Demand deposits modeling

a case study in a European bank

by

Alizée Aupérin

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Friday October 16, 2020 at 3:00 PM.

Student number:	4941691	
Project duration:	March 1, 2020 – October 16, 2020	
Thesis committee:	Prof. dr. ir. C. W. Oosterlee,	TU Delft, supervisor
	Dr. ir. J. Anderluh,	TU Delft
	Ir. P. Lucas,	The company
	Ir. M. A. Dumouchel,	The company

This thesis is confidential and cannot be made public until October 16, 2022.

TU Delft

Abstract

Demand deposits modeling is of top importance for banking institutions and usually represents a large part of a bank portfolio. Even though these products seem rather simple at first glance, demand deposits are without a fixed maturity, generating uncertainties in the model. A significant amount of academic literature on this subject is available. However, we are experiencing negative interest rates for a few years now, that may have affected customer behavior and have led to an excess in deposits-taking for a large majority of bank. In addition to the low-rate environment, this small European bank has been growing rapidly in recent years and new customers have different characteristics from the old ones. Demand deposits modeling is therefore a major challenge for the bank. It can be divided in three steps, respectively the market rates, the deposit volumes and the deposit rates, however, only the market rates and the deposit volumes will be considered in this thesis.

The dynamics of market rate follow a single factor Hull-White model. The mean-reverting parameter and the volatility are calibrated on historical data of the one-month Euribor rates, and then simulated using an exact Monte-Carlo approach. The deposit volume model is based on an Ornstein-Uhlenbeck process, where the constant drift term μ is replaced by a trend that depends on the rates' level. The trend has a different slope whether we are in a normal-rate environment or in a low-rate environment, and the probability of being in either state depends on the market rates. We create age and wealth categories for customer segmentation purpose. We then perform a cluster analysis, either using a k-means method or a hierarchical clustering algorithm, that will be included in the deposit volume model. The clustering forms similar groups of customers, reflecting better customers' diversity. Lastly, we compare two output variables, the average life and the optionality, for different simulations, one without the clustering, and two with the clustering. The best-case scenario for any bank being a high average life associated with a low optionality with regards to demand deposits modeling, the clustering integration in the model leads to more optimal results.

Preface

This thesis has been submitted for the degree of Master of Science in Applied Mathematics at Delft University of Technology. I would like to thank some people that have helped me with my research project. First of all, I would like to express my gratitude to my supervisors: Pierre Lucas and Marc-André Dumouchel from the company, and Kees Oosterlee from TU Delft. Kees, your valuable feedback and accurate proofreading were highly appreciated. You helped me finishing my thesis in time, despite some stress moments during the last weeks. Pierre and Marc, thank you for your daily support and your clear explanations on every single subject. I also would like to thank Jasper Anderluh for being part in my thesis committee.

Last but not least, I would like to thank my family, my boyfriend and my friends. Due to the global pandemic, a large part of my thesis has been written from my parents' place. Working from home was quite hard, however, thanks to your support, I managed to stay focused.

Alizée Aupérin
Saint Pierre, October 2020

Contents

List of Definitions

Accounting balance
The accounting balance is the volume coherent with the balance sheet items each month.

Average Accounting balance
The average accounting balance is defined as the total accounting balance divided by the total number of accounts.

Demand Deposits Account
Demand deposits are a subset of non-maturing deposits. A demand deposit account is an account from which deposited funds have no maturity, can be withdrawn at any time from the bank, and without any notice. It provides the money consumers need for purchasing daily expenses. Banks can pay interest on demand deposit accounts, but this is not a requirement. In this thesis, the bank pays no interest on demand deposit accounts.

Cohort
In this thesis, a cohort will be mostly defined on an annual scale. For instance, the cohort 2010 contains all the customers of the bank that activated their accounts during the year 2010.

Dynamic scenario
A dynamic scenario corresponds to a business-as-usual scenario, where the bank has the possibility of acquiring new customers. Customers, both new and old ones, can add or withdraw money on their account.

Interest Rate Risk
The interest rate risk can be defined as the risk incurred in case of interest rates' variation as a result of all balance sheet and off-balance sheet items.

Intra-Month Volatility
The ALM department performs transactions all over the month, so they want to know the average monthly balance. However, it is the end-of-month balance that is given on the balance sheet. The end-of-month balance can be either underestimated or overestimated. In the case of demand deposits, the end-of-month balance is overestimated compared to the average monthly balance, because customers usually have important payments at the beginning of each month (loan repayment or payment of rent). A monthly adjusted coefficient is then applied to ensure that the overestimation (or underestimation) is not taken into account. This coefficient is assumed to be constant over the life of the product, and has to be determined before starting the deposits volume model.

Liquidity Risk
Liquidity can be defined as "the ability of a bank to fund increases in assets and meet obligations as they come due, without incurring unacceptable losses. The fundamental role of banks in the maturity transformation of short-term deposits into long-term loans makes banks inherently vulnerable to liquidity risk, both of an institution-specific nature and that which affects markets as a whole" [3].

Optional component of the stable balance
The optional component of the stable balance is the proportion of the stable balance that exists due to the rates conjecture. In other words, the stable part of demand deposits (either the accounting balance or the average accounting balance) includes a term that is independent of the level of interest rates and a extra-term that exists only due to the context of low rates. As a matter of fact, because of the low-rates environment that is going on since a few years, banks have an excess of deposits-taking that is only attributable to low interest rates.

Optionality

The optionality is defined as the percentage of optional component in the stable balance. In other words, it is the ratio of the optional component of the stable balance and the stable balance. With regards to non-maturing modeling, the best case scenario for the bank is a high average life, defined in subsection 2.6.2, associated with a low optionality.

Seasonality

The variation of accounting balance is clearly dependent on the month. As a matter of fact, some expenses only appear once a year, at Christmas or during summer holidays for instance. Therefore, the seasonal coefficient of each month has to be estimated, and the monthly data of accounting balances should be deseasonalised.

Stable accounting balance

Let $Vol_{IM} \in [0,1]$ be the intra-month volatility and $S(i)$ be the seasonal coefficient of month i. Then, for this same month, the stable accounting balance $D^S(i)$ is defined as

$$D^S(i) = D(i)\frac{Vol_{IM}}{S(i)}, \tag{1}$$

where $D(i)$ of the accounting balance of a month i.

Stable part of demand deposits

The stable part of demand deposits corresponds to the volume of demand deposits, where the volatile part has been removed. We can distinguish two components into the stable part, that are respectively the optional component of the stable part and the non-optional component of the stable part. The first one, the optional component of the stable part, is defined above. The second component is independent of the level of interest rates. Nowadays, each bank aims to prove that the percentage of accounting balance that depends on the specific rates context among the stable part of accounting balance is as minimal as possible. This will prove that the optionality on demand deposits is low.

Static scenario

A static scenario corresponds to a scenario where the bank cannot acquire any new customers, and where current customers can only withdraw money from their account.

List of Abbreviations

AAB	Average Accounting Balance ($AAB = AAB^S + ABB^{Vol}$)
AB	Accounting Balance ($AB = AB^S + AB^{Vol}$)
AAB^O	Optional Stable Part of Average Accounting Balance
AAB^{Non-O}	Non-Optional Stable Part of Average Accounting Balance
AAB^S	Stable Part of Average Accounting Balance ($AAB^S = AAB^O + AAB^{Non-O}$)
AAB^{Vol}	Volatile Part of Average Accounting Balance
AB^O	Optional Stable Part of Accounting Balance
AB^{Non-O}	Non-Optional Component of Accounting Balance
AB^S	Stable Part of Accounting Balance ($AB^S = AB^O + AB^{Non-O}$)
AB^{Vol}	Volatile Part of Accounting Balance
AL	Average Life
$D(t)$	Deposits Volume at time t (either the average accounting balance or the accounting balance)
EURIBOR	EURo InterBank Offer Rate
N^{CA}	Number of customer Accounts
NMDs	Non Maturity Deposits
$O_{\%}$	Optionality
r^A	Attrition rate

1 Introduction

1.1. Background

Banks play an important role in our modern economic society. They receive deposits from their clients, let them withdraw money whenever they want, and they also provide loans. They act as a financial intermediary by channelling funds between savers and borrowers. Banks make profit, among other things, from the spread between the interest rate received from overdue loans and the paid deposit rate. Each bank is the only responsible of the fair prices of its services in accordance with the financial markets and the competition, as well as its marketing and financial policy.

