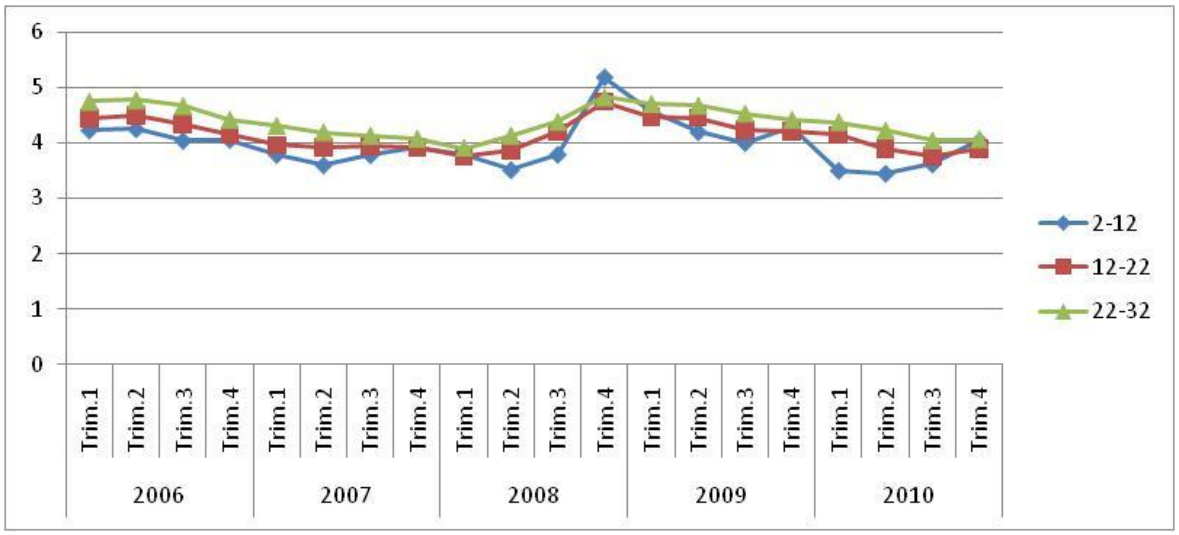

Ilustración 5: evolución del promedio de transacción de letras hipotecarias con plazos entre 2-12, 12-22 y 22-32 años.

Por otro lado, la Ilustración 6 nos muestra ahora la evolución de las desviaciones estándar de las transacciones para los mismos grupos de plazos. En ésta podemos ver que las transacciones de plazos más cortos presentan en promedio un aumento de estas desviaciones, mientas que las de plazos más largos no tienen un aumento marcado. Este aumento de volatilidad se puede observar también en la Ilustración 7 donde se muestra el índice VIX, el cual mide la volatilidad implícita en las opciones financieras, el cual llega a su máximo en 5 años el 13 de Octubre del año 2008.

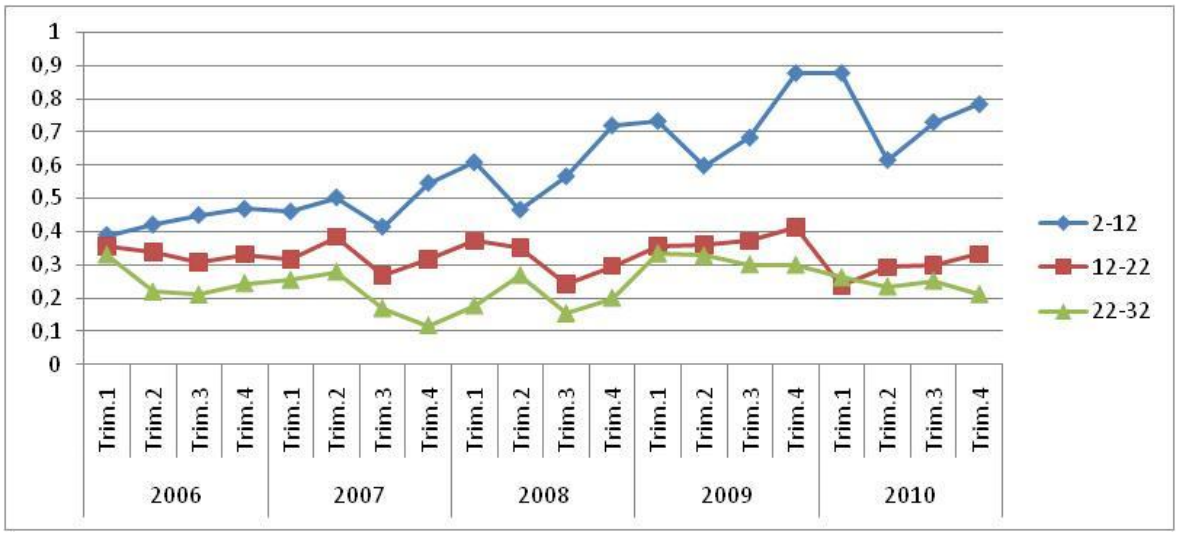

Ilustración 6: evolución de la desviación estándar de las transacciones de letras hipotecarias con plazos entre 2-12, 12-22 y 22-32 años.

Ilustración 7: evolución del índice VIX desde comienzos del año 2006 hasta la fecha.

Todo lo expuesto anteriormente nos da indicios de cómo la crisis financiera introdujo al mercado cambios, o nuevos escenarios, que hacen más compleja y desafiante la tarea de realizar cualquier tipo de análisis sobre los mercados financieros.

6.3 Implementación

Para implementar la metodología primero es necesario definir el número de factores que se van a utilizar para el modelo dinámico. Dada la similitud del procedimiento con el de la estimación de una curva libre de riesgo, se utilizaran los mismos 3 factores que se utilizan en Cortazar, Schwartz, & Naranjo (2007).

Para la implementación del modelo de árboles de regresión, se utilizarán como variables explicativas todas aquellas descritas en la sección 5.2, donde se describe la segunda etapa del modelo propuesto, dejando que el algoritmo decida por si mismo cuáles son las variables que tomará en cuenta y cuáles no.

Como métrica de evaluación, se usa el ajuste de cada una de las etapas del modelo, medido como el error promedio, para conocer el sesgo, y el error absoluto medio, para conocer el poder de ajuste. Se ocupa esta medida ya que el error o desviación de la curva promedio es utilizado como *input* en forma directa para la construcción del árbol de regresión.

Por otro lado, dada la gran importancia que tiene la curva promedio sobre la cual se obtienen las desviaciones, es de vital importancia que sea lo más representativa de los datos posible. Por esto, se mide la calidad de los parámetros estimados obteniendo la estructura de volatilidad implícita del modelo y comparándola con la volatilidad empírica de los datos.

Finalmente a modo de validación, se comparan los resultados obtenidos por el modelo propuesto con aquellos obtenidos por las metodologías alternativas.

6.3.1 Estimación del modelo

Los parámetros del modelo dinámico (primera etapa) se estiman usando la metodología descrita en la sección 2.1.2.2, donde se explica la extensión del filtro de Kalman para una ecuación de medida no línea, usando las tasas internas de retorno TIR de las transacciones dentro de la muestra.

Posteriormente, la estimación del árbol de regresión (segunda etapa) se hace con la implementación del algoritmo de árboles de regresión existente en el *software SQL Server Business Intelligence*. Para evitar la sobre estimación del modelo y evitar así un pobre desempeño fuera de muestra, se fuerza al algoritmo a tener un mínimo de un 1,5% de casos en cada nodo hoja.

El experimento que se realizará utilizará datos de un año para la estimación de todos los parámetros (modelo dinámico y árbol de regresión) para luego usar el semestre siguiente como período de estimación. Este experimento se repetirá cuatro veces, calibrando primero con el año 2006 y estimando el primer semestre del año 2007, luego lo mismo para el año 2007 y primer semestre del 2008, para el año 2008 y primer semestre del 2009 y finalmente para el año 2009 y primer semestre del 2010.

La Tabla 8 muestra los parámetros obtenidos para el modelo dinámico para los cuatro experimentos que se hicieron. Lo primero que se puede observar es la alta similitud entre ellos. Dado que lo que se está midiendo es el comportamiento promedio del mercado de las letras hipotecarias, es razonable pensar que éstas no sufrieron grandes cambios en sus dinámicas. Se puede observar también que en todos los casos las variables de estado correspondientes a explicar el corto plazo son tanto x_2 y x_3 debido a que los valores de κ_2

y κ_3, que representan la velocidad de reversión de la variable de estado a cero, son parecidos y mayores que κ_1, siendo ésta última la que se encarga de explicar el largo plazo. Por otro lado, en todos los casos, las correlaciones de las dos variables largas con la variable corta son similares y bajas, mientras que la correlación entre ellas es mayor. Esto resultado resulta intuitivo dado que ambas variables realizan un trabajo similar.

La Tabla 9 nos muestra la variación porcentual de todos los parámetros de un año a otro. En esta podemos observar que los mayores cambios se dan en las correlaciones de la variable x_1 con la variable x_2, de la variable x_1 con la variable x_3. También en el parámetro δ de la tendencia a largo plazo, lo que nos dice que el nivel de la curva obtenida con las transacciones en el largo plazo cambia año a año. Finalmente, el parámetro υ también sufre grandes cambios. Éste representa la raíz cuadrada de la varianza de los errores de medición, por lo que se puede deducir que el *spread* de las transacciones sobre o bajo la curva promedio tiende a variar de un año a otro.

Tabla 8: parámetros estimados para el modelo dinámico para los cuatro años.

Parámetro	2006	2007	2008	2009
κ_1	2,01278	2,01284	2,01280	2,01278
κ_2	5,98586	5,98576	5,98578	5,98584
κ_3	6,00020	6,00022	6,00022	6,00021
σ_1	0,50868	0,50926	0,50909	0,50867
σ_2	0,50222	0,50229	0,50230	0,50225
σ_3	0,52918	0,52925	0,52926	0,52921
ρ_{12}	0,00120	0,00130	0,00125	0,00118
ρ_{31}	0,00127	0,00137	0,00133	0,00125
ρ_{32}	0,03451	0,03454	0,03455	0,03452
λ_1	-0,18507	-0,18295	-0,18341	-0,18471
λ_2	0,81255	0,81326	0,81310	0,81267
λ_3	-0,18748	-0,18677	-0,18693	-0,18735
δ	0,09856	0,09430	0,09531	0,09782
υ	0,00352	0,00358	0,00344	0,00430

Tabla 9: variación de los parámetros estimados para el modelo dinámico para los cuatro años

Parámetro	2006-2007	2007-2008	2008-2009
κ_1	0,003%	-0,002%	-0,001%
κ_2	-0,002%	0,000%	0,001%
κ_3	0,000%	0,000%	0,000%
σ_1	0,115%	-0,034%	-0,082%
σ_2	0,013%	0,003%	-0,010%
σ_3	0,012%	0,003%	-0,009%
ρ_{12}	7,825%	-3,773%	-5,695%
ρ_{31}	7,774%	-3,752%	-5,652%
ρ_{32}	0,093%	0,020%	-0,070%
λ_1	-1,156%	0,247%	0,705%
λ_2	0,088%	-0,020%	-0,052%
λ_3	-0,381%	0,088%	0,224%
δ	-4,513%	1,050%	2,573%
υ	1,660%	-4,272%	20,122%

Por otro lado, la estimación del modelo de árboles de regresión (etapa 2) dio como resultados árboles similares al del año 2008 mostrado en la Ilustración 8. En ésta se muestran sólo 3 niveles de extensión y la oscuridad de cada nodo representa la densidad de letras con esas características relativa al resto de los nodos. Por otro lado, la Ilustración 9 muestra un ejemplo del camino que seguiría una transacción de letra de crédito emitida por el Banco del Estado a una tasa de 5,5%.

