PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE

ESCUELA DE INGENIERIA

MERGING DYNAMIC FINANCIAL MODELS WITH DATA MINING TECHNIQUES FOR PRICING FIXED INCOME SECURITIES IN LOW-LIQUIDITY MARKETS

JOSÉ IGNACIO VILLARROEL MOYA

Thesis subimtted to the Office of Research and Graduate Studies in partial fulfillment of the requirements for the Degree of Master of Science in Engineering

Advisor:

GONZALO CORTÁZAR SANZ

Santiago de Chile, (August, 2013)

PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE

ESCUELA DE INGENIERIA

MERGING DYNAMIC FINANCIAL MODELS WITH DATA MINING TECHNIQUES FOR PRICING FIXED INCOME SECURITIES IN LOW-LIQUIDITY MARKETS

JOSE I. VILLARROEL

Members of the Committee:

GONZALO CORTAZAR SAINZ

SERGIO MATURANA

HECTOR ORTEGA

LORENZO NARANJO

LUIS FERNANDO ALARCON

Thesis submitted to the Office of Research and Graduate Studies in partial fulfillment of the requirements for the Degree of Master of Science in Engineering

Santiago de Chile, (August, 2013)

To my family for giving me all the opportunities I have ever wanted and all the support I needed for fulfilling my dreams.

ACKNOWLEDGEMENTS

I must first recognize the support of my Supervisor, Gonzalo Cortazar. I thank him for his time, dedication and advices during this investigation. His support was key in the success of this work.

I must also thank Hector Ortega, for helping me with the development of this research and for giving me continuous and valuable feedback in order to complete it.

I would like also to thank everyone in FinlabUC for their valuable help with everything I needed whenever I needed it.

Also, I would want to thank the financial support provided by CONICYT through the Fondecyt project 1130352.

Finally, I want to make a special recognition to my family and friends. I especially thank my parents and my brothers for giving me all the support, especially when the road seemed very difficult. Without them, this thesis would not have been possible.

TABLE OF CONTENTS

ABSTRACT

A two-stage model using both dynamics financial models and data mining techniques for valuing fixed income securities is proposed. Dynamic models have been shown to be particularly useful in low-liquidity emerging markets when estimating an average term structure for an average fixed income security. However, due to the complexity of some fixed income securities, it is difficult to incorporate all variables into these mathematical models. Data mining techniques are used in order to increase the accuracy of predictions and to obtain a better understanding of market valuation patterns. We implement this model in the Chilean mortgage notes market. Results show that average estimation error is reduced by 120 basis points average by incorporating data mining techniques in average with very low computational costs.

RESUMEN

En esta tesis se propone una metodología de dos fases que utiliza tanto modelos dinámicos como técnicas de minería de datos para valorizar activos de renta fija. Los modelos dinámicos han mostrado ser particularmente útiles en mercados emergentes con baja liquidez para estimar una estructura de tasas de interés promedio para un activo típico. Sin embargo, debido a la complejidad de algunos activos de renta fija, es muy difícil incorporar todas las variables en estos modelos matemáticos. Así, se implementan técnicas de minería de datos para aumentar la precisión de las estimaciones y para obtener un mejor entendimiento de los patrones de valorización que utiliza el mercado. Implementamos este modelo en el mercado de letras hipotecarias chileno. Los resultados muestran que el error de estimación se reuce en 120 puntos base, en promedio, cuando se incorporan técnicas de minería de datos con bajos costos computacionales.

1 ARTICLE BACKGROUND

1.1 Introduction

Through the years financial markets have developed, providing liquidity to agents that need it, helping assign resources in a more efficient way and giving access to financing that otherwise would sometimes not be possible.

With this development, accurately pricing of all securities traded in financial markets have become crucial. Every asset has diverse variables that determine its price. Mathematical and financial models have been developed trying to estimate the factors and valuating the securities.

This work focuses on fixed income securities from low-volume trade markets. This presents two main issues to be assessed. First, traditional pricing models have many issues when used in low-volume traded markets as there is little liquidity and information, thus not being able to get an accurate term structure. Second, this kind of securities has many particular characteristics, often difficult to incorporate in traditional pure mathematical models.

In particular, this thesis presents a model that covers both issues presented above for valuating fixed income securities in two stages:

i. A dynamic mathematical model for estimating the average term structure.

ii. A data mining computational model in order to determine the remaining spread according to particular characteristics of each instrument.

Using data mining techniques gives a better understanding of market valuation patterns in order to determine which factors are taken into account in securities prices.

The rest of this work is structured as follows. Section 1.2 presents the main objectives; section 1.3 presents a short literature and conceptual review that serves as a

theoretical framework and section 1.4 states further research. Following this, section 2 contains the main article of this thesis. Within this, Section 2.1 introduces the problem, Section 2.2 describes the model; Section 2.3 looks at the data; Section 2.4 examines the empirical results. Finally, Section 2.5 concludes.

1.2 Main Objective

This thesis has the goal of proposing a two-stage methodology for finding a more accurate pricing model for fixed income securities in low traded volume markets. This will reduce the estimation error and will give a better understanding of the factors and patterns that market incorporates when it assigns a price to a particular instrument.

The main idea is to complement two different approaches. On one side, the financial one that uses dynamic models for estimating the average term structure. This kind of model gives a good estimation, particularly in low-liquidity markets. However, it is difficult to find a model that explains accurately the remaining spread for each transaction. The second approach is a computational one, used to estimate the remaining spread. Data mining techniques are implemented with multiple data about every security. This has the benefit to discover patterns in transactions and improve the final price estimation. The kind of algorithm gives valuable information about how the market decides when pricing securities.

Finally, a second objective is to implement this methodology in the Chilean mortgage-backed notes. This market is particularly interesting due to the nature of the instruments. These are callable notes issued by multiple institutions with diverse coupon rates. This implementation allows to test whether it is possible to incorporate all the information of a particular transaction in a cost-efficient and easy to understand way. It is expected that data mining algorithms will improve the model's fit and estimation will be robust through time.

1.3 Literature Review

The key for valuating fixed-income securities is the interest rate term structure which determines the yield for each maturity. With that curve, every payment can be discounted and thus price the security.

This term structure can be estimated through different models. Two kinds of models are the most used ones. First, static models that estimate the interest rate curve with contemporaneous data only and multiple factors. The second types of models are dynamic ones; the main difference is that these models take into account also historical data and have a more stable solution.

1.3.1 Term structure: static models

Static models were first developed by Nelson & Siegel (1987) and Svensson (1994). These models assume a parsimonious structure for the interest rate and it depends of a limited number of parameters, which are estimated each period with all the transactions available. These models get a successful result in deep and liquid markets, because information can be found for every duration. This is not the case for the Chilean market where liquidity is very low and most of the time, it is not possible to find transactions for every duration.

Nelson & Siegel (1987) defines the forward rate for a given time T as:

$$f(T) = \beta_0 + \beta_1 e^{\frac{T}{\tau}} + \beta_2 \frac{T}{\tau} e^{\frac{T}{\tau}} \qquad (1.1)$$

where $f(T)$ is the forward rate, T the time to maturity and β_0, β_1, β_2 and τ are parameters that must be estimated every day.

In this equation β_0, β_1 and β_2 represents the short, medium and long term. This specification can thus adapt to different structures.