Over the past few years, banks have drastically changed. They suffered from the ”subprime” crisis in 2008, that was primarily due to a high debt load of American customers followed by a fall in real estate prices in the United States. This crisis then spread into a global economic shock and a succession of bank failures. This financial and economic crisis led to new regulations through the publication of Basel III to prevent from a new financial collapse. In a few words, this international accord introduced, in 2010, reforms to improve the regulation, supervision and risk management within the banking sector. One of the most important innovations was the introduction of the Liquidity Coverage Ratio (LCR) and the Net Stable Funding Ratio (NSFR). Apart from those new regulations, we have witnessed the trivialization of low interest rates. Therefore, all departments of the banks have been affected by those changes, and so was the Asset-Liability Management (ALM) department.

ALM departments in banks are, amongst others, in charge of liquidity and interest rate risk management. The liquidity risk is the risk that a company or a bank fails to meet short term financial demands and the interest rate risk is the potential for investment losses that result from a change in interest rates. Those risks arise from assets and liabilities mismatches. To begin with, banks' assets and liabilities can be divided in two groups, that are respectively the maturing products and the non-maturing products. On one side, maturing products have predetermined characteristics and the cash flows and interest rates behavior are for instance already defined in a contract. Among the maturing products, we can point the personal loans, the mortgages,… and so forth. On the other side, the non-maturing products, such as demand deposits or credit cards, do not have contractual defined characteristics.

Nowadays, we are in a low interest rate environment in Europe, and negative market rates became commonplace. This is one more consequence of the ”subprime” crisis. Therefore, the strategy of investing money in short term market rates, that are now often negative, can be a bad one for the banks. Then, the challenge for ALM departments in a bank is to model both non-maturing and maturing products in the best possible way.

This thesis will only focus on demand deposits. As the name suggests, demand deposits do not have a predetermined maturity and can be withdrawn at any time. Even though these products seem rather simple at first glance, modeling demand deposits is such a challenging task for two main reasons. As a matter of fact, customers can withdraw or raise money on their account whenever they want and without any penalty or notice. In the meantime, banks can adjust the deposit rate. Modeling demand

deposits is then a top priority for many banks as it can represent a large amount of funding.

Demand deposits modeling is a three-step process: we need one model for the market rates, one model for the deposit volumes and finally one for the deposit rates. Those three models can be either completely separated or linked through one single model. In this thesis, demand deposits modeling will constitute three different steps, but we will assume that the deposit rates are always zero.

1.2. The role of ALM within the banking sector

Asset-Liability Management (ALM) is a generic term that is used nowadays in a variety of fields. It is used both by banking and insurance companies. In this thesis, ”ALM” will only refer to ”ALM in the banking sector”.

Asset-liability management aims to manage the balance between the income derived from the placements of funds and the cost of the sources of funds. ALM endeavors to measure the levels of liquidity, interest rate and exchange rate risk. The goal is to preserve the long-term profitability of the bank despite the fluctuations of the two main risk factors of the banking activity - liquidity and interest rate - in the context of possible changes in the balance sheet structure related to changes in customer demands.

1.3. Research question

This thesis is a case study in a small European bank that benefits from a rapid growth in recent years. This rapid growth has led to a loss of homogeneity among customers. In what follows, the bank will refer to this small European bank.

The objective of the thesis is to model demand deposits. As aforementioned, we aim to:

- find a model for the market rates,
- find a model for the deposit volumes,
- calibrate all the parameters,
- find a way to segment customers,
- integrate this segmentation in the model.

We do not need a model for the deposit rates as they are assumed to be zero in our model.

All the objectives mentioned above can be summarized in a unique research question: ”How could the bank model demand deposits taking into account its specificities ?”

1.4. Practical implementation and data

The computations have been carried out in R. No codes or original data will be shown, all the data exposed in this thesis have been scaled for confidentiality reasons. However, it does not have any influence on the approach used.

The data used is extracted from internal software of the bank, that gives every month:

- the customers’ date of birth (YYYYMM),
- the actual date (YYYYMM),
- the year of account activation (YYYY),
- the semester of account activation (either 1 or 2),
- the level of wealth 4 months after the account activation,
- the actual accounting balance,

- the number of accounts that have the same features (same date of birth, same activation date, same level of wealth 4 months after the activation date).

As regards the historical period, we will use data from January 2010 to May 2020.

1.5. Outline

The thesis shall be presented as follows. The following chapter provides all the theoretical concepts needed for the model development. Chapter 3 will explain the methodology as well as the hypothesis of demand deposits modeling. The fourth chapter will discuss customer segmentation using a cluster analysis. The fifth chapter will interpret and discuss the results for the different proposals. Finally, the last chapter will conclude the thesis.

2

Theoretical concepts

This chapter intends to define and explain all the theoretical background needed for a good understanding of the model.

2.1. Spot rates

2.1.1. Definition

Definition 1. Continuously compounded mode. The continuously compounded interest rate $R(t,T)$ is defined as:

$$R(t,T) = -\frac{\ln P(t,T)}{T-t}. \tag{2.1}$$

Definition 2. Simple mode. The simple interest rate $L(t,T)$ is defined as:

$$L(t,T) = \frac{1-P(t,T)}{P(t,T)(T-t)}. \tag{2.2}$$

The interbank rates, such as the EURIBOR, are defined as in definition 2.

Definition 3. Yearly compounded mode. The yearly compounded interest rate $Y(t,T)$ is defined as:

$$Y(t,T) = \frac{1}{P(t,T)^{1/(T-t)}} - 1. \tag{2.3}$$

Definition 4. The short rate $r(t)$ is defined as the limit when T goes to t to a spot rate (either $R(t,T)$, $L(t,T)$ or $Y(t,T)$).

2.1.2. Euribor rates

The Euribor, that stands for EURo InterBank Offer Rate, is a spot reference rate constructed from the average interest rate at which eurozone banks offer unsecured short-term lending on the inter-bank market. It refers to a set of five money market rates, corresponding to five different maturities: the one-week, one-month, three-month, six-month and twelve-month rates. To compute these rates, quotes offered in the interbank market by a panel of banks, except the extreme values (15% highest and lowest rates) are gathered. In this thesis, the one-month rate (also known as BOR1m) will be used, both to calibrate the Hull-White one-factor model and the two-state deposit volume model.

The graph in figure 2.1 presents the evolution of the one-month rate over the last twenty years. As one can see, for a few years now, the one-month Euribor rates are negative. [1]

[1] https://www.euribor-rates.eu/en/

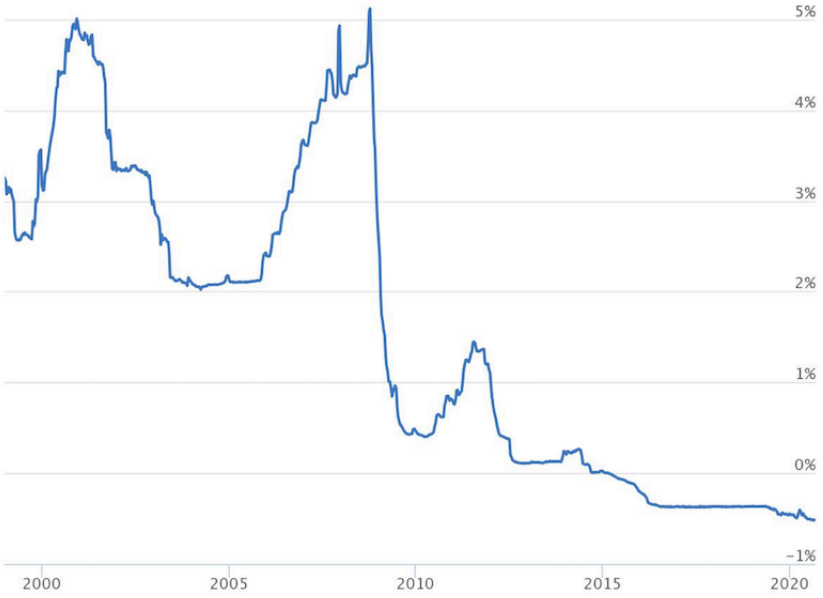

Figure 2.1: Evolution of the one-month rate from January 1999 to September 2020.