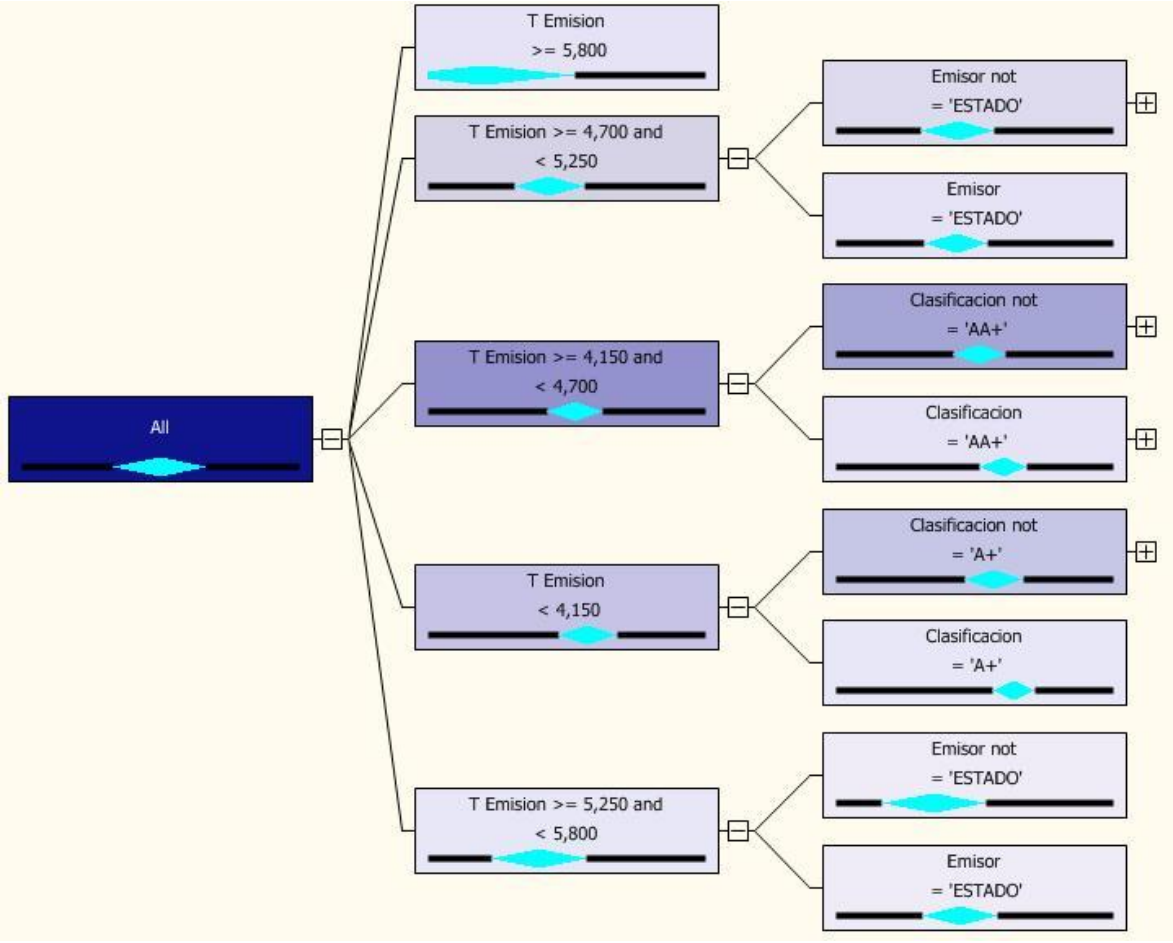

Ilustración 8: árbol de regresión estimado para el año 2009 extendido sólo en tres niveles.

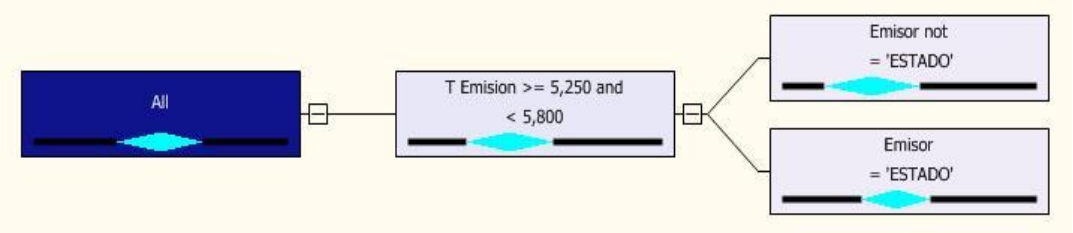

Ilustración 9: ejemplo del camino (completo) seguido por una LCH emitida por el Banco del Estado a una tasa de 5,5 % para el año 2009

La Ilustración 8 nos da ya algunas luces del funcionamiento del mercado Chileno de letras hipotecarias (para el año 2009). De ésta podemos observar que la primera variable que se toma en cuenta es la tasa de emisión. Esto es razonable ya que de ésta depende principalmente el prepago: si la tasa de emisión es baja, entonces las condiciones de mercado que se tienen que dar para que uno de estos instrumentos caiga en prepago son más estrictas, mientras que si la tasa de emisión es alta, la probabilidad que esto ocurra es mayor. Por otro lado, nos dice también que las letras que tienen tasas de emisión más bajas tienden a tener un comportamiento más parecido a los bonos que no tienen probabilidad de prepago: estos dependen principalmente de su clasificación de riesgo. Esto se ve representado en la Ilustración 8 donde los dos nodos que separan a las transacciones de letras con menor tasa de emisión tienen como segundo atributo de decisión a la clasificación de riesgo, mientras que los otros no.

Finalmente, la Ilustración 9 nos muestra el ejemplo de camino completo que tomaría una LCH que fue emitida por el Banco del Estado de Chile a una tasa del 5,5%. En este caso particular, el árbol nos dice que los instrumentos emitidos por este banco a esa tasa tienen un comportamiento similar en su desviación de la curva promedio y que ninguna

otra variable (como su clasificación o plazo) aporta información significativa para su determinación.

Más adelante se completará este análisis con uno cuantitativo del aporte de cada nodo del camino que toma cada papel en la determinación de su desviación del promedio.

6.4 Ajuste del modelo a las transacciones

A continuación se analiza el ajuste del modelo a las transacciones observadas dentro del período de la muestra. Para hacer esto el primer paso consiste en observar cómo la curva promedio logra captar el comportamiento de las transacciones de LCH. En segundo lugar, observar cómo el árbol realiza la corrección (junto al ajuste por sesgo diario). Debido a que esta corrección consiste en sumarle a la curva promedio la predicción hecha por el árbol, es difícil ver los resultados gráficamente. Para resolver este problema, la corrección del árbol en vez de sumársela a la curva se le restará a cada transacción de forma que ésta se acerque a la curva y sea más fácil ver gráficamente el resultado. Finalmente se hace un análisis cuantitativo de los ajustes por etapa del modelo.

De la Ilustración 10 a la Ilustración 17 se muestra un ejemplo dentro del período de calibración y otro fuera de éste para cada uno de los cuatro experimentos que fueron hechos. En el eje de las ordenadas se encuentran la TIR de transacción, la TIR de transacción corregida y las estructuras de tasas cero[9] que encuentra el modelo para cada día. En el eje de las abscisas la duración de los papeles. Las transacciones se grafican de acuerdo a su duración para poder comparar las transacciones (que presentan cupones

[9] La tasa cero se obtiene asumiendo que todos los pagos se efectúan concentrados en la duración.

dependiendo de su tasa de emisión) con la curva cero obtenida (que es una curva cero). El ANEXO D: muestra que usar la duración en este caso no presenta ningún problema dado que dos letras con igual TIR de transacción pero con diferente tasa de emisión (y por ende cupones de distinta magnitud entre ellas) van a presentar igual duración.

Los ejemplos escogidos intentan ser representativos de toda la muestra al ser seleccionados lo más uniformemente posible en el tiempo y tratando de mostrar distintas formas que toma la curva promedio.