Svensson (1994) adds another parameter to the equation, resulting in the following equation for the forward rate:

$$f(T) = \beta_0 + \beta_1 e^{\frac{T}{\tau_1}} + \beta_2 \frac{T}{\tau_1} e^{\frac{T}{\tau_1}} + \beta_3 \frac{T}{\tau_2} e^{\frac{T}{\tau_2}} \qquad (1.2)$$

This structure is more flexible, it has more freedom degrees but it can possibly add unnecessary volatility to the model.

It is shown in Cortazar et al. (2007) that, when cross-section data is missing, these kind of models behaves poorly. This led to the use of dynamic models.

1.3.2 Term structure: Dynamic models

Dynamic models are capable of incorporating historical data. Some examples of these models are Vasicek (1977) and Brennan & Schwartz (1979) whom specify the interest rate term structure with one and two stochastic factors respectively. Langetieg (1990) extends this to a multifactorial model.

This work is based on Cortázar et al. (2007) which proposes a multifactorial generalized Vasicek model for the interest rate with N stochastic factors that have a mean-reversion process as showed in

$$r = 1'x + \delta_0 \qquad (2)$$

where each factor x follows the subsequent dynamic:

$$dx_t = -Kx_t dt + \Sigma dw_t \qquad (3)$$

being K and Σ both diagonal matrixes where each component k_{ii} and σ_{ii} corresponds to the mean-reversion speed and the variance of each x_i factor, respectively. Hence, the risk-adjusted process is:

$$dx_t = -(\lambda_i + Kx_t)dt + \Sigma dw_t \tag{4}$$

where λ_i is the risk-premia for each factor and dw_t is a vector of correlated Brownian motion processes such that:

$$(dw_t)(dw_t)' = \Omega dt \tag{5}$$

Finally, in order to obtain the security price, the Itô's Lemma is applied to obtain a pure discount bond price:

$$P(x_t, \tau) = e^{(u(\tau_i)'x_t + v(\tau_i))} \tag{6}$$

where:

$$u_i(\tau) = -\left(\frac{1 - e^{-\kappa_i \tau}}{\kappa_i}\right) \tag{7}$$

$$v(\tau) = \sum_{i=1}^{n}\left(\tau - \frac{1 - e^{-\kappa_i \tau}}{\kappa_i}\right) - \delta\tau + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\sigma_i \sigma_j \rho_{ij}}{\kappa_i \kappa_j}\left(\tau - \frac{1 - e^{-\kappa_i \tau}}{\kappa_i} - \frac{1 - e^{-\kappa_j \tau}}{\kappa_j} + \frac{1 - e^{-(\kappa_i + \kappa_j)\tau}}{\kappa_i + \kappa_j}\right) \tag{8}$$

We can apply this formulation to a coupon bond using the following expression:

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

Once the price is obtained, the yield to maturity is:

$$P(x_t, \tau) = \sum_i^{Coupons} C_i e^{-yt_i}$$

1.3.3 Data mining

Since 1960, computational calculus power and storage capacities have evolved rapidly. Calculations that took weeks or months are now performed in a matter of seconds. With large datasets available, analysis techniques were developed in order to get valuable information at reasonable costs.

Data mining techniques can be subdivided into two main categories: descriptive and predictive.

Examples of descriptive algorithms are clustering and association rules. Clustering is an algorithm that groups objects in a way that objects in the same group are very similar and different with objects in other groups. Association rules consists in finding relations between variables that may apparently not be related from a large dataset.

Some predictive algorithms are neural networks and genetic algorithms. Neural networks try to emulate neurons structure with connections between different layers that receive input data and output the prediction. Genetic algorithms, on the other hand, are algorithms that mutate over time trying to adapt into a better fit of the results.

The first step for using data mining algorithms is to select the attributes that will be used for training and testing the different techniques. It is not always obvious which features to use that will add predictive power without being too costly. Several techniques have been developed to address the issue of feature selection as shown first in Blum & Langley (1997) and later, extending the analysis to problems with more than 40 features, in Guyon & Elisseeff (2003).

Variable selection is crucial for many reasons. Having the least features incorporated into the analysis will facilitate data visualization and understanding; will also reduce measurement and storage capacities. Analysis will be considerably faster. Finally, and not so obvious, having an optimal set of variables will help improving predictive power. It is important to notice that selecting the most useful variables is not the same as choosing the most relevant ones. There could be many relevant variables but selecting useful features will help produce a better predictive algorithm even if it has to exclude some relevant variables. Discussions about relevance and usefulness are presented in Kohavi & John (1997) and Blum & Langley (1997).

Different techniques for selecting features have been developed. The first common approach is variable ranking as used in several papers, e.g. Bekkerman et al. (2003) Cauana & de SA (2003), Forman (2003), Westo et al. (2003). This technique ranks all variables in order of relevance in terms of predictive power. However, these algorithms have the problem that does not discard redundant features. Common ranking techniques are correlation criteria, single variable classifiers and information theoretic ranking criteria. The problem with ranking is that it tends to produce a redundant set of features that could impact in prediction power and processing time.

The second common approach is variable subset selection. These algorithms try different subset of variables and test them in order to get the optimal subset with more predictive power, punishing the ones with more variables for getting the minimal subset. This method has proven to be better than ranking because it can search through all the space of subsets of variables but it could be very expensive in computational costs terms if the number of features is considerable. Commonly used variable subset selection techniques include wrapper method, filters and embedded methods. Wrapper implementation can be seen in Kohavi & John (1997)

The particular data mining technique implemented in this thesis is Classification and Regression Trees (CART) algorithms, first developed by Breiman et al. (1984). This

technique has the advantage that its result is very transparent and easy to understand as the output is a decision tree that can be translated into a list of instructions that ends up in the estimation of the specific spread. Different applications of CART algorithms can be found in the literature. Kovacic (2010) uses them for the early prediction of a student success through college. Financial examples can be found in Sorensen et al. (1999) or Seshadri (2003) who use classification and regression trees to build quantitative investment strategies for getting above-normal returns in stock markets. The three most common implementations for these trees are CART, M5P and RPTree. All of them can be found in data mining software.

The construction of the regression tree is done in three steps. First, the algorithm builds a very big tree (the maximal tree), which describes better the data. Second stage produces several pruned trees, trying to get an optimal sub-tree that predicts reasonably well but it is not overfitted.

Overfitting is a very common issue in data mining algorithms. It happens when the model that is being used is trained and produces a result that predicts perfectly the training set but it does not predict good when a test set is provided. This is due to the loss of generalization of the algorithm when it tries to fit perfectly with the training set. In order not to overfit, the tree is pruned, losing accuracy in the training set but earning prediction power outside of it.

Finally, third stage selects through cross-validation the optimal tree from the pruned subset of trees in terms of prediction power and complexity.

The optimal tree is then used to predict the specific spread and it is trained every year with new data.

1.3.4 Future Research

As described earlier, this two stages pricing model can be used to price any fixed income security. This work used a particular instrument, the mortgage-backed notes

because it was an interesting case with many variables that are hard to take into account with traditional models. A first line of future research would be to use this model for pricing another type of fixed-income security. This model applies to diverse instruments, so it would make sense to use it for pricing –for instance- corporate bonds, as there would be sufficient information that distinguishes one transaction from another that could feed the data mining algorithm.

Another line of future research would be to build or implement an algorithm or statistical test that would compare and show the evolution of variables from the CART trees and its relevance. Even though this work shows what features are the most important and it is easy to see the decision path for a particular year when calculating the remaining spread for a transaction; it is difficult to determine the absolute importance of every variable and how it changes from one period to another due to the complexity of the trees and its size. Implementing an algorithm that calculates the relevance of variables in every period would be interesting for understanding even better the behavior of the market.