2.1.3. Yield curve

The yield curve illustrates interest rates (either R, L, and Y) at a fixed time t and for a range of maturities. This curve is constructed from bond prices observed on financial markets. The reference curve is the "zero-coupon" yield curve. Later on, this curve will be used as input in the market rates model. However, zero-coupon bonds are often notional bonds, so we compute the theoretical value of a zero-coupon from a coupon bond to construct the yield curve. Because of this, the yield curve can be less reliable and we will see later how to reconstruct the yield curve using polynomial interpolation.

2.2. Forward rates

A forward rate is an interest rate applicable to a financial transaction that will take place in the future. Similarly to the spot rates, there are three modes for computing a forward rates: either the continuously compounded mode, the simple one or the yearly compounded one. They all use the same property: the proceeds from investing at rate s_1 during the period $(0, t_1)$ and then reinvesting those proceeds at rate $f_{1,2}$ during the period (t_1, t_2) is equal to the proceeds from investing at rate s_2 during the period $(0, t_2)$. Note that s_1 and s_2 are spot rates and $f_{1,2}$ is the forward rate we want to compute.

Definition 5. Continuously compounded mode. The continuously compounded forward rate $f_R(t_1, t_2)$ in time period (t_1, t_2) is defined as:

$$f_R(t_1, t_2) = \frac{R(0, t_2)t_2 - R(0, t_1)t_1}{t_2 - t_1}. \tag{2.4}$$

Definition 6. Simple mode. The simple forward rate $f_R(t_1, t_2)$ in time period (t_1, t_2) is defined as:

$$f_L(t_1, t_2) = \frac{1}{t_2 - t_1}\left(\frac{1 + L(0, t_2)t_2}{1 + L(0, t_1)t_1} - 1\right). \tag{2.5}$$

Definition 7. Yearly compounded mode. The yearly compounded forward rate $f_Y(t_1, t_2)$ in time period (t_1, t_2) is defined as:

$$f_Y(t_1, t_2) = \left(\frac{(1 + Y(0, t_2))^{t_2}}{(1 + Y(0, t_1))^{t_1}}\right)^{\frac{1}{t_2 - t_1}} - 1. \tag{2.6}$$

2.3. Market interest rate models

The market interest rate is a key component in demand deposit modeling, as it is assumed to be an input in the deposit volume model as well as in the deposit rate model. Even though any deposit rate model will not be considered in the model, understanding the different market interest rate models that can be used is a top priority. In what follows, both the Vasicek model and the single factor Hull-White model will be presented.

2.3.1. Vasicek model

In 1977, Vasicek [18] introduced a short interest rate model where the interest rate curve is a function of a unique state variable: the short rate. This model was quite unique on its publication as it was the first one to capture mean-reversion. In its most simplified version, the dynamics of the short rate, under the risk-neutral measure **Q**, are described by an Itô process as follows:

$$\begin{cases} \mathrm{d}r(t) &= k(b - r_t)\mathrm{d}t + \sigma \mathrm{d}W_t \\ r(0) &= r_0, \end{cases} \tag{2.7}$$

where $W(t)$ is a Brownian motion, k is the speed of reversion, b is the long-term mean level and σ is the volatility.

Using Itô's lemma,

$$\begin{aligned} \mathrm{d}\big(\mathrm{e}^{kt} r(t)\big) &= k\,\mathrm{e}^{kt} r(t)\mathrm{d}t + \mathrm{e}^{kt}\,\mathrm{d}r(t) + 0\mathrm{d}t \\ &= kb\,\mathrm{e}^{kt}\,\mathrm{d}t + \sigma\,\mathrm{e}^{kt}\,\mathrm{d}W(t), \end{aligned} \tag{2.8}$$

so, $\forall t \geq 0$,

$$r(t) = r(0)\,\mathrm{e}^{-kt} + b(1 - \mathrm{e}^{-kt}) + \int_0^t \sigma\,\mathrm{e}^{-k(t-u)}\,\mathrm{d}W_u. \tag{2.9}$$

Starting from equation (2.9) above, both the expectation and the variance can be easily computed:

$$\begin{cases} \mathrm{E}(r(t)) &= r(0)\,\mathrm{e}^{-kt} + b\big(1 - \mathrm{e}^{-kt}\big), \\ \mathrm{Var}(r(t)) &= \frac{\sigma^2}{2k}\big(1 - \mathrm{e}^{-2kt}\big) \end{cases} \tag{2.10}$$

Proposition 1. *Zero-coupon bond price. The zero-coupon bond price at time t with maturity T can be written as:*

$$P(t,T) = A(t,T)\,\mathrm{e}^{-B(t,T)r(t)}, \tag{2.11}$$

where

$$\begin{cases} B(t,T) &= \frac{1-\mathrm{e}^{-k(T-t)}}{k} \\ A(t,T) &= \exp\Big[\big(b - \frac{\sigma^2}{2k^2}\big)\big(B(t,T) - T + t\big) - \frac{\sigma^2}{4k}B^2(t,T)\Big], \end{cases} \tag{2.12}$$

Proof. See Appendix A.1. □

From proposition 1, we can derive the continuously compounded interest rate $R(t,T)$.

Proposition 2. *Zero-coupon rates. The zero-coupon rate at time t is maturity T is of the following form*

$$R(t,T) = \big(b - \frac{\sigma^2}{2k^2}\big) + \frac{1 - \mathrm{e}^{-k(T-t)}}{k(T-t)}\big(r(t) - \big(b - \frac{\sigma^2}{2k^2}\big)\big) + \frac{\sigma^2}{4k^3}(1 - \mathrm{e}^{-k(T-t)})^2. \tag{2.13}$$

Proof. See Appendix A.2. □

A major drawback of this model is that it is inconsistent with the market-implied zero-coupon yield curve. As a matter of fact, the model is endogenous and we can see in equation (2.13) that the zero-coupon yield curve is fully defined with the model. So we cannot use market data.

2.3.2. Hull-White model

In 1990, J.C Hull and A. White introduced two extensions of the Vasicek [18] and of the Cox, Ingersoll and Ross [8] model, known as the Hull-White [11] one factor models. In what follows, we will only consider the extended Vasicek model, as it will be the one used in demand deposits modeling. It will be denoted as HW1f. Contrary to most older models, as Vasicek and Cox, Ingersoll and Ross, this model can provide a perfect fit to the initial term structure of interest rates. This attractive property, that will be proved, was first pioneered by Ho and Lee in 1986 [24].

When introduced, the HW1f model, as well as the Vasicek model, had the drawback of allowing interest rates to be negative. However, as stated earlier, negative interest rates are really popular nowadays. For instance, the Euribor rates are being negative for a few years now, because of the monetary policy of the European Central Bank. Then, the HW1f model is now very useful as it can deal with negative interest rates much better than others, as the Cox-Ingersoll-Ross model for instance.

This model assumes that the short rate is normally distributed and subject to mean reversion. In the most general form of the HW1f model, under the risk-neutral measure **Q**, the short rate follows the dynamics:

$$\mathrm{d}r(t) = [\theta(t) - a(t)r(t)]\mathrm{d}t + \sigma(t)\mathrm{d}W(t), \tag{2.14}$$

where θ, a and σ are deterministic functions and W is a Brownian motion. The three parameters θ, a and σ respectively represent the drift rate, the mean-reversion rate and the volatility factor.