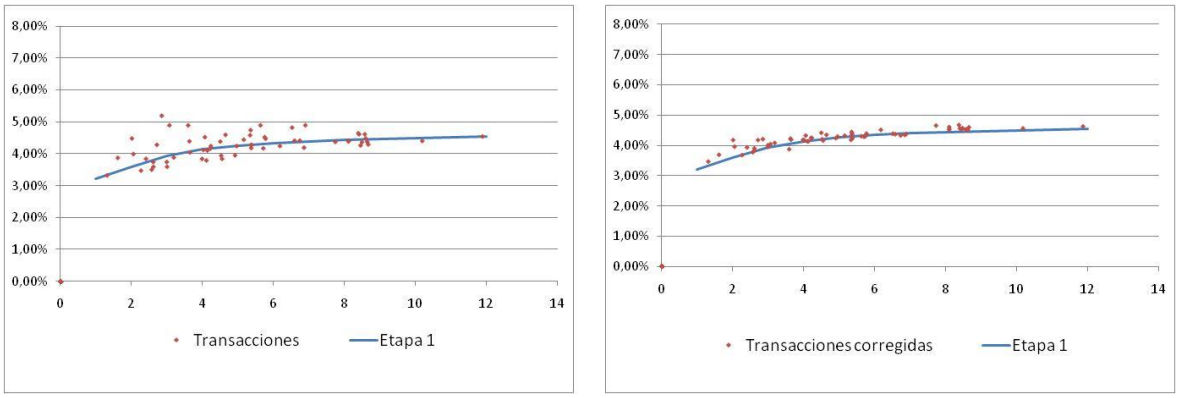

Ilustración 10: las tres etapas del modelo para el día 21-07-2006

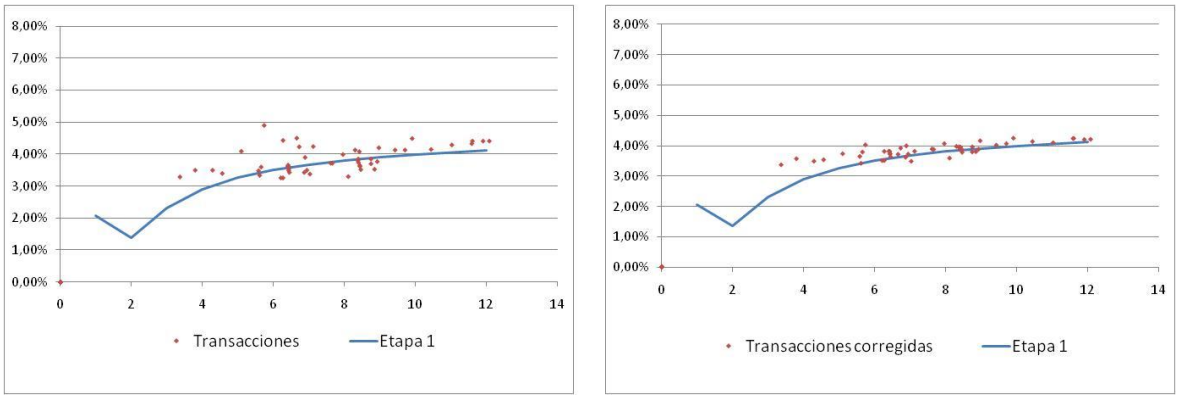

Ilustración 11: las tres etapas del modelo para el día 10-04-2007

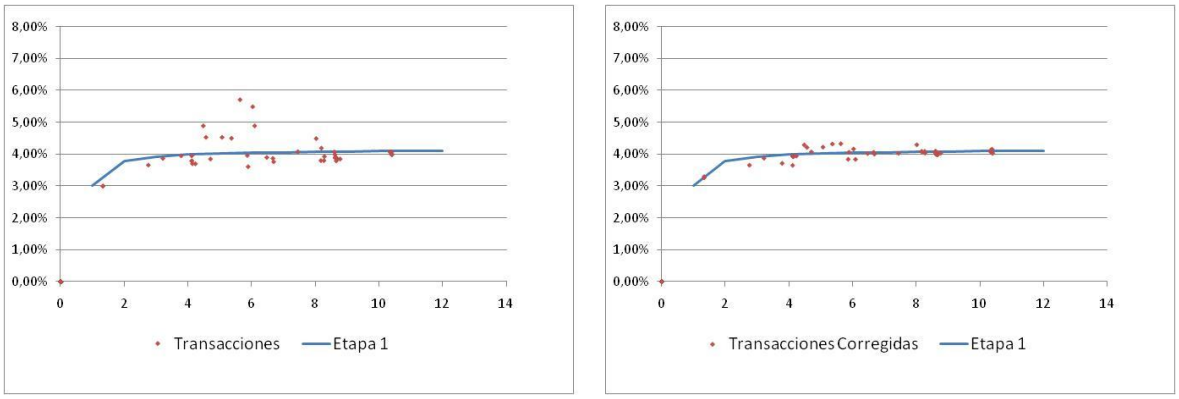

Ilustración 12: las tres etapas del modelo para el día 11-07-2007

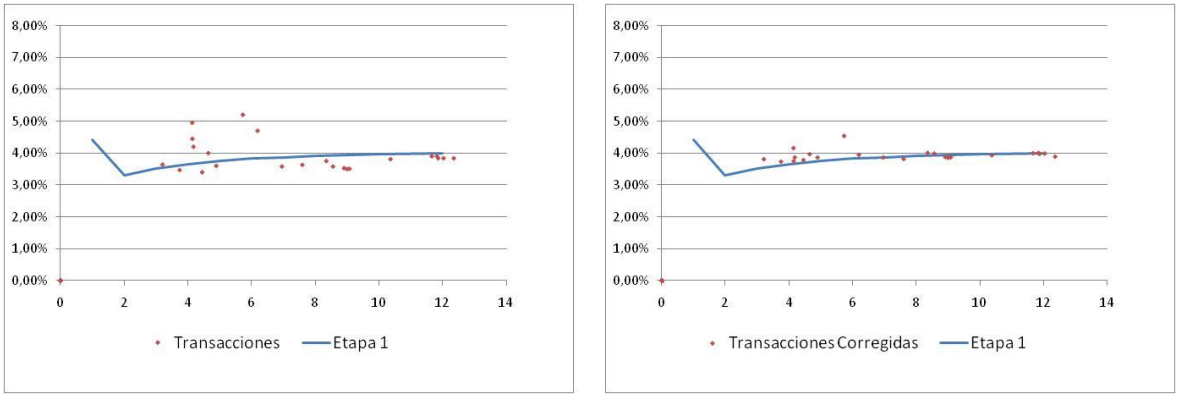

Ilustración 13: las tres etapas del modelo para el día 27-02-2008

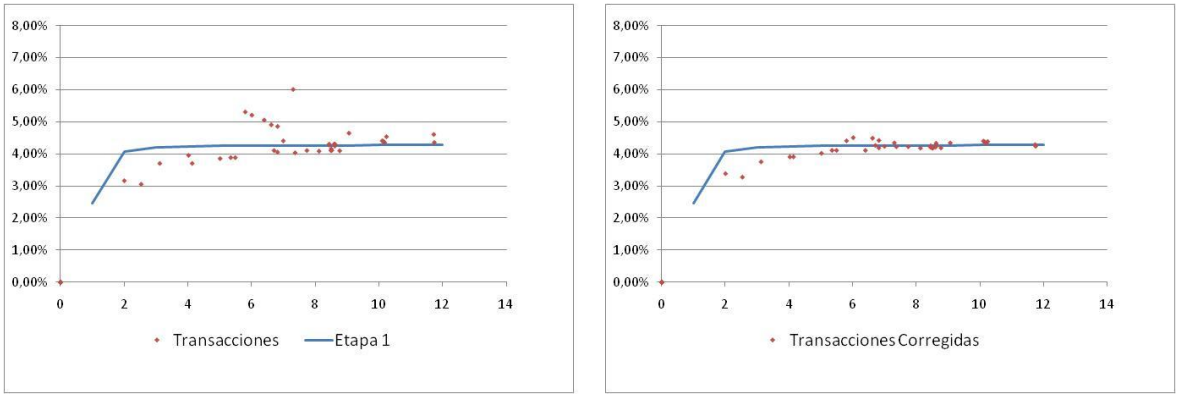

Ilustración 14: las tres etapas del modelo para el día 14-08-2008

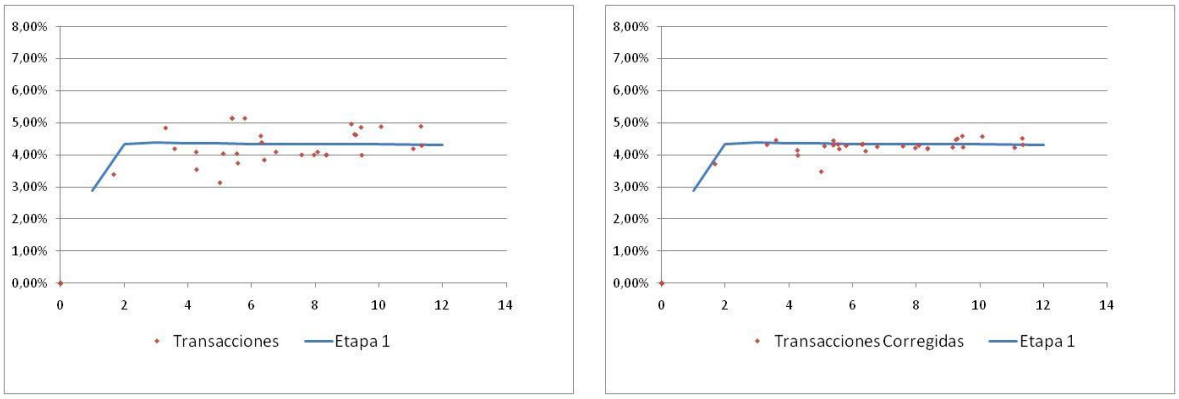

Ilustración 15: las tres etapas del modelo para el día 02-04-2009

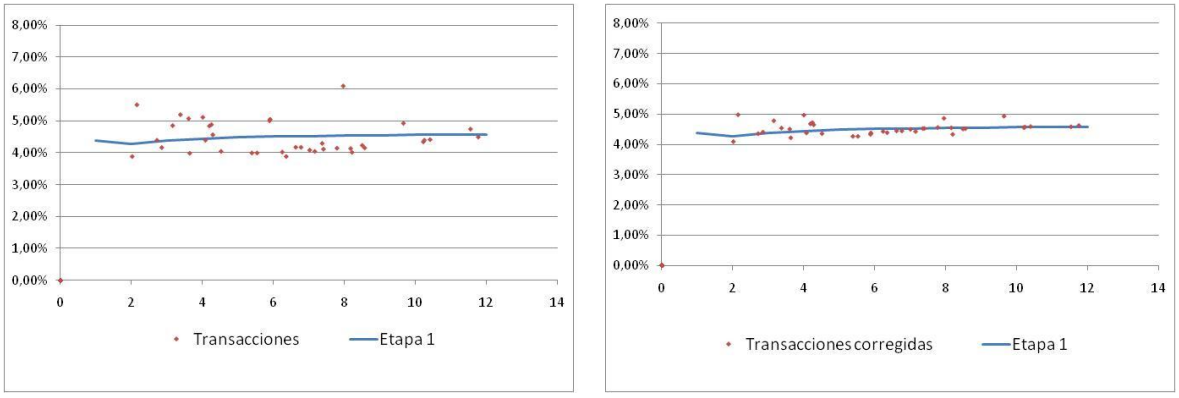

Ilustración 16: las tres etapas del modelo para el día 22-07-2009

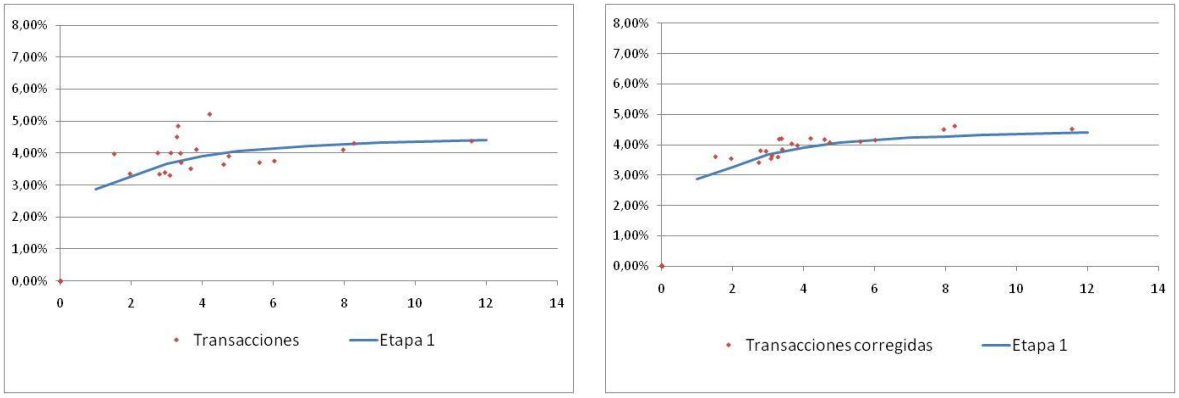

Ilustración 17: las tres etapas del modelo para el día 03-02-2010

Lo primero que se puede destacar son los cuadros de la izquierda en cada ejemplo. En estos se observa cómo la curva promedio tiende, en general, a seguir al promedio de las transacciones, lo cual es de suma importancia ya que para poder trabajar posteriormente sobre las desviaciones se requiere una buena base. Es importante destacar también que esta curva puede estar sesgada, sin significar esto que el modelo esté dando resultados incorrectos o que los parámetros estén mal calibrados, sino que, el modelo al permitir errores de medición, tiende a no considerar el total de la nueva información presentando por ende una especie de inercia. También cabe destacar el nivel de dispersión de las transacciones, el cual es muy alto, pudiendo llegar a tener diferencias entre transacciones de plazos similares de hasta 200 puntos base. Esta característica, como se explicó anteriormente, era deseada ya que es lo que se intenta corregir con el árbol de regresión.

En segundo lugar, se puede apreciar cómo las transacciones al ser corregidas se tienden a mover en dirección a la curva promedio. Esto representa el resultado de la segunda y tercera etapa del modelo. En general se observa con claridad cómo la nube de transacciones (puntos) al pasar del cuadro de la izquierda al de la derecha (representando así la corrección del árbol de regresión y del sesgo diario) tiende a ser más delgada, siguiendo siempre la forma de la curva.

Para ver lo anterior mediante un análisis cuantitativo, la Tabla 10 presenta los errores que se cometen en cada etapa del modelo para los cuatro experimentos que se realizaron. Cabe recordar que el experimento consta de calibrar los parámetros del modelo con un año de datos (datos dentro de la muestra o *In-Sample*) para luego realizar estimaciones durante el primer semestre del año siguiente (datos fuera de la muestra o *Out-of-Sample*).

Tabla 10: error absoluto medio para los cuatro experimentos realizados y las tres etapas del modelo

Período	Tipo	Etapa 1	Etapa 2	Etapa 3
2006-2007	**In-Sample**	0,263%	0,113%	0,095%
	Out-of-Sample	0,271%	0,149%	0,126%
2007-2008	**In-Sample**	0,248%	0,124%	0,101%
	Out-of-Sample	0,233%	0,177%	0,137%
2008-2009	**In-Sample**	0,231%	0,161%	0,140%
	Out-of-Sample	0,285%	0,179%	0,167%
2009-2010	**In-Sample**	0,304%	0,168%	0,139%
	Out-of-Sample	0,270%	0,196%	0,147%

En la tabla se observa en un primer lugar el error de ajuste que tiene la curva promedio con las transacciones antes de ser corregidas. Para el período dentro de muestra llega en promedio cerca de los 26 puntos bases mientras que en el período fuera de muestra sube a cerca de 27 puntos bases en promedio. Estos errores son consistentes con lo que se observa en los ejemplos mostrados anteriormente, donde se podían encontrar diferencias entre transacciones de plazo similar de hasta 200 puntos base (siendo imposible lograr un buen ajuste sólo con un modelo dinámico).

En segundo lugar, se observa como el árbol de regresión logra disminuir dicho error al apoyar a la curva promedio. Porcentualmente, la etapa 2 del modelo logra disminuir el error dentro de período de muestra y fuera de éste en un 45,6% y en un 44,6% promedio respectivamente a lo largo de los 4 experimentos, resultado que nuevamente es consistente con lo visto anteriormente.

En tercer lugar, la tabla muestra que también existe una disminución del error por parte del ajuste mediante el sesgo diario en la etapa 3 del modelo, llegando finalmente en promedio a 12 y 14,4 puntos base de error para los períodos dentro y fuera de muestra

respectivamente. Esta última etapa del modelo realiza una mejora porcentual en promedio de 16,25% y 17,46% para los períodos dentro y fuera de muestra respectivamente. Como podemos observar, esta mejora es más significativa en el período fuera de muestra, lo que se explica debido a que el árbol de regresión no incorpora nueva información después de ser calibrado, por lo que se desactualiza con mayor facilidad y por ende, el apoyo de la corrección final se hace más significativo.

Finalmente, se puede observar que el tercer experimento (2008-2009) es aquel que presenta el mayor error. Esto puede estar relacionado con la crisis financiera que se discutió en el punto 6.2, donde se muestra que a finales del año 2008 el promedio de las transacciones sufre un aumento brusco, y la desviación estándar, de las transacciones más cortas principalmente, también. A pesar de lo anterior, las diferencias entre los errores de ajuste de este experimento con los otros no son de gran magnitud, lo que nos da luces de que el modelo logra una buena respuesta ante escenarios globales como lo es una crisis financiera.