2 MERGING DYNAMIC FINANCIAL MODELS WITH DATA MINING TECHNIQUES FOR PRICING FIXED INCOME SECURITIES IN LOW-LIQUIDITY MARKETS

The term structure of interest rates is the key factor in bond valuation. The two most popular models, proposed by Nelson and Siegel (1987) and Svensson (1994), assume a parsimonious term structure with limited number of parameters and the use of only contemporary data. These models have been very useful when a complete[1] data panel is available, but behave poorly when there is a missing data problem which is typical of emerging markets (Cortazar et al. 2007).

To better handle missing observations in the valuation of corporate bonds, the use of dynamic models, which include also the use of historical observations, have been proposed (Cortazar et al. 2012). There is a long tradition of dynamic models developed to value options and other instruments with prices contingent on volatility. One of the first is Vasicek (1977) who models the spot interest rate using one stochastic factor. Several extensions to these models have been proposed later on. For example, Brennan and Schwartz (1979) analyze a two factor model (short and long rate). Litterman and Scheinkman (1991) use principal component analysis to show that three factors would be required to explain 99% of yield variance. Later, Cortazar et al. (2007) generalize the

[1] i.e. there is data for all dates and maturities.

Vasicek model to *n* unobservable stochastic factors and estimate the parameters of an incomplete data panel using the Kalman filter (Kalman 1960).

Given that typically liquidity in corporate bonds is much lower than in risk free governmental securities, the traditional approach to value risky securities is to first calibrate a risk free term structure and then to add a spread. The main component of this spread is credit risk, which depends on the default probability (Merton 1974; Jarrow and Turnbull 1995; Duffie and Singleton 1999). There are many approaches to estimate this spread. Among them for low-liquidity markets Cortazar et al. (2012) obtain term structures of corporate bond spreads based on credit ratings using a Vasicek framework.

Even though the above models are based on a strong theoretical background and have proven to be very effective when trying to explain the aggregate behavior of a specific family of instruments (e.g. treasury bonds, corporate bonds), securities with particular characteristics may not be well priced using only an "average" term structure. Moreover the requirements for a model to be able to adequately price each specific security would be very stringent having to include many relevant characteristics, such as call probability, liquidity premium, and default risk, etc. This is the case when computer algorithms become useful.

Computer science has evolved in an exponential way in the past decades. Nowadays, extensive calculations that include massive amounts of data are done in a matter of seconds. Along with calculation power, storage capacity has strongly improved. Extensive literature has been developed in managing big and diverse datasets

for extracting valuable information in all fields of knowledge. This subfield of computer science, known as *Data* Mining, has been used for many purposes. For instance, West (2000) uses five different neural networks algorithms for building a credit scoring model in order to discriminate between "good credits" and "bad credit" bank customers. Other applications have been implemented, such as Adomavicius and Tuzhilin (2001) who build computer software that can create customer profiles and predict behavior as well as characterize groups of costumers using association rules and classification algorithms; or Creamer and Freund (2010) who rank accounting and corporate variables according to their impact on performance. Also data mining can be helpful for scientific purposes. For example, Gao et al. (2010), use classification techniques to analyze a combination of internal and external weather conditions to predict and optimize comfort levels based on heating, ventilation and air conditioning.

In this paper we propose merging these approaches in a two-stage procedure: first, use an *n*-factor Vasicek dynamic model to obtain a term structure for pricing the average security, and second, use Classification and Regression Trees (CART) algorithms for estimating the specific spread for a particular fixed income instrument using large data-sets including detailed characteristics of all securities.

To illustrate our model we implement it for the family of mortgage-backed securities that trade in the low-liquidity Chilean fixed-income market. This kind of instruments has been the focus of financial markets for the past few years after the sub-prime crisis in 2008. They are difficult to price because of their diversity and particular

features such as: coupon rate, prepayment probability, default risk (credit rating), issuer, among others.

We use daily data for five sample periods from 2005 to 2011. The model is calibrated with the first year as the in-sample data and then tested out-of-sample with the following semester.

The rest of the paper is laid out as follows: Section 2 describes the model; Section 3 takes a look at the data; Section 4 examines the empirical results; and finally Section 5 concludes.

1. THE MODEL

In this section we present the model to price a mortgage-backed security decomposing its yield into two parts, estimated using different approaches, and an error term:

$$Yield_{MN} = Yield_{ATS} + s_{specific} + \varepsilon \qquad (9)$$

where $Yield_{MN}$ stands for the yield of the mortgage note; $Yield_{ATS}$ is the component explained by the average term structure; $s_{specific}$ is the additional specific

spread for that note which depends on its particular features; and ε represents the model error.

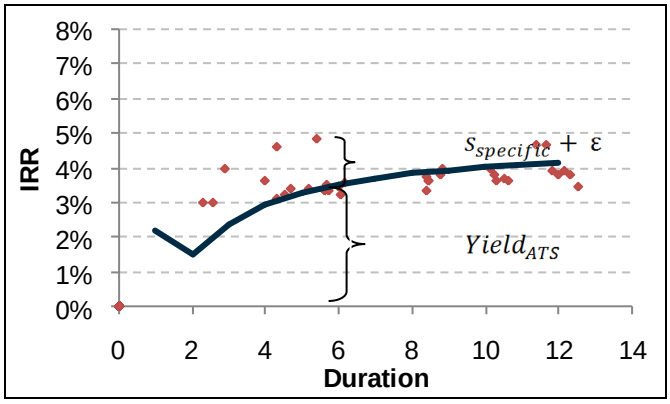

Figure 1 - The model: The line represents the Average Term Structure estimated using the financial model and the points are yields of different mortgage notes. The difference between the line and the points is the specific spread, which will be estimated using data mining, plus the error.

We propose estimating the components of the yield in two stages. The first stage, the estimation of the average term structure, is done using a multivariate dynamic financial model that captures the average behavior of the market. The second stage incorporates classification and regression trees that, using the particular information on the security, estimate the specific spread in order to get a more accurate pricing fit.

1.1 Stage One: Estimating the average term structure

In order to represent the behavior of the market as a whole, we estimate an average term structure using Cortazar et al. (2007) which corresponds to an n-factorial Vasicek model, in which the short term interest rate is defined as follows:

$$r = 1'x + \delta_0 \quad (10)$$

where each factor x follows the process:

$$dx_t = -Kx_t dt + \Sigma dw_t \quad (11)$$

being K and Σ both diagonal matrixes, where each component k_{ii} and σ_{ii} corresponds to the mean-reversion speed and the variance of each x_i factor, respectively. Hence, the risk-adjusted process is:

$$dx_t = -(\lambda_i + Kx_t)dt + \Sigma dw_t \quad (12)$$

where λ_i is the risk-premia for each factor and dw_t is a vector of correlated Brownian motion processes, such that:

$$(dw_t)(dw_t)' = \Omega dt \quad (13)$$

Finally, the pure discount bond price is obtained using Itô's Lemma:

$$P(x_t, \tau) = e^{(u(\tau_i)'x_t + v(\tau_i))} \quad (14)$$

where:

$$u_i(\tau) = -\left(\frac{1 - e^{-\kappa_i \tau}}{\kappa_i}\right) \tag{15}$$

$$v(\tau) = \sum_{i=1}^{n}\left(\tau - \frac{1 - e^{-\kappa_i \tau}}{\kappa_i}\right) - \delta\tau + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\sigma_i \sigma_j \rho_{ij}}{\kappa_i \kappa_j}\left(\tau - \frac{1 - e^{-\kappa_i \tau}}{\kappa_i} - \frac{1 - e^{-\kappa_j \tau}}{\kappa_j} + \frac{1 - e^{-(\kappa_i + \kappa_j)\tau}}{\kappa_i + \kappa_j}\right) \tag{16}$$

Thus the price of a coupon bond is:

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

The yield to maturity, *y*, can be obtained from the following equation:

$$\boldsymbol{P(x_t, \tau)} = \sum_{\boldsymbol{i}}^{\boldsymbol{Coupons}} \boldsymbol{C_i e^{-y t_i}}$$

To estimate this model the Extended Kalman Filter is applied. This method uses a state-space representation, allows for errors in the measurement of the state variables, and handles a non linear measurement equation required for coupon-paying instruments. Also the Kalman filter may be successfully used with incomplete data panels, (Cortazar and Naranjo (2006)).