In this master thesis, only θ will be considered to be time-dependent, while a and σ will remain constant. Let us now prove that this model can fit the initial term structure of interest rates. [9]

Proposition 3. *The model HW1f perfectly fits the market-implied zero-coupon yield curve if we set* $\theta(t) = \frac{\partial}{\partial t}F^M(0,t) + aF^M(0,t) + \frac{\sigma^2}{2a}(1 - \mathrm{e}^{-2at})$*, where* F^M *is the market-implied instantaneous forward rate.*

Proof. The dynamics of the short rate following a HW1f model are

$$\mathrm{d}r(t) = (\theta(t) - ar(t))\mathrm{d}t + \sigma\mathrm{d}W(t). \tag{2.15}$$

Using Itô's lemma, we have: $\mathrm{d}(r(t)\,\mathrm{e}^{at}) = ar(t)\,\mathrm{e}^{at}\,\mathrm{d}t + \mathrm{e}^{at}\,\mathrm{d}r(t) + 0\mathrm{d}t$.

Then,

$$\begin{aligned} \mathrm{d}(r(t)\,\mathrm{e}^{at}) &= ar(t)\,\mathrm{e}^{at}\,\mathrm{d}t + \mathrm{e}^{at}\,\mathrm{d}r(t) \\ &= ar(t)e^{at}\mathrm{d}t + \mathrm{e}^{at}\left[(\theta(t) - ar(t))\mathrm{d}t + \sigma\mathrm{d}W(t)\right] \\ &= \mathrm{e}^{at}\,\theta(t)\mathrm{d}t + \mathrm{e}^{at}\,\sigma\mathrm{d}W(t). \end{aligned} \tag{2.16}$$

Taking any $s \geq t$, $r(s)\,\mathrm{e}^{as} - r(t)\,\mathrm{e}^{at} = \sigma\int_t^s \mathrm{e}^{au}\,\mathrm{d}W_u + \int_t^s \mathrm{e}^{au}\,\theta(u)\mathrm{d}u$, so

$$r(s) = r(t)\,\mathrm{e}^{-a(s-t)} + \int_t^s \mathrm{e}^{-a(s-u)}\,\theta(u)\mathrm{d}u + \int_t^s \sigma\,\mathrm{e}^{-a(s-u)}\,\mathrm{d}W_u. \tag{2.17}$$

Then, the zero-coupon bond price with maturity T is defined by:

$$P(t,T) = \mathrm{E}\Big(\exp(-\int_t^T r(s)\mathrm{d}s)|\mathcal{F}_t\Big). \tag{2.18}$$

In particular,

$$\begin{aligned}\int_t^T r(s)\mathrm{d}s &= \int_t^T r(t)\,\mathrm{e}^{-a(s-t)}\,\mathrm{d}s + \int_t^T\int_t^s \mathrm{e}^{a(s-u)}\,\theta(u)\mathrm{d}s\mathrm{d}u + \int_t^T\int_t^s \sigma\,\mathrm{e}^{-a(s-u)}\,\mathrm{d}W_u\mathrm{d}s \\ &= r(t)\left[-\frac{1}{a}\,\mathrm{e}^{-a(s-t)}\right]_t^T + \int_t^T\left(\int_u^T \mathrm{e}^{-a(s-u)}\,\theta(u)\mathrm{d}s\right)\mathrm{d}u \\ &\quad + \int_t^T\left(\int_u^T \sigma\,\mathrm{e}^{-a(s-u)}\,\mathrm{d}s\right)\mathrm{d}W_u \text{ , by interchanging the two integrals} \\ &= \frac{r(t)}{a}\big(1-\mathrm{e}^{-a(T-t)}\big) + \int_t^T \frac{\theta(u)}{a}\big(1-\mathrm{e}^{-a(T-u)}\big)\mathrm{d}u + \int_t^T \frac{\sigma}{a}\big(1-\mathrm{e}^{-a(T-u)}\big)\mathrm{d}W_u \end{aligned} \quad (2.19)$$

The zero-coupon bond price can be written in the following form,

$$P(t,T) = A(t,T)\exp\big(-B(t,T)r(t)\big), \quad (2.20)$$

where $A(t,T)$ and $B(t,T)$ are two functions to be determined.
Now, let us set

$$B(t,T) = \frac{1}{a}\big(1-\mathrm{e}^{-a(T-t)}\big). \quad (2.21)$$

Then,

$$\begin{aligned}P(t,T) &= \mathrm{E}\Big(\exp\big(-r(t)B(t,T) - \int_t^T \theta(u)B(u,T)\mathrm{d}u - \int_t^T \sigma B(u,T)\mathrm{d}W_u\big)|\mathcal{F}_t\Big) \\ &= \exp\big(-r(t)B(t,T) - \int_t^T \theta(u)B(u,T)\mathrm{d}u\big)\mathrm{E}\Big(\exp\big(-\int_t^T \sigma B(u,T)\mathrm{d}W_u\big)|\mathcal{F}_t\Big) \\ &= \exp\big(-r(t)B(t,T) - \int_t^T \theta(u)B(u,T)\mathrm{d}u\big)\exp\big(\frac{1}{2}\int_t^T \sigma^2B^2(u,T)\mathrm{d}u\big), \\ &\quad \text{by Itô's isometry}\end{aligned} \quad (2.22)$$

Simultaneously,

$$\begin{aligned}\int_t^T \sigma^2B^2(u,T)\mathrm{d}u &= \int_t^T \frac{\sigma^2}{a^2}\big(1-\mathrm{e}^{-a(T-u)}\big)^2\mathrm{d}u \\ &= \frac{\sigma^2}{a^2}\left((T-t) - \frac{2}{a}\left[\mathrm{e}^{-a(T-u)}\right]_t^T + \frac{1}{2a}\left[\mathrm{e}^{-2a(T-u)}\right]_t^T\right) \\ &= \frac{\sigma^2}{a^2}\left(T-t-B(t,T) + \frac{1}{2a}\left(1-\mathrm{e}^{-2a(T-t)} - 2 + \mathrm{e}^{-a(T-t)}\right)\right) \\ &= \frac{\sigma^2}{a^2}\left(T-t-B(t,T)\right) - \frac{\sigma^2}{2a}B^2(t,T)\end{aligned} \quad (2.23)$$

Then, the zero-coupon bond price is given by

$$P(t,T) = A(t,T)\exp\Big(-B(t,T)r(t)\Big), \quad (2.24)$$

where

$$\begin{cases} A(t,T) &= \exp\Big(-\int_t^T \theta(u)B(u,T)\mathrm{d}u - \frac{\sigma^2}{2a^2}\big(T-t-B(t,T)\big) - \frac{\sigma^2}{4a}B^2(t,T)\Big), \\ B(t,T) &= \frac{1}{a}\Big(1-\mathrm{e}^{-a(T-t)}\Big). \end{cases} \tag{2.25}$$

Once the zero-coupon bond price is computed, one can derive the instantaneous forward rate as follows

$$\begin{aligned} F(0,T) &= -\frac{\partial}{\partial T}\ln P(0,T) && (2.26)\\ &= -\frac{\partial}{\partial T}\ln A(0,T) + \frac{\partial}{\partial T}\big(B(0,T)r(0)\big) \\ &= \int_0^T \theta(u)\frac{\partial}{\partial T}B(u,T)\mathrm{d}u + \frac{\sigma^2}{2a^2}\underbrace{\Big(\frac{\partial}{\partial T}B(0,T)-1\Big)}_{=-aB(0,T)} + \frac{\sigma^2}{4a}\Big(2B(0,T)\frac{\partial}{\partial T}B(0,T)\Big) \\ &\quad + r(0)\frac{\partial}{\partial T}B(0,T) \\ &= \int_0^T \theta(u)\frac{\partial}{\partial T}B(u,T)\mathrm{d}u + r(0)\frac{\partial}{\partial T}B(0,T) + \frac{\sigma^2}{2a}B(0,T)\Big(\frac{\partial}{\partial T}B(0,T)-1\Big), && (2.27) \end{aligned}$$

as well as the derivative of F with respect to T

$$\begin{aligned} \frac{\partial}{\partial T}F(0,T) &= \theta(T)\underbrace{\frac{\partial}{\partial T}B(T,T)}_{=1} + \int_0^T \theta(u)\frac{\partial^2}{\partial T^2}B(u,T)\mathrm{d}u + \frac{\sigma^2}{2a}\frac{\partial}{\partial T}B(0,T)\Big(\frac{\partial}{\partial T}B(0,T)-1\Big) \\ &\quad + \frac{\sigma^2}{2a}B(0,T)\frac{\partial^2}{\partial T^2}B(0,T) + \frac{\partial^2}{\partial T^2}B(0,T)r(0) \\ &= \theta(T) + \int_0^T \theta(u)\frac{\partial^2}{\partial T^2}B(u,T)\mathrm{d}u + \frac{\partial^2}{\partial T^2}B(0,T)r(0) \\ &\quad - \frac{\sigma^2}{2a}\Big(\frac{\partial}{\partial T}B(0,T) - \Big(\frac{\partial}{\partial T}B(0,T)\Big)^2 - \Big(\frac{\partial^2}{\partial T^2}B(0,T)\Big)B(0,T)\Big). && (2.28) \end{aligned}$$