Otro resultado interesante que se puede obtener es cómo se distribuyen los errores al agrupar por plazo. En la Tabla 11 aparece el error absoluto medio del modelo agrupado por el plazo de las transacciones. De ésta se puede concluir rápidamente que los errores están concentrados, en todas las etapas, en los plazos más cortos. A pesar de lo anterior, esta estructura puede estar explicada por la existencia de mayor volatilidad en los plazos cortos y no realmente porque el modelo sea peor. Por esta razón la Tabla 12 muestra los errores absolutos medios divididos por la desviación estándar de las transacciones dentro de cada plazo.

Tabla 11: error absoluto y cantidad de transacciones para períodos dentro y fuera de muestra para cada experimento.

		2006		2007		2008		2009	
Muestra	**Plazo (años)**	**MAE**	**N° Trans.**	**MAE**	**N° Trans.**	**MAE**	**N° Trans.**	**MAE**	**N° Trans.**
In-Sample	**2-5**	0,231%	810	0,212%	678	0,352%	378	0,362%	311
	5-8	0,135%	2143	0,143%	1354	0,231%	747	0,192%	882
	8-11	0,091%	3633	0,122%	1872	0,160%	808	0,180%	803
	11-14	0,078%	3389	0,092%	1505	0,133%	740	0,140%	843
	14-17	0,084%	2025	0,078%	1186	0,102%	871	0,102%	873
	17-20	0,075%	3961	0,066%	2380	0,096%	1539	0,092%	1051
	20-23	0,096%	56	0,081%	53	0,126%	103	0,094%	204
	23-26	0,069%	839	0,069%	859	0,101%	686	0,093%	522
	26-29	0,085%	133	0,071%	252	0,108%	239	0,099%	291
	29-32	0,075%	390	0,066%	434	0,072%	340	0,071%	330
Out-of-Sample	**2-5**	0,284%	388	0,326%	233	0,346%	158	0,240%	167
	5-8	0,173%	812	0,213%	366	0,229%	406	0,199%	329
	8-11	0,119%	1153	0,145%	414	0,195%	369	0,169%	247
	11-14	0,117%	945	0,118%	402	0,161%	378	0,142%	281
	14-17	0,099%	681	0,096%	429	0,124%	421	0,100%	243
	17-20	0,107%	1530	0,093%	707	0,126%	471	0,090%	328
	20-23	0,080%	46	0,164%	51	0,144%	125	0,108%	74
	23-26	0,079%	482	0,112%	357	0,159%	282	0,116%	104
	26-29	0,080%	225	0,107%	178	0,108%	171	0,166%	90
	29-32	0,106%	188	0,100%	122	0,096%	127	0,101%	38

En esta se puede observar que el índice creado es mucho más estable que el error por si sólo lo que nos dice que el modelo tiende a tener resultados similares a lo largo de toda la estructura sin estar sesgado en este sentido. Para analizar de manera más fácil el comportamiento de este indicador la Ilustración 18 nos muestra el comportamiento promedio (de los cuatro años) del indicador para los períodos dentro y fuera de muestra en conjunto con el promedio, también dentro y fuera de muestra, de la cantidad de transacciones ocurridas. De esta se puede observar que efectivamente el indicador es más estable que el error absoluto medio y que existe una correlación negativa entre la cantidad

de transacciones y el índice. Esto último tiene sentido ya que los plazos que presenten más transacciones van a pesar más al momento de calibrar tanto el modelo dinámico como el árbol de regresión. Esto último no es un problema ya que la diferencia del indicador cuando hay muchas transacciones y pocas no es significativa.

Tabla 12: error absoluto para períodos dentro y fuera de muestra ajustados por desviación estándar para cada experimento.

		2006	2007	2008	2009
Muestra	**Plazo (años)**	$MAE/_{\sigma}$	$MAE/_{\sigma}$	$MAE/_{\sigma}$	$MAE/_{\sigma}$
In-Sample	**2-5**	1,11	0,90	0,96	1,08
	5-8	0,97	0,91	1,01	1,10
	8-11	0,98	0,92	0,96	1,00
	11-14	0,97	0,88	0,94	1,03
	14-17	1,05	1,07	1,03	0,97
	17-20	1,18	1,16	1,06	0,96
	20-23	1,36	1,28	1,10	1,31
	23-26	1,08	1,28	1,22	0,89
	26-29	1,32	0,82	1,00	1,07
	29-32	1,17	1,21	1,10	1,00
Out-of-Sample	**2-5**	1,08	1,05	1,07	1,23
	5-8	1,13	0,96	1,15	1,23
	8-11	1,05	0,93	1,11	1,04
	11-14	0,94	0,88	0,96	1,26
	14-17	1,17	1,20	0,87	1,14
	17-20	1,20	1,17	1,01	1,12
	20-23	1,08	1,16	1,11	1,30
	23-26	1,28	1,33	0,82	1,01
	26-29	1,38	1,18	0,95	0,82
	29-32	1,05	1,29	1,22	1,84

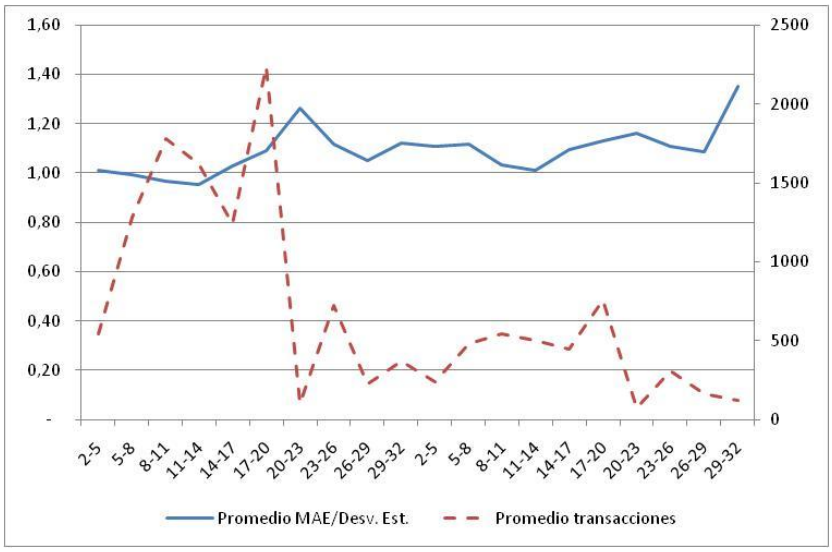

Ilustración 18: promedio del indicador $^{MAE}/_{\sigma}$ y de la cantidad de transacciones a lo largo de todos los experimentos.

6.4.1 Ajuste de la estructura de volatilidad

La estructura de volatilidad es la curva que relaciona las volatilidades para distintos plazos. En un modelo dinámico (como lo es la primera etapa), un buen ajuste de esta estructura es muy importante para lograr una correcta representación de los datos, puesto que la volatilidad de las variables de estado está directamente ligada a la fórmula de valorización. Por otro lado este ajuste determina que la actualización en serie de tiempo de los datos, cuando no existen observaciones, sea consistente con las varianzas y correlaciones observadas históricamente, logrando predicciones del filtro más creibles ante la ausencia de observaciones, pues es una medida de la estabilidad de las mismas.

El ANEXO C: muestra que para el modelo planteado, dada una dinámica Vasicek en las variables de estado, la estructura de volatilidad instantánea para los retornos de un bono, puede ser obtenida según la siguiente fórmula:

$$\sigma_p^2(t) = \sum_{i=1}^{N}\sum_{j=1}^{N}\sigma_i\sigma_j\rho_{ij}\left(\frac{e^{-k_i t}-1}{k_i}\right)\left(\frac{e^{-k_j t}-1}{k_j}\right)$$

(6-1)

Por otro lado, para calcular la volatilidad de las observaciones se usa la aproximación propuesta en Cortazar, Schwartz, & Naranjo (2007), donde se plantea una metodología para aproximar la estructura de volatilidad empírica en un mercado con paneles de datos incompleto. Para ello se agrupan las transacciones en intervalos según el plazo, calculando el promedio diario de cada intervalo y luego se obtiene la desviación estándar de las diferencias de observaciones diarias para cada uno de estos grupos. Adicionalmente se calcula la duración promedio del intervalo, generando la estructura como si fuera la volatilidad de un bono de descuento con madurez igual a esta duración. Debido a que los instrumentos usados presentan más de un pago en el tiempo, estos se agrupan por su duración de Macaulay, para considerar de antemano su representación como bono cero cupón. De esta forma se puede comparar directamente la estructura de volatilidad con la de la estructura generada por el modelo. Dado el proceso de reversión a una media de largo plazo, la estructura de volatilidad debería presentar una forma decreciente.

En la Ilustración 19 podemos ver la estructura de volatilidad tanto para la curva promedio como para las transacciones, para cada uno de los períodos dentro de muestra. En ella se observa, en cada caso, un buen ajuste de la volatilidad del modelo a la volatilidad de las transacciones.

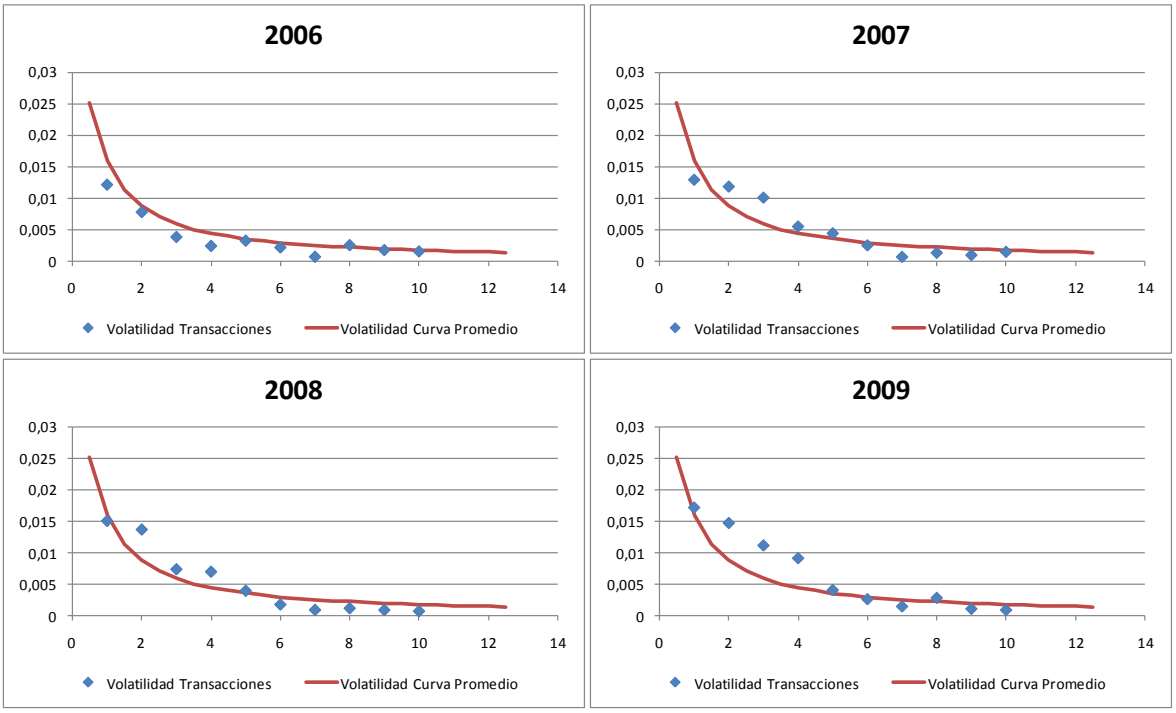

Ilustración 19: volatilidad de la curva promedio y de las transacciones para cada período dentro de muestra.

6.5 Comparación con metodología alternativa

Las metodologías alternativas que se utilizarán serán aquellas descritas en el punto 5.4, siendo una de ellas usar el último *spread* sobre la curva libre de riesgo y la otra utilizar una regresión lineal para reemplazar la labor del árbol. Es importante recordar que éstas sólo pretenden evaluar el desempeño del modelo de árbol de regresión y no los modelos dinámicos.

6.5.1 Mantención del último *spread* específico

La forma de obtener este resultado, detallada en el punto 5.4.1, es corregir el error cometido por la curva promedio utilizando el último *spread* específico del mismo papel.

Vale decir, para cada transacción se busca la transacción más cercana de ese mismo papel y se usa como *spread* la diferencia entre la transacción y dicha curva.

Cabe destacar que debido a que no hay parámetros que tengan que ser estimados, los resultados fuera de muestra presentados para esta metodología alternativa en realidad no lo son, por lo que podría dar la intuición de que tienen buen comportamiento fuera de muestra o malo dentro de ésta.