The measurement equation of the Standard Kalman Filter is:

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

where z_t is a $(m_t \; x \; 1)$ vector that has all the observable variables in t; H_t is a $(m_t \; x \; n)$ matrix where n is the model's number of state variables x_t; d_t is a vector of constants that represents the distance between the observable and non-observable variables; and v_t is a non-correlated stochastic variable with mean zero and variance R_t.

The transition equation which defines state variables dynamics is:

$$x_t = A_t x_{t-1} + c_t + \varepsilon_t \qquad \varepsilon_t \sim N(0, Q_t) \qquad 17)$$

From equations (9) and (10), we obtain the prediction equations:

$$x_{t|t-1} = A_t \hat{x}_{t-1} + c_t \qquad 18)$$

$$P_{t|t-1} = A_t P_{t-1} A'_t + Q_t \qquad 19)$$

Finally, the optimal state variable estimate is the predicted updated using new information:

$$\hat{x}_t = \hat{x}_{t/t-1} + P_{t/t-1} H_t^T F_t^{-1} v_t \qquad 20)$$

$$P_t = P_{t/t-1} - P_{t/t-1} H_t^T F_t^{-1} H_t P_{t/t-1} \qquad 21)$$

where:

$$F_t = H_t P_{t/t-1} H_t^T + R_t \qquad 22)$$

$$v_t = z_t - (H_t \hat{X}_{t/t-1} + d_t) \qquad 23)$$

Cortazar et al. (2007) show how to apply the Extended Kalman filter by linearizing the measurement equations.

Parameter estimates can be obtained by maximizing the log-likelihood function:

$$\log L(\psi) = -\frac{1}{2}\sum log|F_t| - \frac{1}{2}\sum v_t^T F_t^{-1} v_t \qquad 24)$$

where ψ represents a vector of unknown parameters.

1.2 Stage Two: Estimating the specific spread

Most families of fixed-income securities include instruments with a diverse set of characteristics which trade at prices that deviate from the family average. Hence we propose including a second stage which estimates this deviation, $s_{specific}$, from the average term structure obtained in the first stage. This spread should depend on some undetermined function of the many specific characteristics the instrument has.

Instead of attempting to build a complex financial model that takes into account all relevant variables, the approach includes data mining techniques to estimate the specific spread. In particular, we propose using the Classification and Regression Trees (CART) because the output is an easy-to-interpret set of rules that explains the way the market considers different variables for pricing mortgage notes. This algorithm is often

used in the economic and financial context because of its *transparency* feature: the decision tree can be represented as a set of decisions in plain English. It is also very fast and efficient in terms of computational resources.

CART algorithms were introduced in Breiman et al. (1984). In the literature, several implementations of CART algorithms have been proposed and applied in many areas. Kovacic (2010) uses them for the early prediction of a student success through college. Financial examples can be found in Galindo and Tamayo (2000) that make a comparative analysis of different statistical and machine learning modeling methods of classification on a mortgage loan data set for credit risk assessment, and in Sorensen et al. (1999) and Seshadri (2003), who use classification and regression trees to build quantitative investment strategies for getting above-normal returns in stock markets.

We choose the three most common implementations for classification and regression trees: CART, M5P and RPTree. Our goal is to obtain an estimation of the spread as accurate as possible, and to compare prediction power. Models use each year of data as the training set and the subsequent semester as the test set. All parameters for each model are optimized according to standard considerations for this kind of algorithms.

The algorithm determines the best combination of variables that explains the specific spreads and creates a decision model that calculates them for any specific transaction. The algorithm also prunes the tree to optimize predictive power. Prune parameters are also optimized for every period.

2. IMPLEMENTATION DATA

The dataset consists of five years of mortgage-backed note transactions from the Chilean market covering from 01/01/2006 to 06/30/2011. Table 1, presents a summary of the number of transactions in each period and the total amount traded for each semester in billions of Chilean billion Pesos (CLP bn).

Table 1. Summary for the Chilean mortgage-backed notes market. *Transactions* is the number of transactions in the period and *Amount* is the amount traded in CLP bn.

		Transactions	Amount (CLP bn)
2011			
	1 Semester	2119	289
2010			
	1 Semester	2750	185
	2 Semester	2815	176
2009			
	1 Semester	2890	445
	2 Semester	3217	415
2008			
	1 Semester	3234	501
	2 Semester	3173	486
2007			
	1 Semester	6422	994
	2 Semester	4127	447
2006			
	1 Semester	9536	1455
	2 Semester	7820	1176

Table 1 shows that the total number of transactions and amount traded has been decreasing over time denoting a very low-liquidity market which justifies the use of a Kalman Filter estimation.

For each transaction in the dataset the yield, the Macaulay duration and a list of additional features is gathered. For example in this market *callability* adds uncertainty to the payoffs, thus investors demand an extra spread. We present the complete list of attributes for each transaction that could affect spreads that were included in the data base:

- *Note issuer:* the financial institution that issues the note.
- *Maturity:* the remaining time until the last payment of the note.
- *Coupon:* the interest rate that the note-issuer has agreed to pay.
- *Risk Grade:* the risk classification assigned by rating agencies.
- *Turnover:* measure of liquidity calculated as the amount traded, divided by the total amount outstanding of the security.
- *Macaulay Duration:* duration measure.
- *Presence:* relative-size factor calculated as the outstanding amount divided by the total amount of mortgage notes available in the Chilean market.

- *Historical call probability:* Three indices, each one calculated with the historical prepayment information during one, five and ten years, as a predictor of future early exercises.

3. EMPIRICAL RESULTS

In order to calibrate the parameters of the multifactorial Vasicek model, we use transactions from one year (January to December) and then evaluate, out-of-sample, the following semester. We repeat this process for each period from 01/01/2006 to 12/31/2011.

To assess the estimation accuracy, we use the mean average error (MAE) between the observed and the estimated yield.

3.1 Stage One: Dynamic Model for the Average Term Structure

In this section we present the results of the model to obtain the average term structure. In Figures 2 to 5 we can see the estimated average term structure and the actual transactions both in and out-of-sample, on the left and right side, respectively. Each figure shows the yield of securities for different durations..