(2.29)

$B(t,T)$ is already defined, so we can compute the first and second derivatives, giving

$$\frac{\partial}{\partial T}B(t,T) = \mathrm{e}^{-a(T-t)}, \tag{2.30}$$

$$\frac{\partial^2}{\partial T^2}B(t,T) = -a\,\mathrm{e}^{-a(T-t)}. \tag{2.31}$$

Then,

$$F(0,T) = \int_0^T \theta(u)\,\mathrm{e}^{-a(T-u)}\,\mathrm{d}u + r(0)\,\mathrm{e}^{-aT} - \frac{\sigma^2}{2a^2}\big(1-2\,\mathrm{e}^{-aT}+\mathrm{e}^{-2aT}\big) \tag{2.32}$$

$$\frac{\partial}{\partial T}F(0,T) = \theta(T) - a\int_0^T \theta(u)\,\mathrm{e}^{-a(T-u)}\,\mathrm{d}u - ar(0)\,\mathrm{e}^{-aT} + \frac{\sigma^2}{a}\Big(\mathrm{e}^{-aT}-\mathrm{e}^{-2aT}\Big) \tag{2.33}$$

Combining $F(0,T)$ and $\frac{\partial}{\partial T}F(0,T)$, we get that

$$aF(0,T) + \frac{\partial}{\partial T}F(0,T) = \theta(T) - \frac{\sigma^2}{2a}(1-\mathrm{e}^{-2aT}),$$

and finally

$$\theta(T) = aF(0,T) + \frac{\partial}{\partial T}F(0,T) + \frac{\sigma^2}{2a}(1 - \mathrm{e}^{-2aT}). \tag{2.34}$$

The model is calibrated on market data, if the instantaneous forward rate $F(0,T)$ and the market-implied instantaneous forward rate $F^M(0,T)$ are equal for every T. So, eventually,

$$\forall\, T \geq 0, \theta(T) = aF^M(0,T) + \frac{\partial}{\partial T}F^M(0,T) + \frac{\sigma^2}{2a}(1 - \mathrm{e}^{-2aT}) \tag{2.35}$$

□

The two remaining parameters a and σ are calibrated using historical market data (the one-month Euribor rates). As a matter of fact, we start by computing the empirical variances of the data set and of the series of differences between two consecutive values. Simultaneously, we compute the theoretical variances of $r(t+\Delta t) - r(t)$ and of $r(t)$, which leads us to a system of two equations with two unknowns.

Let us start by computing the two theoretical variances.

From equation (2.17), the short rate can be written as follows

$$r(t) = r(0)\,\mathrm{e}^{-at} + \int_0^t \mathrm{e}^{-a(t-u)}\,\theta(u)\mathrm{d}u + \int_0^t \sigma\,\mathrm{e}^{-a(t-u)}\,\mathrm{d}W_u. \tag{2.36}$$

As the first two terms are deterministic, the expectation and the variance of the short rate can be easily determined

$$\begin{aligned}
\mathrm{E}(r(t)) &= r(0)\,\mathrm{e}^{-at} + \int_0^t \mathrm{e}^{-a(t-u)}\,\theta(u)\mathrm{d}u + \underbrace{\mathrm{E}\Big(\int_0^t \sigma\,\mathrm{e}^{-a(t-u)}\,\mathrm{d}W_u\Big)}_{=0} \\
&= r(0)\,\mathrm{e}^{-at} + \int_0^t \mathrm{e}^{-a(t-u)}\,\theta(u)\mathrm{d}u \qquad (2.37) \\
\mathrm{Var}(r(t)) &= \sigma^2\mathrm{Var}\Big(\int_0^t \mathrm{e}^{-a(t-u)}\,\mathrm{d}W_u\Big) \\
&= \sigma^2\Bigg(\mathrm{E}\Big(\int_0^t \mathrm{e}^{-2a(t-u)}\,\mathrm{d}u\Big) - \underbrace{\mathrm{E}^2\Big(\int_0^t \mathrm{e}^{-a(t-u)}\,\mathrm{d}W_u\Big)}_{=0}\Bigg), \text{ by Ito isometry} \\
&= \frac{\sigma^2}{2a}\big(1 - \mathrm{e}^{-2at}\big) \qquad (2.38) \\
& \qquad (2.39)
\end{aligned}$$

So, the short rate clearly follows a normal distribution, where the expectation and the variance are explicitly given above.

Similarly, one can derive the theoretical variance of $r(t+\Delta t) - r(t)$. Using the Euler discretization,

$$r(t+\Delta t) - r(t) = (\theta(t) - ar(t))\Delta t + \sigma\sqrt{\Delta t}Z, \tag{2.40}$$

where $Z \sim \mathcal{N}(0,1)$.

Then,

$$\begin{aligned}\mathrm{Var}\big(r(t+\Delta t)-r(t)\big) &= a^2(\Delta t)^2\mathrm{Var}(r(t))+\sigma^2\Delta t\mathrm{Var}(Z)-2a(\Delta t)^{\frac{3}{2}}\sigma\big(\mathrm{E}(Zr(t))-\underbrace{\mathrm{E}(Z)}_{=0}\mathrm{E}(r(t))\big)\\ &= a^2(\Delta t)^2\mathrm{Var}(r(t))+\sigma^2\Delta t-2a(\Delta t)^{\frac{3}{2}}\sigma\mathrm{E}(Zr(t)) \qquad (2.41)\end{aligned}$$

Using the equation for the short rate

$$r(t)=r(0)\,\mathrm{e}^{-at}+\int_0^t \mathrm{e}^{-a(t-u)}\,\theta(u)\mathrm{d}u+\frac{\sigma}{\sqrt{2a}}\sqrt{1-\mathrm{e}^{-2at}}Z, \qquad (2.42)$$

where $Z\sim\mathcal{N}(0,1)$, $\mathrm{E}(Zr(t))$ can be computed as follows

$$\begin{aligned}\mathrm{E}(Zr(t)) &= \mathrm{E}\Big(\big(r(0)\,\mathrm{e}^{-at}+\int_0^t \mathrm{e}^{-a(t-u)}\,\theta(u)\mathrm{d}u\big)Z+\frac{\sigma}{\sqrt{2a}}\sqrt{1-\mathrm{e}^{-2at}}ZZ\Big)\\ &= \big(r(0)\,\mathrm{e}^{-at}+\int_0^t \mathrm{e}^{-a(t-u)}\,\theta(u)\mathrm{d}u\big)\underbrace{\mathrm{E}(Z)}_{=0}+\frac{\sigma}{\sqrt{2a}}\sqrt{1-\mathrm{e}^{-2at}}\underbrace{\mathrm{E}(Z^2)}_{=1}\\ &= \frac{\sigma}{\sqrt{2a}}\sqrt{1-\mathrm{e}^{-2at}} \qquad (2.43)\end{aligned}$$

If $t\longrightarrow+\infty$, then

$$\mathrm{Var}(r(t)) \longrightarrow \frac{\sigma^2}{2a} \qquad (2.44)$$

$$\mathrm{Var}(r(t+\Delta t)-r(t)) \longrightarrow \sigma^2\Delta t+\frac{\sigma^2a(\Delta t)^2}{2}-\sqrt{2a}\sigma^2(\Delta t)^{\frac{3}{2}} \qquad (2.45)$$

Let us denote by x the empirical variance of $r(t+\Delta t)-r(t)$ and by y the empirical variance of $r(t)$. Both of them are easily computed using any sofware.