La Tabla 13 muestra la comparación del resultado obtenido por el modelo planteado en esta investigación junto al de la metodología alternativa. Como se puede observar, al modelo propuesto es ampliamente superior tanto dentro de muestra como fuera de muestra (comparando con el período dentro de muestra de la metodología alternativa). En general, la metodología alternativa es un 56% peor dentro de la muestra y un 27% peor fuera de ella aproximadamente.

Tabla 13: comparación del modelo con la primera metodología alternativa.

Período	Tipo	Etapa 1	Etapa 2	Etapa 3	Metodología Alternativa
	In-Sample	0,263%	0,113%	0,095%	0,146%
2006-2007	**Out-of-Sample**	0,271%	0,149%	0,126%	0,167%
	In-Sample	0,248%	0,124%	0,101%	0,177%
2007-2008	**Out-of-Sample**	0,233%	0,177%	0,137%	0,221%
	In-Sample	0,231%	0,161%	0,140%	0,216%
2008-2009	**Out-of-Sample**	0,285%	0,179%	0,167%	0,214%
	In-Sample	0,304%	0,168%	0,139%	0,199%
2009-2010	**Out-of-Sample**	0,270%	0,196%	0,147%	0,213%

Por otro lado, no es posible obtener a partir de esta metodología información de la composición de las transacciones o del mercado en general como es posible obtener a partir del árbol de regresión.

6.5.2 Regresión

La forma de obtener este resultado está detallada en el punto 5.4.2, en el que se explica que para reemplazar la estimación del *spread* específico hecha por el árbol de regresión esta metodología realiza una regresión lineal sobre las características del instrumento.

La Tabla 14 muestra, a modo de ejemplo, los parámetros resultantes luego de aplicar la metodología alternativa sobre el período correspondiente al 2009. Podemos observar que los parámetros que entregó son bastante intuitivos. Por ejemplo, a medida que la tasa de emisión aumenta, el castigo sobre la curva promedio también lo hace. Se puede también observar el efecto contrario con las variables de clasificación A, esto último es fácil de explicar ya que una alta clasificación de riesgo hace que la probabilidad de caer en *default* sea menor por lo tanto sufra un menor castigo.

Tabla 14: parámetros resultantes de la estimación de la metodología alternativa para el año 2010

	Coeficiente	**Error estándar**	**estadístico T**
Intercepto	-0,01955	0,00025	-77,65396
Emisor = ESTADO	-0,00178	0,00016	-11,07996
Emisor = DESARROLLO	-0,00111	0,00013	-8,65275
Emisor = CORPBANCA	-0,00035	0,00016	-2,16975
Plazo	0,00005	0,00001	9,07336
Tasa de emisión	0,00453	0,00005	86,64289
Clasificación = AAA	-0,00130	0,00017	-7,52305
Clasificación = AA-	-0,00117	0,00014	-8,13016
Probabilidad de prepago (N=10)	-0,00159	0,00023	-6,76224
Rotación	-0,00008	0,00002	-4,13431

Por otro lado, podemos observar también que el mayor estadístico t (en valor absoluto) lo tiene el intercepto. Al igual que con el modelo de árboles de regresión esto no

necesariamente aporta mucha información ya que nos puede estar mostrando que hay un período en que el *spread* del mercado hipotecario con el de los bonos libres de riesgo es grande. El segundo mayor estadístico t es el de la tasa de emisión, lo que está en línea con lo obtenido por el árbol de regresión (al ser esa la primera variable siempre).

La Tabla 15 muestra el error de ajuste (bajo la misma medida) que tuvieron ambos modelos para cada año. Se puede observar que el error de la metodología alternativa es consistentemente mayor al del árbol de regresión, incluso cuando se le realiza la misma corrección diaria del sesgo que se le realiza a la metodología del árbol de regresión.

Tabla 15: error de ajuste del modelo de árbol de regresión, de regresión lineal y de regresión lineal corregida por el sesgo diario.

Período	Tipo	Modelo árbol de regresión	Regresión lineal	Regresión lineal-corrección sesgo
2006-2007	**In-Sample**	0,095%	0,151%	0,138%
	Out-of-Sample	0,126%	0,179%	0,159%
2007-2008	**In-Sample**	0,101%	0,153%	0,131%
	Out-of-Sample	0,137%	0,189%	0,146%
2008-2009	**In-Sample**	0,140%	0,180%	0,155%
	Out-of-Sample	0,167%	0,188%	0,179%
2009-2010	**In-Sample**	0,139%	0,185%	0,159%
	Out-of-Sample	0,147%	0,207%	0,157%

Finalmente, podemos decir que esta metodología aporta menos información que el árbol de regresión. A pesar de que ésta nos puede indicar mediante los parámetros que ajustó cuáles variables inciden más en la determinación de las desviaciones o *spreads* específicos, no nos puede decir si una variable importa más o menos dependiendo de otras. Esto sí es indicado por el árbol de regresión, el cual nos dice exactamente la manera en que las variables deben ser observadas y analizadas.

6.6 Análisis de los componentes de la TIR de las LCH

Una vez calibrado y probado el modelo resulta interesante analizar cuantitativamente como las características propias de las LCH y su combinación influyen en la determinación del *spread* específico sobre o bajo la curva de la familia de papeles. Hasta ahora, sólo se había podido obtener información cualitativa de este tipo, pero ahora es posible obtener a partir del modelo de árbol de regresión (el encargado de explicar estos *spreads*) una visión más objetiva.

Para realizar esto, primero es importante recalcar que los instrumentos financieros en general miden sus *spreads* sobre una curva libre de riesgo. Esta curva libre de riesgo es aquella que se puede obtener a partir de los instrumentos más seguros, vale decir, aquellos instrumentos emitidos por entidades gubernamentales (en el caso de Chile el Banco Central de Chile y la Tesorería General de la República) ya que tienen una baja probabilidad de entrar en *default* dado el respaldo que tienen.

Entonces, para analizar las desviaciones de las transacciones de la curva promedio comenzaremos primero por el aporte que hace, a la TIR de la estimación, la curva libre de riesgo, luego el *spread* de la familia de las LCH sobre la anterior y finalmente cada nodo que recorre la transacción a lo largo del camino que le corresponda en el árbol de regresión.

Otro análisis que se puede desprender del modelo es el comportamiento en promedio del mercado de las letras hipotecarias usando el resultado del árbol de regresión. La forma de hacer esto es analizar cómo este árbol, en promedio, asigna los *spreads* específicos a lo largo de cada una de sus ramas.

6.6.1 Obtención de la curva libre de riesgo

Para obtener la curva libre de riesgo se utiliza el mismo procedimiento de Cortazar, Schwartz, & Naranjo (2007) que se ha usado hasta ahora pero con papeles libres de riesgo tanto del Banco Central de Chile como de la Tesorería General de la República. En particular se usan los bonos *bullet* [10]nominados en UF BCU y BTU.

La estimación de parámetros para el mismo período, del 01-01-2006 al 30-06-2010, de transacciones de letras hipotecarias que se analiza se muestra en la Tabla 16, en conjunto con los parámetros encontrados para los cuatro experimentos hechos anteriormente. Se puede observar que la variable de estado x_3 en todos los casos se encarga de explicar el corto plazo (dada su rápida reversión a cero) pero que la variable x_2 en el modelo libre de riesgo ya no se encarga del corto sino del mediano plazo. Por otro lado, podemos observar que el parámetro υ_{LH} (que representa raíz cuadrada de la varianza de los errores de medición) es del orden de 10 veces mayor que sus análogos υ_{BCU} y υ_{BTU} lo que resulta intuitivo ya que la dispersión que se puede observar en estos últimos es mucho menor a la de las LCH. Esto ocurre ya que no existen opciones de ningún tipo y los riesgos por liquidez y *default* son menores para instrumentos libres de riesgo.

[10] Un bono *bullet* es aquel que paga sólo intereses en sus cupones salvo en el último que paga el valor nominal más el interés correspondiente a ese período.

Tabla 16: parámetros curva libre de riesgo y curva promedio para los cuatro años.

Parámetro	Curva libre de Riesgo	2006	2007	2008	2009
κ_1	0,0144	2,01278	2,01284	2,01280	2,01278
κ_2	0,4313	5,98586	5,98576	5,98578	5,98584
κ_3	6,9609	6,00020	6,00022	6,00022	6,00021
σ_1	0,0101	0,50868	0,50926	0,50909	0,50867
σ_2	0,0932	0,50222	0,50229	0,50230	0,50225
σ_3	0,1292	0,52918	0,52925	0,52926	0,52921
ρ_{12}	0,3419	0,00120	0,00130	0,00125	0,00118
ρ_{31}	-0,0818	0,00127	0,00137	0,00133	0,00125
ρ_{32}	-0,2519	0,03451	0,03454	0,03455	0,03452
λ_1	-0,0038	-0,18507	-0,18295	-0,18341	-0,18471
λ_2	0,0378	0,81255	0,81326	0,81310	0,81267
λ_3	-0,2686	-0,18748	-0,18677	-0,18693	-0,18735
δ	0,0181	0,09856	0,09430	0,09531	0,09782
υ_{LH}		0,00352	0,00358	0,00344	0,00430
υ_{BCU}	0,0007				
υ_{BTU}	0,0004				

En la Ilustración 20 se muestra un ejemplo del resultado obtenido de la curva libre de riesgo para el día 03-05-2006.

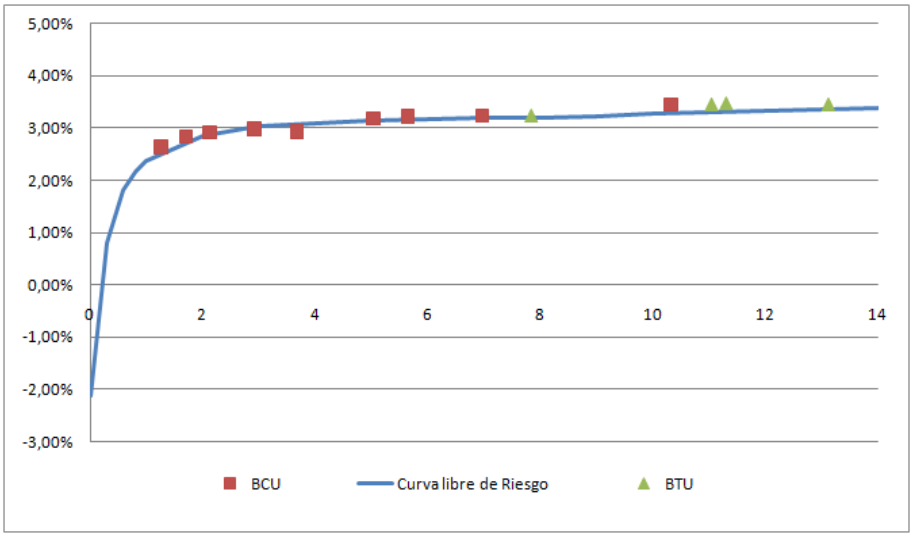

Ilustración 20: curva libre de riesgo y transacciones de bonos BTU y BCU para el día 03-05-2006.

6.6.2 Estudio de los *spreads* específicos

Dada la curva libre de riesgo, podemos realizar el análisis cuantitativo de estos spreads o desviaciones de la media de las transacciones.

Para lograr este objetivo, se descompone la TIR estimada para una transacción de la siguiente manera:

$$TIR_{estimada} = TIR_{Gob} + S_{Familia\ LH} + \sum_{i=1}^{n} (\eta_i | \eta_{i-1}) \quad (6\text{-}2)$$

donde TIR_{Gob} es la tasa interna de retorno de un papel libre de riesgo equivalente al mismo plazo, $S_{Familia\ LH}$ es el *spread* de la curva promedio por sobre la libre de riesgo, η_i es el spread que agrega el nodo i sobre η_{i-1} siendo η_0 la curva promedio y n el número de nodos que visita una letra desde la raíz del árbol de regresión hasta llegar al nodo hoja que le corresponde.

Para encontrar η_i se deben tomar en cuenta todas las transacciones (del período de entrenamiento) que llegaron al nodo i para así encontrar la regresión que va a estimar el aporte que le hace a la transacción que se está estudiando.

Para poder explicar mejor la forma en que se realiza el procedimiento y las conclusiones que se pueden obtener, se presentan a continuación dos ejemplos simples.

En la Tabla 17 se presentan las principales características de las transacciones que serán usadas. Se escogieron dos transacciones lo más distintas posibles para hacer más claro el procedimiento. La primera tiene una tasa de emisión relativamente alta (el promedio es cercano al 4,7%) mientras que la segunda relativamente baja. Por otro lado, el

plazo de la primera es cerca de 2,4 veces el de la segunda y fueron emitidas por entidades diferentes.