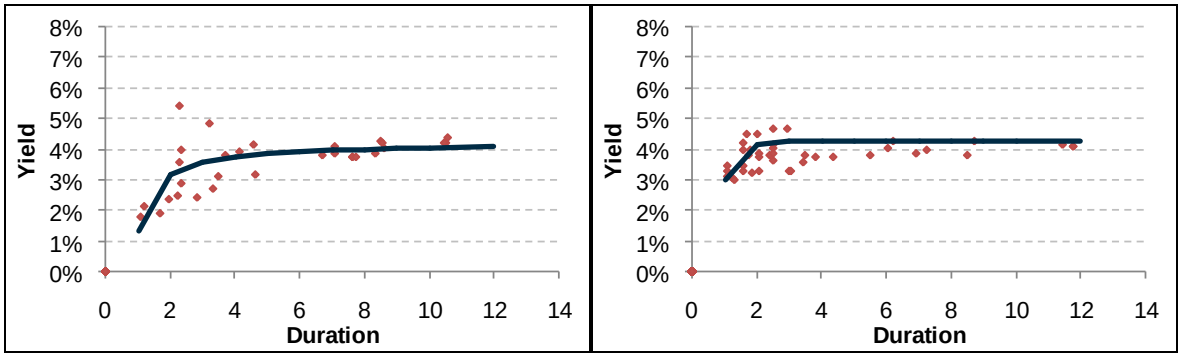

Figure 2. In-sample 2010 (left) and out-of-sample first semester of 2011 (right) average term structure and actual transactions

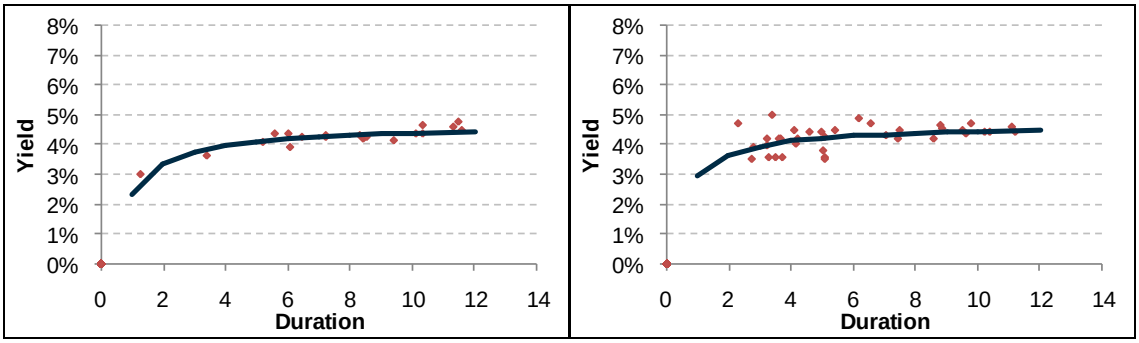

Figure 3. In-sample 2009 (left) and out-of-sample first semester 2010 (right) average term structure and actual transactions

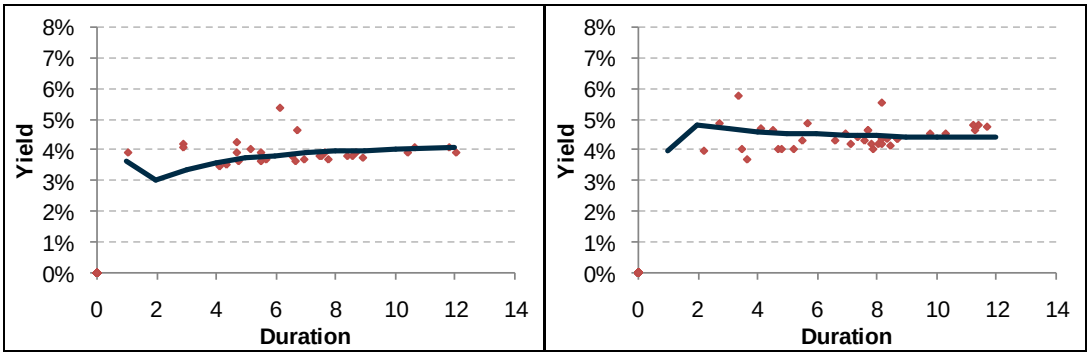

Figure 4. In-sample 2008 (left) and out-of-sample 2009 (right) average term structure and actual transactions

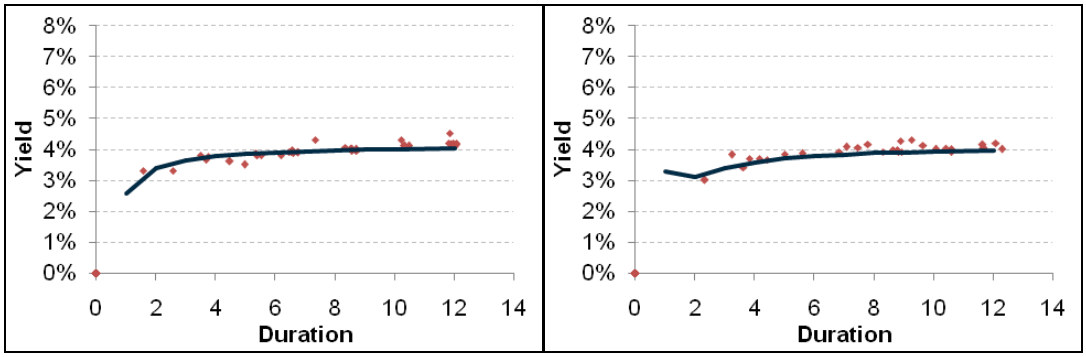

Figure 5. In-sample 2007 (left) and out-of-sample first semester 2008 (right) average term structure and actual transactions

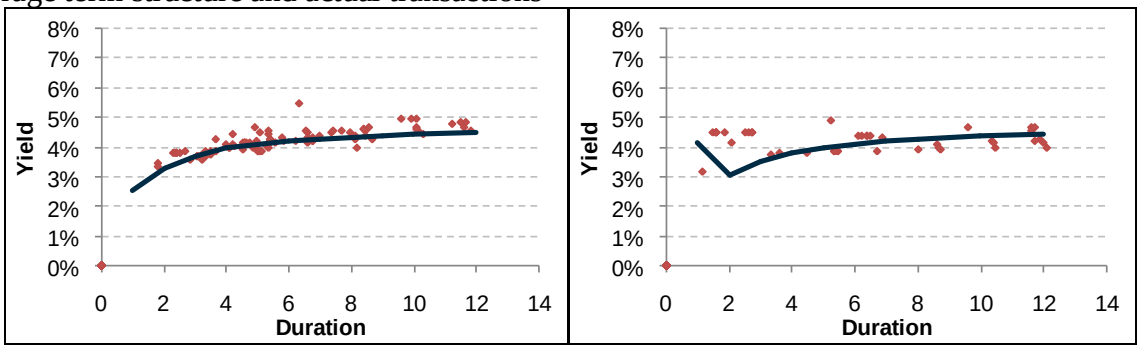

Figure 6. In-sample 2006 (left) and out-of-sample first semester 2007 (right) average term structure and actual transactions

It can be seen that the term structure explains relatively well the average behavior of the market. The underlying dynamic model that generates the structure and the

estimation procedure incorporate both historic and current information in order to have a more stable solution.

At this stage of the analysis, all the transactions are considered similar, only diverging in their durations. It is impossible at this level to discriminate between different issuers, credit risk rating, prepayment probability, etc. It is easy to see that the term structure adapts well to the observations, even though notes have different characteristics. Table 2 shows the mean average error (MAE) for different periods, both in and out-of-sample.

Table 2 MAE of the estimation for stage 1 for different time periods.