The system of equations is then given as follows

$$\begin{cases} x &= \sigma^2\Big(\frac{(\Delta t)^2a}{2}+\Delta t-\sqrt{2a}(\Delta t)^{\frac{3}{2}}\Big),\\ y &= \frac{\sigma^2}{2a}\end{cases}$$

$$\Rightarrow\begin{cases} x &= \sigma^2\Big(\frac{(\Delta t)^2}{2}\frac{\sigma^2}{2y}+\Delta t-\sqrt{2\frac{\sigma^2}{2y}}(\Delta t)^{\frac{3}{2}}\Big),\\ a &= \frac{\sigma^2}{2y}\end{cases}$$

$$\Rightarrow\begin{cases} x &= \frac{(\Delta t)^2}{4y}\sigma^4+\Delta t\sigma^2-\frac{\Delta t^3}{y}^{\frac{1}{2}}\sigma^3,\\ a &= \frac{\sigma^2}{2y}\end{cases}$$

$$\Rightarrow\begin{cases} 0 &= \frac{(\Delta t)^2}{4y}\sigma^4-\frac{\Delta t^3}{y}^{\frac{1}{2}}\sigma^3+\Delta t\sigma^2-x,\\ a &= \frac{\sigma^2}{2y}\end{cases} \qquad (2.46)$$

The first equation of the system can be solved using any software. Among the four possible values that satisfy the equation, the complex numbers are left out. Then, among the remaining values for σ, we will keep the one that lets a be positive around 2-5 %. As a matter of fact, it is a typical value and a is always of that order.

2.4. Clustering for customer segmentation

Customer segmentation consists of analysing customers' behavior and then dividing them into small groups that have similarities. This can be done using clustering in machine learning. The logic goes like this: one needs to find the smallest variations among customers within each group, and the largest variations within each different group. Common cluster analysis methods are the k-means algorithm [12] and the hierarchical cluster analysis [19]. Studying the complexity of each method to pick the right one can be extremely important while dealing with a large data set. However, in this thesis, only 100 observations will be considered for the clustering, so the complexity of each method should not be a criterion for the selection.

2.4.1. K-means algorithm

The K-means algorithm is an unsupervised clustering algorithm that gathers the initial data set into K number of clusters. Choosing the ideal number of clusters K is a distinct step that can be done using the Elbow method [15] or the gap statistic method [21]. Both of the two methods are explained in subsection 2.4.3. It will be assumed here that the number K is already picked. Then, given the data set $X_1, X_2, X_3, \ldots X_n$, the algorithm works as follows [12]:

1. Pick K points as the initial centroids from the data set. This step can be done either randomly or using any selection method.
2. Calculate the Euclidean distance of each point X_i in the data with the identified cluster centroids.

   Definition 8. Given two points $\mathbf{p} = (p_1, p_2, \ldots, p_n)$ and $\mathbf{q} = (q_1, q_2, \ldots, q_n)$, the Euclidean distance between $\mathbf{p}$ and $\mathbf{q}$ is equal to

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

3. Assign each data point X_i to the closest centroid using the distance found in step 2.
4. Determine the new centroid by taking the average of the points in each cluster group.
5. Repeat steps 2 to 4 until the centroids do not change.

The K-means algorithm is really simple to implement.

2.4.2. Hierarchical cluster analysis

The hierarchical clustering [19] is an algorithm that builds a hierarchy of clusters. There are two types of strategies, either the agglomerative one or the divisive one. This method can be visualised using a dendrogram, a dendrogram being a tree-like diagram that records the sequences of merges and splits.

The divisive algorithm is a top down approach that works as follows. All the data sets form at the beginning one cluster, and recursively, they are split into two clusters, and so on, until each observation is in its own cluster. On the contrary, the agglomerative method is a bottom up approach that operates in an inverted way. Initially, each observation represents its own cluster. The clusters are then recursively merged by similarity until one unique cluster is formed, or until K clusters are formed if we want a specific number of clusters.

Before starting to split or merge two clusters, the metric and the linkage criterion used have to be chosen. Each method has pros and cons.

Metric

The chosen metric calculates, at each step, the distance between two points and puts all the distances in a similarity matrix. Common metrics are the Euclidean distance, the Manhattan distance and the maximum distance. The definition of the Euclidean distance has already been given in definition 8.

Definition 9. Manhattan distance. Given two points $\mathbf{p} = (p_1, p_2, ..., p_m)$ and $\mathbf{q} = (q_1, q_2, ..., q_m)$, the Manhattan distance between $\mathbf{p}$ and $\mathbf{q}$ is equal to

$$d(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} |q_i - p_i|. \tag{2.47}$$

Definition 10. Maximum distance. Given two points $\mathbf{p} = (p_1, p_2, ..., p_m)$ and $\mathbf{q} = (q_1, q_2, ..., q_m)$, the maximum distance between $\mathbf{p}$ and $\mathbf{q}$ is equal to

$$d(\mathbf{p}, \mathbf{q}) = \max_i |q_i - p_i|. \tag{2.48}$$

Linkage criterion

Once the similarity matrix is computed, the chosen linkage criterion determines how the step of merging or splitting will be done. Common linkage methods are the single-linkage algorithm, the complete linkage algorithm and Ward's method [13].

Definition 11. Similarity/distance between two clusters with the single-linkage algorithm. The similarity S between two clusters C_1 and C_2 is here defined as follows:

$$S(C_1, C_2) = \min_{\mathbf{p_i} \in C_1, \mathbf{p_j} \in C_2} d(\mathbf{p_i}, \mathbf{p_j}). \tag{2.49}$$

Definition 12. Similarity/distance between two clusters with the complete-linkage algorithm. The similarity S between two clusters C_1 and C_2 is then defined as follows:

$$S(C_1, C_2) = \max_{\mathbf{p_i} \in C_1, \mathbf{p_j} \in C_2} d(\mathbf{p_i}, \mathbf{p_j}). \tag{2.50}$$

Definition 13. Similarity/distance between two clusters with Ward's method. This method was introduced by Ward [13] in 1963. The similarity S between two clusters C_1 and C_2 is now defined as follows:

$$S(C_1, C_2) = \sum_{\mathbf{p_i} \in C_1, \mathbf{p_j} \in C_2} \frac{d^2(\mathbf{p_i}, \mathbf{p_j})}{Card(C_1) Card(C_2)}, \tag{2.51}$$

where $Card(C_k)$ is the cardinality of cluster k.

2.4.3. Optimal number of clusters

Several methods can be used to determine the optimal number of clusters. Although this can be determined nearly immediately on any programming software, understanding the theory behind is quite important. We will present here the Elbow method [15] and the gap statistic method [21].

The Elbow method is a heuristic method that was apparently firstly described by Thorndike in 1953. This method is the easiest way to answer the following question: "Into how many families should the specimens be grouped ?" [15]. Intuitively, the greater K we pick, the less average distance we will have within families. However, at some point, choosing any value greater than K will not reduce the average distance within families significantly. To determine this optimal number of clusters K, we compute the K-means algorithm for a range of K. Typical range is to try K from 1 to 10, but this is just arbitrary. Then, for every K in the range, we perform a cluster analysis using the K-means method and we compute a clustering score. The most commonly used "clustering score" is the total within-cluster sum of square (WCSS). $WCSS$ can be defined as

$$WCSS = \sum_{i=1}^{K} \left(\sum_{p_j \in C_i} d(\mathbf{p_j}, \mathbf{c_i}) \right), \tag{2.52}$$

where C_i refers to the cluster i, $\mathbf{c_i}$ is the centroid of cluster i and $d(\mathbf{x_i}, \mathbf{x_j})$ is the Euclidean distance between the two points x_i and x_j.

Then, we plot the clustering score, the $WCSS$ as a function of the number of clusters K. Eventually, the optimal K will be the value where the $WCSS$ falls rapidly before reaching a plateau. In other words, we pick K at the elbow. However, such an elbow does not clearly appear sometimes and it can be quite confusing.