Tabla 17: datos para las dos transacciones de los ejemplos.

	Ejemplo 1	**Ejemplo 2**
Fecha	12-04-2007	12-11-2009
Emisor	Banco Estado	Banco de Chile
TIR Transacción (%)	4,2	3,7
Plazo (años)	14,7	6,1
Tasas Emisión (%)	5,8	4
Clasificación	AA+	AAA
Probabilidad de prepago (N=1)	0,04	0,00
Probabilidad de prepago (N=5)	0,29	0,08
Probabilidad de prepago (N=10)	0,29	0,08
Presencia	154	16
Rotación (%)	0,32	4,31
Tasa curva promedio (%)	3,72	4,14
Tasa libre de riesgo (%)	2,69	2,51

La Ilustración 21 describe los caminos que siguen ambas transacciones a lo largo del árbol de regresión. Se puede observar que en ambos caso la tasa de emisión es la primera variable a considerar, mientas que para la tasa de emisión baja el plazo es la segunda y para la tasa de emisión alta es el emisor.

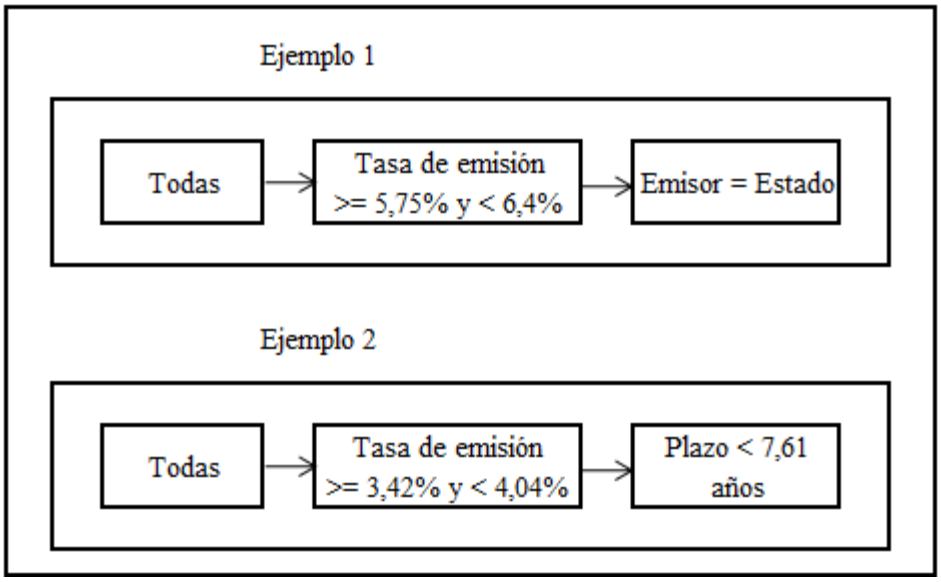

Ilustración 21: camino seguido por las transacciones de ambos ejemplos dentro del árbol de regresión.

El procedimiento descrito anteriormente para analizar cuantitativamente los *spreads* específicos para el primer ejemplo se resume en la Tabla 18. En ésta se muestra el aporte a la TIR que hacen tanto la curva libre de riesgo, la curva promedio y cada nodo del árbol de regresión. También se muestra el aporte porcentual dentro de la TIR y finalmente el aporte porcentual de los tres últimos sobre el spread.

Tabla 18: resultado del ejemplo N°1

	Aporte	**Aporte %**	**Aporte % spread**
$TIR_{Transacción}$	4,200%		
TIR_{rf}	2,690%	64,05%	
$S_{familia}$	1,029%	24,51%	68,18%
$\eta_{Tasa\ emisión}$	0,540%	12,86%	35,77%
$\eta_{Emisor}\|\eta_{tasa\ emisión}$	-0,081%	-1,92%	-5,34%
$S_{sesgo\ día}$	0,007%	0,16%	0,45%
ε	0,014%		

Lo primero que debemos observar es que el mayor aporte porcentual a la TIR lo realiza la curva libre de riesgo, pero esto no nos aporta mucha información ya que ese valor va a depender principalmente del nivel que ésta tenga y no de la transacción en sí. Por esta razón se debe observar el aporte porcentual que hacen las variables al *spread* sobre la curva libre de riesgo.

Para continuar con el análisis debemos observar que la curva promedio, luego de hacer su aporte (1,029%) a la estimación del *spread* deja un porcentaje sin explicar que corresponde al *spread* específico. Éste es mayor que cero ($4{,}2\% - (2{,}69\% + 1{,}029\%) = 0{,}481\%$), lo que nos dice sobre estimó el precio de la LCH ya que al descontar los flujos a una tasa menor el resultado es mayor. A partir de esto, lo que deberíamos esperar del árbol de regresión es que corrija aquel resultado al alza, lo que ocurre efectivamente. El aporte que hace la tasa de emisión (primer nodo para el camino que toma esta transacción) es positivo y mayor que cero (0,540%). Luego, el aporte que hace el hecho de que la LCH haya sido emitida por el Banco del Estado dado que tenía una tasa de emisión relativamente alta fue de -0,081%. Esto nos dice que a pesar de que la transacción fue castigada por tener una tasa de emisión alta (y por ende una mayor facilidad para realizar una operación de prepago) es perdonada por haber sido emitida por el Banco del Estado de Chile (indicándonos que este banco tiene, por ejemplo, menor probabilidad de caer en *default* o que los deudores de ese banco tienen un comportamiento frente a la opción de prepago menos activa). Finalmente, el error cometido por el modelo, luego de corregir por el sesgo del *spread* específico promedio de ese día, fue de 1,4 puntos base.

Por otro lado, es interesante observar cómo se conforma porcentualmente el *spread* de la transacción sobre la curva libre de riesgo. Podemos observar que más del 68% lo

aporta la curva promedio. Esto nos puede estar diciendo que en general, existe un alto spread promedio del mercado hipotecario sobre el mercado de los instrumentos sin riesgos o que las características de la letra no son muy relevantes a la hora de explicar el *spread*.

Finalmente, es importante notar que, a pesar de que el *spread* porcentual aportado por la variable tasa de emisión es más del doble que el de la variable emisor, no podemos concluir que la primera es más relevante que la segunda. Esto ya que la segunda variable está condicionada a la primera.

La Tabla 19 nos muestra el resumen del procedimiento para el segundo ejemplo de transacción.

Tabla 19: resultado del ejemplo N°2

	Aporte	**Aporte %**	**Aporte % spread**
$TIR_{Transacción}$	3,700%		
TIR_{rf}	2,510%	67,84%	
$S_{familia}$	1,634%	44,17%	137,33%
$\eta_{Tasa\ emisión}$	-0,293%	-7,92%	-24,62%
$\eta_{plazo}\|\eta_{tasa\ emisión}$	-0,074%	-2,01%	-6,25%
$S_{sesgo\ día}$	-0,079%	-2,12%	-6,61%
ε	0,002%		

Nuevamente conviene observar el porcentaje sin explicar por la curva promedio, el cual es de -0,444% $(3{,}7\% - (2{,}51\% + 1{,}634))$. Dado que éste fue negativo la curva promedio subestimó el precio final de la LCH que se estaba transando. Luego, era de esperar que el árbol corrigiera la tasa estimada a la baja, siendo esto efectivamente lo que ocurrió. Luego de los aportes que hacen la tasa de emisión y el plazo a la TIR (-0,293% y -

0,074% respectivamente) y de la corrección por el sesgo *spread* específico de ese día (0,079%) el error cometido por el modelo fue de 2 puntos base.

Por otro lado, resulta interesante ver que la segunda variable que se toma en cuenta es el plazo. Esto se puede explicar debido a que dada esa tasa, el comportamiento de la LCH debiera ser más parecido al de un bono sin opción de prepago, en cual caso la variable más importante es efectivamente el plazo. Por otro lado, también es interesante ver cómo esta variable influye poco en la determinación del *spread*, lo que nos sugiere, que la variable plazo es captada en su mayoría por el modelo dinámico de la curva promedio.

6.6.3 Estudio del comportamiento promedio del mercado

El objetivo de este análisis es observar cómo el árbol de regresión realizó, en promedio, las correcciones sobre la curva promedio (*spread* específico), para tener así una intuición de la forma en que el mercado responde frente a este mercado.

Para realizar en estudio, tomaremos el árbol de regresión obtenido en el experimento correspondiente a calibrar con el año 2008, el cual se ilustra (hasta 3 niveles de expansión) en la Ilustración 8. En este podemos observar que la primera variable que se toma en cuenta para encontrar los *spreads* específicos es la tasa de emisión de la letra hipotecaria. Luego, hay dos tipos de decisiones, las que filtran por emisor o por clasificación de riesgo.

La Tabla 20 presenta el resumen de las estimaciones que hace el árbol para el primer nivel de extensión. En ésta se puede observar que el castigo promedio que hace el árbol va disminuyendo a medida que disminuye la tasa a la que fue emitida la letra hipotecaria (hasta que se convierte en un premio para aquellas con tasas de emisión menor a 4,7%). Es

importante recordar que este castigo o premio, según corresponda, es medido relativo a la familia papeles (ya que estamos midiendo el *spread* específico sobre la curva promedio).

Tabla 20: resumen estudio promedio del árbol de regresión para el año 2008 extendido un nivel.

Condición	**Castigo promedio**	**Desviación estándar**
Tasa emisión > 5.8	0,641%	0,003491
Tasa emisión >= 5.25 and Tasa emisión < 5.8	0,209%	0,002412
Tasa emisión >= 4.7 and Tasa emisión < 5.25	0,127%	0,001386
Tasa emisión >= 4.15 and Tasa emisión < 4.7	-0,067%	0,001285
remisión < 4.15	-0,151%	0,000750

Este mayor castigo (o premio) debido a una mayor (o menor) tasa de emisión es un resultado razonable debido a que si un crédito hipotecario fue emitido a una alta tasa, las condiciones de mercado que se tienen que dar para re financiar ese crédito (entre ellas una menor tasa de interés) son menos exigentes, por lo tanto la probabilidad de que se realice el refinanciamiento aumenta y por lo tanto también aumenta el *spread* asociado a esa característica. Por otro lado, si un crédito fue emitido a una baja tasa, las condiciones que se tienen que dar (entre ellas una tasa aún menor) son más difíciles de conseguir, por ende el riesgo de que se concrete un refinanciamiento es menor (y así el *spread* asociado).

También resulta interesante observar cómo, al igual que el castigo promedio, la desviación estándar de ese castigo baja. Este resultado también es razonable debido a que, como se discutió en el párrafo anterior al emitir letras a una menor tasa de emisión, la probabilidad de que sean refinanciadas es menor, por lo que su comportamiento debiera parecerse más al de una deuda sin esa opción, la cual debería tender a desviarse menos de la media.

Para continuar el análisis, la Tabla 21 nos muestra el mismo resultado anterior pero para dos niveles de extensión. En esta podemos encontrar dos resultados importantes. El primero se ve en el grupo de las letras con tasas de emisión entre 5,25% y 5,8%, donde el castigo al ingresar la segunda condición, respecto al castigo tomando en cuenta sólo la primera, disminuye considerablemente al ser el Banco del Estado el emisor y aumenta al no serlo (con la primera condición el castigo era de 0,209%). Esto nos dice que las letras emitidas por el Banco del Estado tienen algún comportamiento que hace que baje el castigo como por ejemplo una menor frecuencia de prepago.

Tabla 21: Resumen estudio promedio del árbol de regresión para el año 2008 extendido un nivel

Condición 1	Condición 2	Promedio	Desviación estándar
Tasa emisión > 5.8	**Tasa emisión > 5.8**	0,641%	0,003491
Tasa emisión >= 5.25 and Tasa emisión < 5.8	**Tasa emisión >= 5.25 and Tasa emisión < 5.8 and Emisor = ESTADO**	0,101%	0,001816
	Tasa emisión >= 5.25 and Tasa emisión < 5.8 and Emisor <> ESTADO	0,299%	0,002482
Tasa emisión >= 4.7 and Tasa emisión < 5.25	**Tasa emisión >= 5.25 and Tasa emisión < 5.8 and Emisor = ESTADO**	0,128%	0,001658
	Tasa emisión >= 5.25 and Tasa emisión < 5.8 and Emisor <> ESTADO	0,126%	0,001209
Tasa emisión >= 4.15 and Tasa emisión < 4.7	**Tasa emisión >= 4.15 and Tasa emisión < 4.7 and Clasificación = AA+**	-0,208%	0,000703
	Tasa emisión >= 4.15 and Tasa emisión < 4.7 and Clasificación <> AA+	-0,037%	0,001179
Tasa emisión < 4.15	**Tasa emisión < 4.15 and Clasificación = A+**	-0,285%	0,000631
	Tasa emisión < 4.15 and Clasificación <> A+	-0,139%	0,000632

El segundo resultado importante es el que se observa en el grupo de las letras con tasas de emisión entre 4,15% y 4,7%, donde el premio aumenta considerablemente si la clasificación de riesgo es alta (la clasificación AA+ es la segunda mayor). Este resultado es

muy razonable ya que al tener mayor clasificación el *spread* asociado al *default* de uno de estos papeles debería bajar. Un resultado análogo a éste se encuentra en las letras que tienen tasa de emisión menor a 4,15%, donde el premio es menor si la clasificación de riesgo es baja (la clasificación A+ es la segunda menor dentro de la muestra).