	MAE (%)		
	in sample	out of sample	total
2010 - 2011	0.429	0.286	0.390
2009 - 2010	0.304	0.270	0.296
2008 - 2009	0.231	0.285	0.248
2007 - 2008	0.248	0.233	0.245
2006 - 2007	0.263	0.271	0.265

From Table 2 we can see that the term structure gives a reasonably good and stable estimation of the price for the market average. Even though market liquidity measured as number of transactions declines over time, the estimation error at this stage remains relatively low. Also it is important to highlight that the in and out-of-sample

errors are relatively similar. These results are consistent with a robust model with a stable average solution.

3.2 Stage Two: Classification and Regression Trees for the Specific Spread

After the estimation of the average term structure, the specific spread for each transaction is estimated using data mining techniques. In particular, CART algorithms, which have proven to be even more effective than other tree algorithms when classifying data (Yadav et al. 2012), are used.

Feature selection is crucial at this stage of the analysis. Precise identification of characteristics that are important and have predictive power allows the model to: (i) improve prediction performance, (ii) provide faster and more cost-effective predictors and (iii) give a better understanding of the underlying process that generated the tree (Guyon and Elisseeff 2003).

Different approaches were used for selecting the most relevant attributes. First, a ranking algorithm was used. Ranking has proven to be simple, scalable and to have good empirical results. Different authors use attribute ranking as a baseline method (e.g. Bekkerman et al. 2003, Caruana and de Sa 2003, Weston et al. 2003).

Table 3 shows the output obtained from the ranking algorithm. According to this selection criterion the best features for constructing the classification trees are coupon, historical call probability and Macaulay duration.

Table 3. Results of the ranking algorithm for each feature and period

	Ranking					Average
	2010 - 2011	2009 - 2010	2008 - 2009	2007 - 2008	2006 - 2007	
Issuer	3	7	7	8	9	7
Maturity	10	6	6	6	5	7
Coupon	1	1	1	1	1	1
Risk Grade	5	9	10	10	10	9
Turnover	9	10	9	9	8	9
Macaulay Duration	2	5	4	5	3	4
Presence	8	8	8	7	7	8
Call Probability 1 year	4	4	2	2	2	3
Call Probability 5 year	6	3	5	3	6	5
Call Probability 10 year	7	2	3	4	4	4

However, ranking techniques are not used to determine the definitive subset of attributes because it does not necessarily eliminate redundancy between the selected variables. In order to obtain the most relevant subset of features to build the best classificator and to assure optimality, a correlation and a wrapper algorithm are used. Results for both are shown in Table 4 and 5.

Table 4. Results of the correlation algorithm for each feature and period

	Correlation					No. Times Selected
	2010 - 2011	2009 - 2010	2008 - 2009	2007 - 2008	2006 - 2007	
Issuer	No	No	No	No	No	0
Maturity	No	No	No	Yes	No	1
Coupon	Yes	Yes	Yes	Yes	Yes	5
Risk Grade	No	No	No	No	Yes	1
Turnover	No	Yes	No	Yes	Yes	3
Macaulay Duration	No	No	No	No	No	0
Presence	No	No	No	No	No	0
Call Probability 1 year	No	No	No	No	No	0
Call Probability 5 year	No	No	No	No	No	0
Call Probability 10 year	No	No	No	No	No	0

Table 5. Results of the wrapper algorithm for each feature and period

	Wrapper					No. Times Selected
	2010 - 2011	2009 - 2010	2008 - 2009	2007 - 2008	2006 - 2007	
Issuer	Yes	Yes	Yes	Yes	Yes	5
Maturity	No	Yes	Yes	Yes	Yes	4
Coupon	Yes	Yes	Yes	Yes	Yes	5
Risk Grade	Yes	No	No	No	No	1
Turnover	No	No	No	Yes	Yes	2
Macaulay Duration	Yes	No	Yes	No	No	2
Presence	Yes	No	No	No	Yes	2
Call Probability 1 year	No	No	No	No	Yes	1
Call Probability 5 year	No	No	No	No	No	0
Call Probability 10 year	No	No	No	No	No	0

Results show that coupon is the most important feature. This is logical because if the coupon rate is too high, implies that the probability of default or the call probability is also high, thus the spread should increase.

The main difference in both tables is that issuer comes as relevant in the wrapper algorithm, while it is not in the correlation algorithm. Also it is interesting to notice that call probability is not an important feature in any of the approaches, probably because this feature has already been picked up by the coupon variable.

It is important to recall that the wrapper algorithm always finds the best solution because of its brute-force nature. The disadvantage of this technique is that it is very intensive in computational resources and sometimes impossible to use. However, in this particular case, an optimal subset of attributes was found in less than an hour.

After this analysis, features that were proven to be useful for estimating the specific spread of each security are four:

- *Issuer*
- *Maturity*
- *Coupon*
- *Turnover*

To measure model performance the first year of data is used to train the model and the following semester to test the accuracy of the estimations, as we did in stage

one. All construction parameters were also optimized in order to get the most general and accurate tree.

Three implementation algorithms were used: *CART, M5P* and *RPTree*. The output is the specific spread that should be added to the average market yield estimation in order to get a more accurate price.

Table 6 shows the remaining MAE after adding to the average term structure the specific spread estimated by the regression tree using the three algorithms implemented. It can be seen that the difference between algorithms is relatively small. Nevertheless, CART is the best performer in 4/5 of the periods. This is due to the pruning methodology which is optimal. On the other hand, M5P is the worst performer. This algorithm tends to produce overfitted trees that behave very well in-sample but not out-of-sample. Moreover, algorithm speed performance is similar among the three implementations. Both training and estimating with the trees did not present problems in terms of computational power or speed with our amount of data and the software used (Weka and MATLAB).

Table 7 compares the MAE between stage-one and stage-two of the model. It can be seen that using these data mining techniques increases estimation accuracy, reducing about 120bp (average) of the MAE for all the data sets used. As expected, the tree has a larger effect in the in-sample dataset where it tends to reduce error by 134bp average. Anyway, out-of-sample performance is consistently good, especially in the 06 – 07 period where it gets a 117bp error reduction.

Table 6 MAE comparison for each implementation of stage 2.

In-sample period corresponds to one year and out-of-sample is the following semester.

	CART			M5P			RPTree		
	in sample	out of sample	total	in sample	out of sample	total	in sample	out of sample	total
2010 - 2011	0.288	0.242	0.275	0.277	0.210	0.258	0.256	0.240	0.251
2009 - 2010	0.168	0.196	0.175	0.159	0.198	0.169	0.154	0.208	0.167
2008 - 2009	0.161	0.179	0.166	0.149	0.178	0.158	0.142	0.202	0.160
2007 - 2008	0.124	0.177	0.137	0.118	0.172	0.131	0.111	0.179	0.127
2006 - 2007	0.113	0.149	0.122	0.105	0.151	0.117	0.097	0.160	0.114
Average	0.171	0.189	0.175	0.162	0.182	0.167	0.152	0.198	0.164

Table 7 - MAE evolution in stage 1 and 2 and overall change.