The Elbow method is quite simple but maybe too subjective in some cases. Then, the gap statistic method [21] is an alternative method to determine the optimal number of clusters. This method can be applied to both clustering algorithm, the K-means and the hierarchical clustering. As for the Elbow method, we need to decide the range of K to test. The idea under the gap statistic method is to compare the $\log$ of the pooled within-cluster sum of squares w_k with their expected values under an appropriate null reference distribution of the data (i.e a distribution with no obvious cluster) for different values of K. We define w_k as

$$w_k = \sum_{i=1}^{k} \frac{1}{2n_i} \Big(\sum_{p_j, p_{j'} \in C_i} d(\mathbf{p_i}, \mathbf{p_{i'}}) \Big), \tag{2.53}$$

where $n_i = Card(C_i)$ is the cardinality of cluster i, $d(\mathbf{p_i}, \mathbf{p_{i'}})$ is the Euclidean distance between the points p_j and $p_{j'}$.

Then, we compute the gap statistic for every k as:

$$Gap_n(k) = \mathrm{E}_n^* \big(\log(w_k) \big) - \log(w_k), \tag{2.54}$$

where E_n^* is the expectation under a sample size n from the reference distribution and n is the size of the data set.

To estimate $\mathrm{E}_n^* \big(\log(w_k) \big)$, we generate B copies of the reference distribution and we compute the average of the generated copies. We usually choose B around 500, as it is the default value in the software R. Eventually, the optimal number of clusters K will be the value that maximizes $Gap_n(k)$ after taking the sampling distribution into account. However, the concept of maximizing can be interpreted in different ways. Tibshirani et al. [21] proposed to pick K as follows:

1. Assuming that we want to test k from 1 to K_{max}, we cluster the observed data, using any clustering algorithm, for every k from 1 to K_{max} and we compute w_k.

2. We generate B reference data sets, we cluster all of them, for k varying from 1 to K_{max}, we compute w_k^b for $k \in (1, K_{max})$ and $b \in (1, B)$. Then, for every k, we estimate the gap statistic as in (2.54)

$$Gap(k) = \frac{1}{B} \sum_{b=1}^{B} \big(\log(w_k^b) \big) - \log(w_K) \tag{2.55}$$

3. Setting $\bar{w} = \frac{1}{B} \sum_{b=1}^{B} \big(\log(w_k^b) \big)$, we compute the standard deviation as:

$$sd_k = \sqrt{\frac{1}{B} \sum_{b=1}^{B} \big(\log(w_k^b) - \bar{w} \big)^2} \tag{2.56}$$

4. Finally, we will pick the smallest K such that

$$Gap(k) \geq (Gap(k+1) - s_{k+1}), \tag{2.57}$$

where $s_{k+1} = sd_k \sqrt{1 + \frac{1}{B}}$

Other ways to pick the optimal K using the gap statistic method is to choose the first local maximum, or the global maximum. We will only consider the approach proposed by Tibshirani et al. [21] in Chapter 5.

2.5. Dynamics of deposit volume

The dynamics of deposit volume need to be simulated in a static scenario. However, the deposit volume model, as constructed, simulates the evolution of the average accounting balance in a dynamic environment. The issue is then to move from a dynamic scenario, where a customer can add or withdraw money on his account, to a static scenario, where a customer can only withdraw money on his account. For this reason, only a decrease in deposit volume is taken into consideration. Then, two methods are used to estimate a deposit volume decrease. For confidentiality reasons, those methods will only by named as ”Method 1” and ”Method 2”. These are defined below.

Definition 14. Method 1. The technique of this method works as follows. To construct the adjusted path, at every time step t, we will keep the initial value at time t only if it less than the lowest point reached at any time step before t (excluding t). Otherwise, the adjusted value at time t will be the lowest point reached at any time step before t. Once the adjusted path is constructed, we also need to take into account the attrition rate. Mathematically speaking, the adjusted average accounting balance held on an arbitrary account at time $t+1$ is:

$$\begin{cases} D^{M1}(t+1) &= \min(D^{M1}(t); D(t+1)) \cdot (1 - r^A(t+1)), \\ D^{M1}(0) &= D(0), \end{cases} \tag{2.58}$$

where $D^{M1}(t)$ is the deposit volume at time t with the application of the first method, $D(i)$ is the deposit volume at time t of the initial path, $r^A(t)$ is the attrition rate at time t. We recall that the attrition rate at time $t+1$ is the ratio of the number of accounts closed between time t and time $t+1$ and the number of accounts at time t.

Definition 15. Method 2. This second method is rather simple mathematically speaking but maybe less justifiable from a theoretical point of way. Taking any arbitrary path, we want to construct an adjusted path where we apply only the decreases. In case of a rise between $t-1$ and t, the adjusted value at time t will be the adjusted value at time $t-1$. In brief, the adjusted average accounting balance held on an arbitrary account at time $t+1$ is:

$$\begin{cases} D^{M2}(t+1) &= D^{M2}(t) \cdot \min\left(1; \frac{D(t+1)}{D(t)}\right), \\ D^{M2}(0) &= D(0), \end{cases} \tag{2.59}$$

where $D^{M2}(t)$ is the deposit volume at time t with the application of the second method and $D(t)$ is the deposit volume at time t of the initial path.

In the figure 2.2, one arbitrary initial path and the effect of the two methods are represented.

As we can see in figure 2.2, the method 2 is much more discriminant than method 1, that is more conservative. However, both the evolution of the average accounting balances and the attrition rates are needed in the method 1, while the method 2 only requires data on the average accounting balances. Applying the method 1 requires additional analysis that will not be done in this thesis. For this reason, only the method 2 will be applied to move from a dynamic scenario to a static scenario later on.

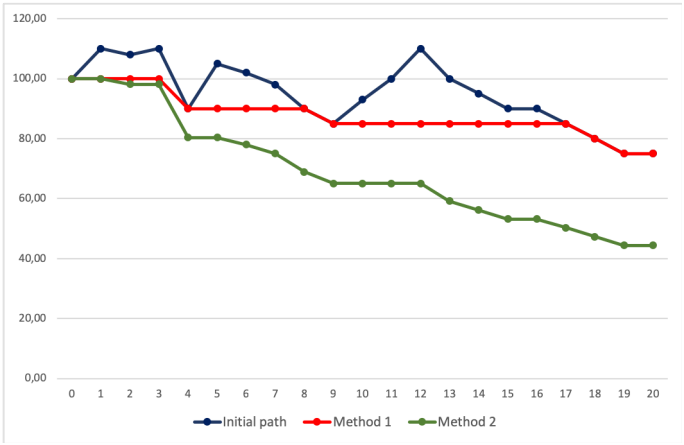

Figure 2.2: Application of two methods on an arbitrary path to move from a dynamic scenario to a static scenario

2.6. Specific definitions for the output of the model

2.6.1. Linear adjustment 15 years

A linear adjustment 15 years is an adjustment that can be made on every path so that they will all reach 0 before or exactly after fifteen years. This adjustment will be done after the deposit volume model simulation for several reasons. Hedging instruments are usually much less liquid after 15 years. Also, the time period used to calibrate the parameters is usually less than 15 years. In this thesis, it will be only 10 years, so forecasting deposit volume after 15 years might too be ambitious and too risky.

Let $D(t)$ be the deposit volume at time t of an arbitrary account. Then, the deposit volume at time t on this account after the linear adjustment, denoted as $D^{LA}(t)$, is as follows:

$$D^{LA}(t) = D(t) - D(15) \cdot \left(\frac{t}{15}\right)^2, \tag{2.60}$$

where t is the time in years and $D(15)$ is the deposit volume at 15 years for this arbitrary account.

The effect of this adjustment is presented in the graph in figure 2.3.