Un último punto interesante es el que se observa en el grupo de las letras con tasa de emisión entre 4,7% y 5,25% donde el castigo cambia en sentido opuesto al primer resultado expuesto (aumenta al ser emitido por el Banco del Estado). Este resultado, a pesar de parecer contradictorio, no lo es ya que la diferencia entre ambos es despreciable (igual a 0,2 puntos base). Esto nos dice que estas dos características (tasa de emisión y emisor = Banco del Estado) son suficientes y debe realizarse un estudio más profundo tal como se puede observar en la Ilustración 8 (donde se muestra gráficamente el árbol obtenido para el año 2008).

7 CONCLUSIONES

Esta tesis propone una metodología tanto para la estimación de precios de transacciones de papeles con riesgo y opciones, como para la extracción de información cualitativa y cuantitativa de su comportamiento a partir de datos de transacciones. Esta metodología busca resolver el problema de la dificultad de modelar los distintos factores que determinan los *spreads* sobre las curvas libres de riesgo, como lo son las opciones de prepago o los problemas de liquidez o default. Para lograr esto se plantea una combinación entre modelación dinámica de estructuras de tasas de interés mediante métodos de no arbitraje y modelación estática del comportamiento particular de las transacciones usando algoritmos de minería de datos. Particularmente, se propone modelar el comportamiento promedio con la metodología dinámica y las desviaciones de ésta (spreads específicos) mediante el modelo estático de minería de datos. Se propone también una metodología de estimación en 2 etapas donde se busca aprovechar las bondades de cada uno de los modelos.

Para probar el modelo éste se aplica a una muestra de transacciones de letras de crédito hipotecarias, las cuales presentan tanto opciones de prepago como los problemas de liquidez y *default* de la mayoría de los papeles. Los resultados muestran que se puede modelar con éxito el comportamiento promedio de las transacciones logrando estructuras de volatilidad consistentes entre datos y modelo. También muestran que el algoritmo de minería de datos árboles de regresión logra de buena manera hacer estimaciones de *spreads* específicos para tener estimaciones de tasas internas de retorno con bajo error y

dar una intuición explícita del comportamiento del mercado de letras de crédito hipotecarias.

La aplicación del modelo al mercado chileno de letras muestra que éste tiene una inclinación a determinar las desviaciones (o *spreads* específicos) de las transacciones promedio usando en un primer lugar la tasa a la cual los instrumentos son emitidos. Este resultado es razonable dado que la tasa de emisión es un determinante muy importante en la ejecución de un prepago y por ende del castigo que esta merece por la misma razón. Por otro lado, se muestra también una tendencia a diferenciar por el emisor de la letra y por el plazo al vencimiento en algunas ocasiones.

El resultado anterior se complementa con un análisis del comportamiento promedio del mercado en base al resultado del árbol. Este nos muestra cómo van cambiando los castigos y premios estimados a lo largo de sus ramas.

El modelo puede en consecuencia ser un aporte puesto que permite obtener estimaciones estables del comportamiento promedio de un mercado particular, estimaciones de precios de instrumentos y porque ofrece una intuición, tanto cualitativa como cuantitativa, de la forma en que se transan los instrumentos y la influencia que tienen las distintas combinaciones de variables en los *spreads* por sobre las curvas libres de riesgo.

Investigaciones posteriores podrían considerar la aplicación de una metodología de estimación conjunta para el modelo dinámico y estático, de forma de intentar eliminar los errores sistemáticos y evitar la re estimación de los parámetros del primer modelo. Otra forma en que este problema podría solucionarse sería haciendo que el árbol de regresión

pudiera incluir información nueva en cada período, integrando y eliminando así los errores sistemáticos.

Otro aporte importante sería el de la inclusión de los papeles de corta duración, puesto que todos los papeles van a llegar inevitablemente a ese punto. Para lograr esto sería necesario en una primera instancia que la curva promedio reflejara la volatilidad de esos plazos de forma que sea factible trabajar posteriormente con un modelo de estimación. En segunda instancia, sería importante también encontrar las potenciales variables que causan el mayor impacto en ese grupo de forma que el árbol de regresión pueda usarlas para obtener la información del mercado y estimar los *spreads* específicos.

Por otro lado, la aplicación de este modelo a otros mercados distintos al de las letras de crédito hipotecarias sería sin duda un importante aporte. No sólo por la necesidad de constante de encontrar mejores poderes de ajuste sino que por el potencial de información adicional que se podría extraer de la manera en que los distintos mercados financieros realizan sus transacciones.

Finalmente, un posible aporte sería estudiar la forma de usar la información rescatada del árbol de regresión, relativa a la importancia de las características del instrumento, como forma de identificar las variables relevantes en el planteamiento de modelos dinámicos de valorización.

BIBLIOGRAFIA

Adams, N. M., Hand, D. J., & Till, R. J. (2001). Mining for Classes and Patterns in Behavioural Data. *The Journal of the Operational Research Society, Vol. 52, N°9, Special Issue: Credit Scoring and Data Mining* , 1017-1024.

Bedendo, M., Cathcart, L., & El-Jahel, L. (2007). The Slope of the Term Structure of Credit Spreads: An Empirical Investigation. *Journal of Financial Research, 30(2).* , 237–257.

Black, F., & Scholes, M. (1973). The Pricing of Options and Corporate Liabilities. *Journal of Political Economy, Vol. 81* , 637-654.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees.* Wadsworth, Monterey : Chapman and Hall/CRC; 1 edition .

Brennan, M. J., & Schwartz, E. S. (1985). Evaluating Natural Resource Investments. *Journal of Business, Vol. 58* , 133-155.

Brennan, M., & Schwartz, E. (1979). A continuous time approach to the pricing of bonds . *Journal of Banking & Finance, vol 2 (3)* , 133-155.

Chen, L., Lesmond, D., & Wei, J. (2007). Corporate Yield Spreads and Bond Liquidity. *The Journal of Finance, 62(1)* , 119–149.

Collin-Dufresne, P., Goldstein, R. S., & Martin, J. S. (2001). The Determinants of Credit Spread Changes. *The Journal of Finance, Vol. 56, N°6* , 2177-2207.

Cortazar, G., & Naranjo, L. F. (2006). An N-Factor Gaussian Model of Oil Futures. *Journal of Futures Markets, Vol. 26, N°3* , 243-268.

Cortazar, G., Schwartz, E., & Naranjo, L. (2007). Term-Structure Estimation in Markets with Infrequent Trading. *International Journal of Finance & Economics, Vol. 12, N°4* , 353–369.

Cortazar, G., Schwartz, E., & Tapia, C. (2010). Credit Spreads in Illiquid Markets, Documento de Trabajo Departamento de Ingeniería Industrial y de Sistemas, Escuela de Ingeniería, Pontificia Universidad Católica de Chile.

Dai, Q., & Singleton, K. (2002). Expectation puzzles, Time-varying Risk Premia, and Affine Models of the Term Structure. *Journal of Financial Economics, Vol. 63, N°3* , 415–441.

Deconinck, E., Hancock, T., Coomans, D., Massart, D., & Heyden, Y. V. (2005). Classification of drugs in absorption classes using the classification and regression trees

(CART) methodology. *Journal of Pharmaceutical and Biomedical Analysis, Vol. 39* , 91–103.

Dungey, M., Martin, V. L., & Pagan, A. R. (2000). A Multivariate Latent Factor Decomposition of International Bond Yield Spreads. *Journal of Applied Econometrics, Vol. 15, No. 6, Special Issue: Inference and Decision* , 697-715.

Elton, E. J., Gruber, M. J., Agrawal, D., & Mann, C. (2001). Explaining the Rate Spread on Corporate Bonds. *The Journal of Finance, Vol. 56, N°1* , 247-277.

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques, second edition.* San Francisco: Morgan Kaufmann Publishers.

Harvey, A. (1990). Forecasting, Structural Time Series Models and the Kalman Filter. *Cambridge University Press* .

Kalman, R. (1960). A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering, Vol. 82, N°1* , 35–45.

Kau, J. (2005). The effect of mortgage price and default risk on mortgage spreads . *JOURNAL OF REAL ESTATE FINANCE AND ECONOMICS 30 (3)* , 285-295.

Langetieg, T. (1980). A multivariate model of the term structure. *Journal of Finance, Vol. 35, N° 1* , 71-97.

Liu, J., Longstaff, F., & Mandell, R. (2006). The Market Price of Risk in Interest Rate Swaps: The Roles of Default and Liquidity Risks. *The Journal of Business, Vol. 79, N°5* , 2337–2359.

Longstaff, F., & Schwartz, E. (1992). Interest Rate Volatility and the Term Structure: A Two-Factor General Equilibrium Model. . *Journal of Finance, Vol 47, N°4* , 1259–1282.

Longstaff, F., Mithal, S., & Neis, E. (2005). Corporate Yield Spreads: Default Risk or Liquidity? New Evidence from the Credit Default Swap Market . *The Journal of Finance, Vol. 60, No. 5* , 2213-2253.

Merton, R. (1974). On the Pricing of Corporate Debt: The Risk Structure of Interest Rates. *Journal of Finance, Vol 29, N°2* , 449–470.

Nelson, C., & Siegel, A. (1987). Parsimonious modeling of yield curves. *Journal of Business, Vol. 60, N° 4* , 473-489.

O'Connor, N., & Madden, M. G. (2006). A neural network approach to predicting stock exchange movements using external factors. En *Applications and Innovations in Intelligent Systems XIII* (págs. 64-77). Springer London.

Øksendal, B. (2003). *Stochastic Differential Equations: An Introduction With Applications.* Springer.

Rodriguez, R. (1988). Default Risk, Yield Spreads, and Time to Maturity . *The Journal of Financial and Quantitative Analysis, Vol. 23, No. 1* , 111-117 .

Sørensen, C. (2002). Modeling Seasonality in Agricultural Commodity Futures. *Journal of Futures Markets, Vol. 22, N°5* , 393–426.

Svensson, L. (1994). Estimating and interpreting forward interest rates. Working Paper. *National Bureau of Economic Research* .

Vasicek, O. A. (1977). An equilibrium characterization of the term structure. *Journal of Financial Economics, Vol. 5, N° 2* , 177-188.

Wang, J. (2008). Liquidity, default, taxes, and yields on municipal bonds . *JOURNAL OF BANKING & FINANCE 32 (6)* , 1133-1149 .