	Stage 1			Stage 2 Average			MAE Reduction (%)		
	in sample	out of sample	total	in sample	out of sample	total	in sample	out of sample	total
2010 - 2011	0.429	0.286	0.390	0.273	0.231	0.262	-0.155	-0.056	-0.128
2009 - 2010	0.304	0.270	0.296	0.161	0.201	0.170	-0.144	-0.069	-0.126
2008 - 2009	0.231	0.285	0.248	0.151	0.186	0.162	-0.080	-0.099	-0.086
2007 - 2008	0.248	0.233	0.245	0.118	0.176	0.131	-0.131	-0.056	-0.113
2006 - 2007	0.263	0.271	0.265	0.105	0.153	0.118	-0.158	-0.117	-0.147
Average	0.295	0.269	0.289	0.161	0.189	0.169	-0.134	-0.079	-0.120

Figures 7 to 11 show the model-fitting in both Stages 1 and 2. For Stage 2 the specific spread is subtracted to the transaction yield so the errors in the figures at the right represent those unexplained by our model after both stages are performed.

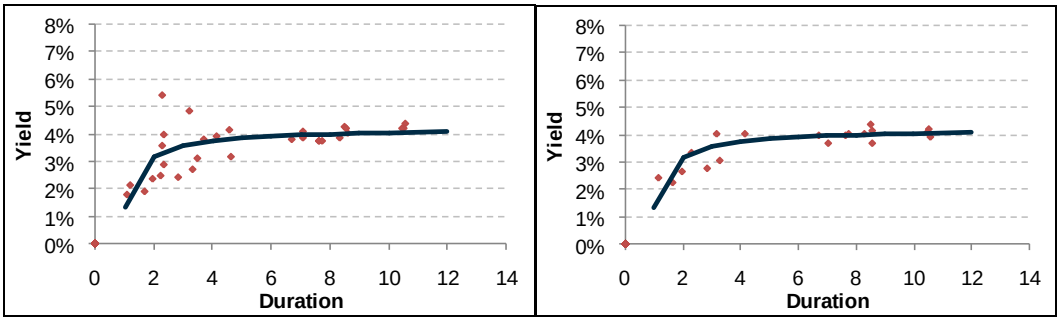

Figure 7. Term structure and modified transactions for the 2010 – 2011 period in Stage 1 (left) and Stage 2 (right). In the figure at the right the estimated specific spread is subtracted to each transaction so the remaining errors represent those unexplained by our model after both stages are performed.

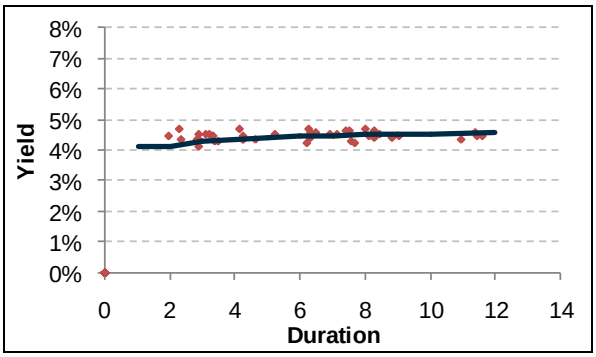

Figure 8. Term structure and modified transactions for the 2009 – 2010 period in Stage 1 (left) and Stage 2 (right). In the figure at the right the estimated specific spread is subtracted to each transaction so the remaining errors represent those unexplained by our model after both stages are performed.

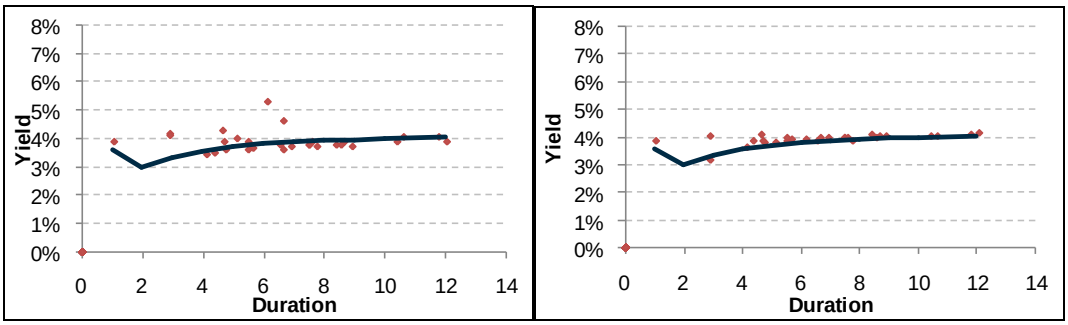

Figure 9. Term structure and modified transactions for the 2008 – 2009 period in Stage 1 (left) and Stage 2 (right). In the figure at the right the estimated specific spread is subtracted to each transaction so the remaining errors represent those unexplained by our model after both stages are performed.

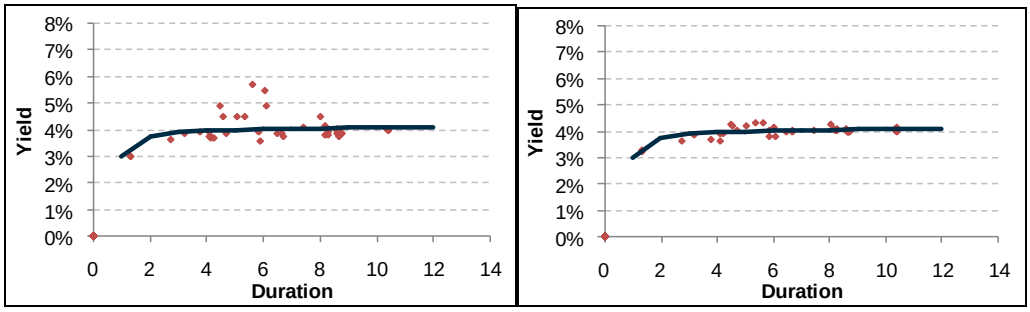

Figure 10. Term structure and modified transactions for the 2007 – 2008 period in Stage 1 (left) and Stage 2 (right). In the figure at the right the estimated specific spread is subtracted to each transaction so the remaining errors represent those unexplained by our model after both stages are performed.

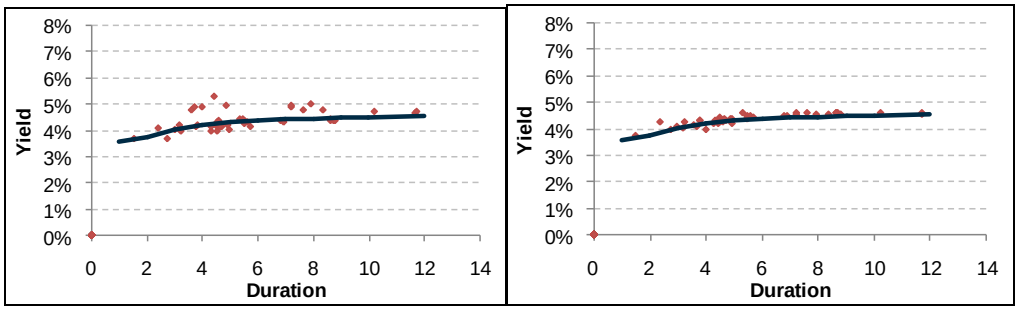

Figure 11. Term structure and modified transactions for the 2006 – 2007 period in Stage 1 (left) and Stage 2 (right). In the figure at the right the estimated specific spread is subtracted to each transaction so the remaining errors represent those unexplained by our model after both stages are performed.

The previous figures illustrate the important accuracy improvement from Stage 1 to Stage 2. Estimation error is reduced substantially. Stability through time is another attribute that can easily be observed from the charts.

To illustrate the decision trees and the variables chosen as the most important ones, Figure 12 shows only the first three levels of the tree trained with 2006 data.