A N E X O S

ANEXO A: DERIVACIÓN DEL VALOR DE UN BONO DE DESCUENTO EN EL MODELO VASICEK MULTIFACTORIAL PARA LA TASA LIBRE DE RIESGO

En un modelo dinámico de tasas de interés, normalmente el valor de la tasa se representa como una combinación lineal de las variables de estado:

$$r = \delta^r{}' x^r + \delta_0^r \tag{A-1}$$

La dinámica luego de ajustar por riesgo, para cada uno de los factores x queda determinada por la expresión:

$$dx = -(\lambda + Kx_i)dt + \Sigma dw \tag{A-2}$$

donde el término dw representa un vector de movimientos Brownianos correlacionados tal que:

$$(dw)'(dw) = \Omega dt \tag{A-3}$$

Cada elemento de la matriz Ω, ρ_{ij} representa la correlación que existe entre el factor i y j. El valor del instrumento P cuya tasa de descuento es r puede obtenerse aplicando el lema de Îto:

$$dP(x,T) = \sum_{i=1}^{N} P_{x_i} dx + P_T dT + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} P_{x_i x_j} dw_i dw_j \tag{A-4}$$

Reemplazando los términos dx por la ecuación (A-2), usando la relación $dT = -dt$, y desechando los términos dt con exponente mayor a uno, la ecuación (A-4) queda de la siguiente forma:

$$dP(x,T) = \left(-\sum_{i=1}^{N} P_{x_i}(\lambda i + \kappa_i x_i) - P_T + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} P_{x_i x_j}\sigma_i\sigma_j\rho_{ij} \right)\partial t + \left(\sum_{i=1}^{N} P_x \sigma_i \right)\partial w \tag{A-5}$$

Por valorización neutral al riesgo, la tendencia de la ecuación anterior debe ser igual a la tasa libre de riesgo r (por arbitraje), con lo cual se obtiene la ecuación diferencial para el valor del instrumento, dada por la siguiente ecuación:

$$-\sum_{i=1}^{N} P_{x_i}(\lambda i + \kappa_i x_i) - P_T + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} P_{x_i x_j}\sigma_i\sigma_j\rho_{ij} = \delta_0 + \sum_{i=1}^{N}\delta_i x_i \tag{A-6}$$

Esta ecuación tiene soluciones analíticas de la forma descrita en (A-7).

$$P(x,T) = e^{(u(T)'x + v(T))} \tag{A-7}$$

Por lo tanto los valores de las derivadas en la ecuación (A-6) en función de los términos $u(T)$ y $v(T)$ anteriores son:

$$P_{x_i} = u_i(T)P$$

$$P_{x_i x_j} = u_i(T)u_j(T)P$$

$$P_{x_i} = \left(\sum_{i=1}^{N} u_{i'}(T) + v'(T) \right)P \tag{A-8}$$

Finalmente, reemplazando estas ecuaciones en (A-6) y agrupando términos similares, se obtienen las expresiones para los términos $u(T)$ y $v(T)$. Esto permite definir una solución cerrada para el valor de un instrumento bajo esta modelación.

$$u_i(\tau) = -\delta_i\left(\frac{1-e^{-k_i\tau}}{k_i}\right)$$

$$v(\tau) = \sum_{i=1}^{n}\delta_i\frac{\lambda_i}{k_i}\left(\tau - \frac{1-e^{-k_i\tau}}{k_i}\right) - \delta_0^r\tau$$

$$+\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\delta_i)^2\frac{\sigma_i\sigma_j\rho_{ij}}{k_ik_j}\left(\tau - \frac{1-e^{-k_i\tau}}{k_i} - \frac{1-e^{-k_j\tau}}{k_j} + \frac{1-e^{-(k_i+k_j)\tau}}{k_i+k_j}\right)$$

(A-9)

ANEXO B: MATRICES DE LA ECUACION DE MEDIDA DEL FILTRO DE KALMAN CON UNA RELACION LINEAL Y UNA LINEALIZADA

En su forma original, el filtro planteado por (Kalman, 1960) supone una relación lineal entre las variables observables y no observables (latentes) del modelo. Esto se traduce en una ecuación de medida dada la siguiente expersión:

$$z_t = H_t x_t + d_t + v_t \qquad v_t : N(0, R_t) \tag{B-1}$$

El valor de un bono de descuento que paga 1 en el período de vencimiento T está representado por:

$$P = e^{-rT} \tag{B-2}$$

Por otro lado, su representación en el espacio de estados, asumiendo que estas variables siguen un proceso de reversión a la media del tipo Vasicek, está dada por:

$$P_s = e^{(u(T)'X + v(T))} \tag{B-3}$$

Igualando (B-2) y (B-3) , se encuentra la relación lineal que existe entre las variables observables r y las variables de estado X :

$$r = -\frac{1}{T}(u(T)'X + v(T)) \tag{B-4}$$

De esta forma, las matrices que usa el filtro para la ecuación de medida son:

$$z_t = \begin{bmatrix} r_t^1 \\ \vdots \\ r_t^{M_t} \end{bmatrix} \quad H_t = \begin{bmatrix} -u_1^1(T)/T & \cdots & -u_N^1(T)/T \\ \vdots & \ddots & \vdots \\ -u_1^{M_t}(T)/T & \cdots & -u_N^{M_t}(T)/T \end{bmatrix} \quad d_t = \begin{bmatrix} -v^1(T)/T \\ \vdots \\ -v^M(T)/T \end{bmatrix}$$

(B-5)

donde N es el número total de factores para modelar la tasa r y M_t es el número de observaciones en el período t. La ausencia de cupones (como en el caso de un bono de descuento), permite despejar una expresión cerrada para la tasa r en función de las variables de estado. En esta se observa la relación lineal entre éstas y la variable observable. Para el caso de un bono con cupones, su valor y su representación en el espacio de estados están dados por las ecuaciones (B-6) y (B-7).

$$P = \sum_{i=1}^{T} C_i e^{-rt_i}$$

(B-6)

$$P_s = \sum_{i=1}^{T} C_i e^{-(u(t_i)'X + v(t_i))}$$

(B-7)

En este caso, a diferencia del caso anterior, no es posible obtener una relación lineal entre la tasa r y las variables de estado X. Luego, para resolver este problema, (Harvey, 1990) propone una extensión del filtro de Kalman que permite usarlo frente a ecuaciones de medida no lineal, linealizando la función que relaciona ambas variables a través de una extensión de Taylor de primer orden de la siguiente manera:

$$f(x_t) = f(\hat{x}_{t|t-\Delta t}) + \left.\frac{\partial f(x_t)}{\partial x_{t'}}\right|_{x_t = \hat{x}_{t|t-\Delta t}} (x_t - \hat{x}_{t|t-\Delta t})$$

(B-8)

donde $f(x)$ representa la función que relaciona el espacio de estados con la variable observada. La linealización se realiza en torno a la predicción que hace el filtro usando la información disponible hasta $t-\Delta t$, de forma de asegurar que la aproximación de primer orden sea suficientemente aceptable.

Igualando las expresiones (B-6) y (B-7) y derivando implícitamente se obtiene:

$$\frac{\partial P}{\partial r}\frac{\partial r}{\partial x} = \sum_{i=1}^{T} C_i u(t_i) e^{(-u(t_i)'X+v(t_i))}$$

$$\frac{\partial P}{\partial r} = \sum_{i=1}^{T} C_i t_i e^{-rt_i}$$

(B-9)

Aplicando esta linealización, se obtiene la nueva ecuación de medida linealizada.

$$z_t = \hat{H}_t x_t + \hat{d}_t + v_t \qquad v_t : \; N(0, R_t)$$

(B-10)

donde

$$\overline{H}_t = \begin{bmatrix} \overline{H}_t^1 \\ \vdots \\ \overline{H}_t^{M_t} \end{bmatrix} \qquad \overline{d}_t = \begin{bmatrix} \overline{d}_t^1 \\ \vdots \\ \overline{d}_t^{M_t} \end{bmatrix}$$

$$\overline{H}_t^m = \frac{1}{\sum_{i=1}^{T_m} C_i^m t_i e^{-r_0 t_i}} \begin{bmatrix} \sum_{i=1}^{T_m} C_i^m u_i^{1,m} e^{-r(u'X+v)} & \cdots & \sum_{i=1}^{T_m} C_i^m u_i^{N,m} e^{-r(u'X+v)} \end{bmatrix}$$

$$\overline{d}_t^m = r_0 - \overline{H}_t^m \hat{x}_{t|t-\Delta t} \qquad r_0 = r(\hat{x}_{t|t-\Delta t})$$

(B-11)

En la ecuación (B-11), los términos $\overline{H}_t^m$ y $\overline{d}_t^m$ representan el término *m-ésimo* de las matrices $\overline{H}_t$ y $\overline{d}_t$, r_0 correspondientes a la tasa obtenida de evaluar la expresión (B-7)

en las variables de estado estimadas por el filtro y luego igualarla a la expresión (B-6), despejando la tasa r por algún método numérico.

ANEXO C: DERIVACIÓN DE LA ESTRUCTURA DE VOLATILIDAD PARA UN BONO MODELADO CON UNA DINÁMICA DE TIPO VASICEK

A continuación se presenta el desarrollo para la obtención de una fórmula general y cerrada para la estructura de volatilidad de algún instrumento modelado bajo una dinámica Vasicek y con reversión a la media. Si se asume un proceso para el precio del bono $P(x,T)$ con una dinámica

$$\frac{dP}{P} = \mu_P dt + \sigma_P dw \tag{C-1}$$

y donde x representa el vector de variables de estado que modelan el precio del instrumento y T es su madurez. Luego, aplicando el lema de Îto, se obtiene:

$$\frac{dP}{P} = \frac{1}{P}\sum_{i=1}^{N} P_{x_i} dx_i + \frac{1}{2}\frac{1}{P}\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\partial^2 P}{\partial x_i \partial x_j} dx_i dx_j - \frac{1}{P} P_T dt \tag{C-2}$$

Por otro lado, de la ecuación (C-2) es claro que al despreciar términos dt con exponente mayor a 1, se obtiene:

$$\left(\frac{dP}{P}\right)\left(\frac{dP}{P}\right) = \sigma_P dt \tag{C-3}$$

Luego considerando que la correlación entre dos variables de estado x_i y x_j es ρ_{ij}, y combinando (C-2) en (C-3) se obtiene la expresión para la volatilidad de un instrumento modelado usando x_i. Esta se presenta en la ecuación (C-4) .

$$\sigma_P^2 = \sum_{i=1}^{N}\sum_{j=1}^{N}\sigma_i\sigma_j\rho_{ij}\left(\frac{1}{P}P_{x_i}\right)\left(\frac{1}{P}P_{x_j}\right) \tag{C-4}$$

Dada la dinámica de Vasicek impuesta para las variables de estado, se demostró que la representación en el espacio de estados para un instrumento de descuento es:

$$P = e^{\sum_{i=1}^{N} u_i x_i + v} \tag{C-5}$$

Luego, reemplazando (C-5) en (C-4) se obtiene la expresión final para la estructura de volatilidad instantánea teórica del modelo presentada en la ecuación (C-6).

$$\sigma_P^2 = \sum_{i=1}^{N}\sum_{j=1}^{N}\sigma_i\sigma_j\rho_{ij}u_i u_j \qquad con \qquad u_i = \frac{e^{k_i}-1}{k_i} \tag{C-6}$$

ANEXO D: DURACION DE UNA LETRA HIPOTECARIA Y COMPARACIÓN SEGÚN PLAZO

La duración de un instrumento con pagos en el futuro se representa según la siguiente ecuación:

$$Duración = \frac{\sum_{i=1}^{Cupones} C_i e^{(-TIR*\tau_i)} * \tau_i}{\sum_{i=1}^{Cupones} C_i e^{(-TIR*\tau_i)}}$$

(D-1)

Luego, como en el caso de las letras se tiene que todos los cupones son iguales para un mismo instrumento, la duración queda especificada de la siguiente manera:

$$Duración = \frac{\sum_{i=1}^{Cupones} e^{(-TIR*\tau_i)} * \tau_i}{\sum_{i=1}^{Cupones} e^{(-TIR*\tau_i)}}$$

(D-2)

De esta forma, podemos afirmar que dos letras hipotecarias con el mismo plazo al vencimiento, y por ende igual cantidad de cupones (pero diferentes en magnitud) tienen la misma duración.

ANEXO E: TABLA DE DESARROLLO DE UNA LETRA DE CREDITO HIPOTECARIO

Para el cálculo de la tabla de desarrollo de una letra de crédito hipotecario se utiliza una tasa de emisión equivalente que depende de la periodicidad de los pagos de la letra. Esta se calcula de la siguiente manera:

$$r_{EQ} = (1 + r_{anual})^{Periodicidad/12} - 1$$

(E-1)

En la ecuación anterior r_{anual} corresponde a la tasa de emisión, r_{EQ} es la tasa equivalente de emisión para cada *Periodicidad* (mensual, trimestral, semestral o anual). Esta tasa es utilizada en el cálculo de los cupones C_i, según el monto del corte M y el número de cupones N, utilizando la fórmula de la anualidad:

$$C_i = M\left(\frac{1}{r_{EQ}} - \frac{1}{r_{EQ}(1 + r_{EQ})^N}\right)^{-1} \qquad i = 1, \dots, N-1$$

(E-2)

Los intereses se pagan sobre el capital insoluto anterior, sin considerar los días efectivos.

$$I_i = K_{i-1}\left(1 + r_{EQ}\right) - K_i$$

(E-3)

La amortización se calcula como la diferencia entre el cupón y los intereses pagados:

$$A_i = C_i - I_i$$

(E-4)

Y el nuevo saldo es el capital insoluto anterior menos la parte amortizada para la actual fecha de pago de cupón.

$$K_i = K_{i-1} - A_i$$

(E-5)

El último cupón de la LCH debe realizar el ajuste correspondiente al redondeo de decimales realizado en los cupones anteriores.