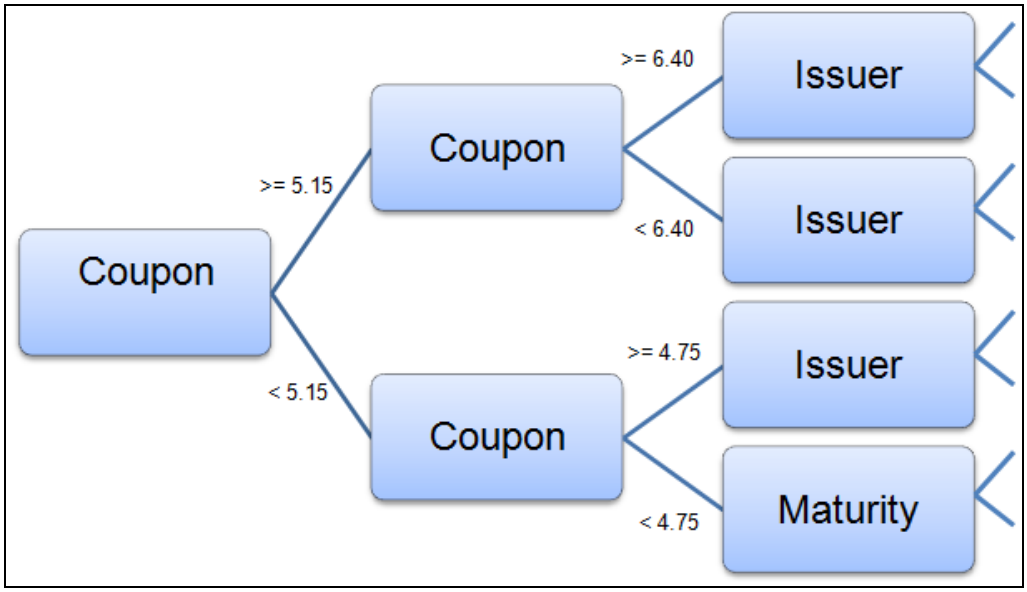

Figure 12 – Decision tree using RPTree for the 2006-2007 period

It can be seen that Coupon is the first variable that is used for a decision. This takes place in every tree that was trained and gives us valuable information about what the market takes into account when pricing different instruments. The use of the coupon rate is in line with the call and default probability since notes that have high coupon rates are more likely to be exercised earlier, especially if market conditions have changed and it is

possible to get financing at lower rates. The other variables take place in different levels and combinations for other trees, so no other feature can be recognized as more relevant than others.

4. CONCLUDING REMARKS

We propose merging a dynamic financial model and a data mining technique in a two-stage estimation method for pricing fixed income securities in a low liquidity market. We implement our model using yields of mortgage-backed notes traded in the Chilean fixed income market.

To explain transactions we first estimate a term structure to explain the behavior of an average security in a family. We propose using a three-factor Vasicek model and estimating its parameters using an Extended Kalman Filter.

On the second stage, we estimate the remaining specific spread for each particular security using classification and regression tree machine learning algorithms. Three common implementations are tested: CART, M5P and RPTree. Several attributes are gathered in order to feed the algorithms and to get the most accurate estimations.

We use five years of data from 2006 to 2011. Each year the model is estimated with the first year of transactions and then tested with the following semester of data.

Dynamic models showed that they can be very useful in term structure estimation. When parameters are estimated through the Extended Kalman Filter method, the

solution is able to capture both the history and the dynamic of the term structure. Nevertheless, due to the particular characteristics of each instrument, it is impossible to have a precise estimation of the final price with this kind of models.

Complementing this, the machine learning algorithms obtain inner data patterns that explain most of the remaining spread. With the available features it is possible to estimate the extra premium that the market demands for a particular transaction. We show that using these data mining approach, instead of only dynamic financial models, reduces estimation error in an average of 120bp.

References

ADOMAVICIUS, G. & TUZHILIN, A. (2001). Expert-Driven Validation of Rule-Based User Models in Personalization Applications. Data Mining and Knowledge Discovery, vol. 5, (1-2), 33-58.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., & STONE, C. J. (1984). Classification and Regression Trees. Wadsworth & Brooks, Monterrey, CA.

BEKKERMAN, R., EL-YANIV, R., TISHBY, N., & WINTER, Y (2003). Distributional Word Clusters vs. Words for Text Categorization. JMLR, 3:1183–1208.

BRENNAN, M., & SCHWARTZ, E. (1979). A Continuous Time Approach to The Pricing of Bonds . Journal of Banking & Finance, vol 2 (3), 133-155.

CARUANA, R. & DE SA, V (2003). Benefiting From the Variables that Variable Selection Discards. JMLR, (3), 1245–1264.

CORTAZAR, G., & NARANJO, L. (2006). An N-Factor Gaussian Model of Oil Futures Prices. *The Journal of Futures Markets*, Vol.26, No. 3, March, 2006, 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. (2012). Credit spreads in illiquid markets: Model and implementation. Emerging Markets Finance & Trade, November–December 2012, 48, 6, 53–72.

CREAMER G. & FREUND Y.(2010) Using Boosting for Financial Analysis and Performance Prediction: Application to S&P 500 Companies, Latin American ADRs and Banks. Computational Economics (2010) 36:133–151

DUFFIE, D. & SINGLETON, K. (1999). Modeling Term Structures of Defaultable Bonds. Review of Financial Studies, 12(4), 687–720.

GALINDO J. & TAMAYO P. (2000) Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications. Computational Economics 15: 107–143, 2000.

GAO, Y., TUMWESIGYE, E. & CAHILL, B. (2010), "Using Data Mining in Optimisation of Building Energy Consumption and Thermal Comfort Management," 2nd International conference on Software Engineering and Data Mining (SEDM).

GUYON, I. & ELISSEEFF, A. (2003). An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, (3), 1157-1182.

JARROW, R. & TURNBULL, S. (1995). Pricing Derivatives on Financial Securities Subject to Credit Risk. Journal of Finance, 50, 53-86.

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

KOVACIC, Z. J. (2010), “Early Prediction of Student Success: Mining Student Enrollment Data”, Proceedings of Informing Science & IT Education Conference.

LITTERMAN, R. & SCHEINKMAN, J., (1991). Common Factors Affecting Bond Returns. The Journal of Fixed Income. 1, 54-61.

LONGSTAFF, F., MITHAL, S., & NEIS, E., (2005), Corporate Yield Spreads: Default Risk or Liquidity? New Evidence from Credit-Default Swap Market, Journal of Finance, Vol. 60, 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 & SIEGEL (1987). Parsimonious Modeling of Yield Curve. The Journal of Business. 60(4):473-489

SESHADRI L. (2003). JPMORGAN US QUANTITATIVE FACTOR MODEL. *JPMORGAN QUANTITATIVE EQUITY AND DERIVATIVES.*

SORENSEN, E. H., MILLER, K. L., & OOI, C. K. (2000). The decision tree approach to stock selection. *The Journal of Portfolio Management*, *27*(1), 42-52

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.

WEST, D. (2000). Neural Network Credit Scoring Models. Computers & Operations Research, vol. 27, (11), 1131-1152.

WESTON, J., ELISSEFF, A., SCHOELKOPF, B., & TIPPING, M. Use of the Zero Norm With Linear Models and Kernel Methods. JMLR, 3:1439–1461, 2003.

YADAV, S., BHARADWAJ B. & PAL S. (2012). Data Mining Applications: A Comparative Study for Predicting Student's performance. International Journal of Innovative Technology & Creative Engineering, Vol 1.