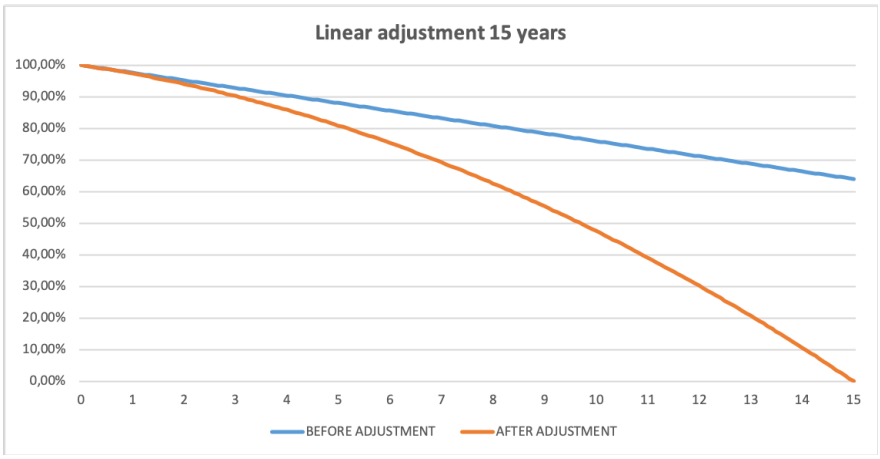

Figure 2.3: Example of a linear adjustment 15 y on an arbitrary path

2.6.2. Average life

The average life is the average effective maturity of the deposits. For a given account i, the average life AL^i will be calculated as follows in the model:

$$AL^i = \frac{\sum_{t=1}^{n} \text{Age}^i(t) \cdot (AB^i(t) - AB^i(t-1))}{AB^i(0)}, \tag{2.61}$$

where $\text{Age}^i(t)$ and $AB^i(t)$ are respectively the age (in years) of account i and the accounting balance of account i at time t.

Eventually, the two essential output parameters will be both the average life after adjustments and the optionality, where the optionality corresponds to the ratio between the optional component of the stable balance and the stable balance.

2.7. Polynomial interpolation

Polynomial interpolation will be used in the market rates model to reconstruct the zero-coupon yield curve. Both the cubic interpolation and the square interpolation are presented below.

2.7.1. Cubic Interpolation

A cubic interpolation builds a set of third-degree polynomials that ensure the continuity of the function and its first derivative at each point x_i. Assume that we have 4 points x_i, and their respective values y_i $\forall i \in [1, 4]$, with $x_1 < x_2 < x_3 < x_4$.

Assume that we want to estimate the value of a fifth point x using a cubic interpolation, such that $x \in [x_2, x_3]$.

First, the approximate value of the function f at the point x is given by

$$f(x) \simeq ax^3 + bx^2 + cx + d. \tag{2.62}$$

Let us denote g the interpolating function. To find the coefficients of the polynomial g, one needs to solve the following matrix equation $AX = Y$:

$$\begin{bmatrix} x_2^3 & x_2^2 & x_2 & 1 \\ x_3^3 & x_3^2 & x_3 & 1 \\ 3x_2^2 & 2x_2 & 1 & 0 \\ 3x_3^2 & 2x_3 & 1 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} y_2 \\ y_3 \\ 0.5\left(\frac{y_3 - y_2}{x_3 - x_2} + \frac{y_2 - y_1}{x_2 - x_1}\right) \\ 0.5\left(\frac{y_4 - y_3}{x_4 - x_3} + \frac{y_3 - y_2}{x_3 - x_2}\right) \end{bmatrix}. \tag{2.63}$$

In other words:

- the interpolating function should pass through the two points (x_2, y_2) and (x_3, y_3).
- the derivative at the point x_2 (respectively x_3) is equal to the average of the slopes between the points (x_1, y_1) and (x_2, y_2) (respectively (x_2, y_2) and (x_3, y_3)), and between the points (x_2, y_2) and (x_3, y_3) (respectively (x_3, y_3) and (x_4, y_4)).

By inverting the matrix A, we obtain the four values of the coefficients and the value $g(x)$ can directly be computed.

The cubic interpolation will be used to reconstruct the zero-coupon yield curve based on specific points of the curve. However, for some points, we will only have 3 points surrounding x, such that $x_1 < x_2 < x < x_3$. Then, we will not use a cubic interpolation anymore but it will be a square interpolation.

2.7.2. Square interpolation

A square interpolation builds a set of second-degree polynomials that ensure the continuity of the function and its first derivative at each point x_i. Assume that we have 3 points x_i, and their respective

values y_i, $\forall i \in [1,3]$, with $x_1 < x_2 < x_3$.

Assume that we want to estimate the value of a fourth point x using a square interpolation, such that $x \in [x_2, x_3]$.

First, the approximate value of the function f at the point x is given by

$$f(x) \simeq ax^2 + bx + c. \tag{2.64}$$

Let us denote g the interpolating function. To find the coefficients of the polynomial g, one needs to solve the following matrix equation $AX = Y$:

$$\begin{bmatrix} x_2^2 & x_2 & 1 \\ x_3^2 & x_3 & 1 \\ 2x_2 & 1 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} y_2 \\ y_3 \\ 0.5\left(\dfrac{y_3 - y_2}{x_3 - x_2} + \dfrac{y_2 - y_1}{x_2 - x_1}\right) \end{bmatrix}. \tag{2.65}$$

Similarly to a cubic interpolation, we invert the matrix A to obtain the three coefficients of g and then we can compute the value $g(x)$.

3

The foundations of the model

In this thesis, demand deposits modeling will be a three-step process. We will then haolume one model for the market rates, one for the deposit volumes and one for the deposit rates. Demand deposits are not remunerated by the bank so the deposit rates are always zero. Therefore, this section will only focus on market rate modeling and deposit volume modeling. We will start by the market rate model, as it will be used as input data for the deposit volume model.

3.1. Market rate model

3.1.1. Model

The first step of demand deposits modeling is to find a stochastic model for the evolution of future interest rates. Common models are the Vasicek model [18], the Cox-Ingersoll-Ross model [8], the Hull-White model [11] or the no-arbitrage models of HJM [10]. Several constraints are taken into consideration for the choice of this model:

- For a few years now, negative rates became commonplace on the European market. So even though this characteristic was fiercely criticized in the past, especially in the Vasicek model, the chosen model should allow negative rates. Log-normal models or the Cox-Ingersoll-Ross model [8] are then excluded.
- The model has to fit exactly today's term structure of interest rates. As seen in Chapter 2, the Vasicek model is inconsistent with market data and is excluded.

Finally, a single factor Hull-White model has been chosen to model the market rates. This model is also used by Jarrow and Van Deventer in 1998 in their demand deposits model [20]. Additional benefit of this model is that it does not allow arbitrage. We did not chose a Hull-White two factor model, as it adds much more complexity regarding the computation of the model. As a matter of fact, the challenge of the thesis lies in the integration of one or two variables to reflect the wide disparity between customers that appeared the last few years. There is therefore no need to add complexity in the market rate model.

The dynamics of a single factor Hull-White model, under the risk-neutral measure **Q**, follows this stochastic differential equation:

$$\mathrm{d}r(t) = (\theta(t) - a_{HW} r(t))\mathrm{d}t + \sigma_{HW}\mathrm{d}W_{HW}(t) \tag{3.1}$$

where θ, a_{HW} and σ_{HW} respectively represent the drift rate, the mean-reversion rate and the volatility factor and where $W_{HW}(t)$ is a Brownian motion. Contrary to Chapter 2, the characters *"HW"* are precised in the parameters' names so that there is no confusion with futures variables. Only θ is time-dependent, and is explicitly given below:

$$\theta(t) = \frac{\partial}{\partial t}F^M(0,t) + a_{HW}F^M(0,t) + \frac{\sigma_{HW}^2}{2a_{HW}}(1 - \mathrm{e}^{-2a_{HW}t}), \tag{3.2}$$

where F^M is the market-implied instantaneous forward rate.

Once the formula for the function θ is explicitly given, we aim to compute the value of $\theta(t)$ for every t over the horizon of simulation (every month from 0y to 40y). To do so, we need the value of the parameters a_{HW} and σ_{HW} as well as the market-implied instantaneous forward rates. But we only have the zero-coupon yield curve as input data, and points on the yield curve are not all reliable. This problem needs to be addressed upstream and then, the market-implied instantaneous forward rates can be calculated from the zero-coupon rates.

Reconstruction of the zero-coupon yield curve

The zero-coupon yield curve, that serves as input data for the market rate model, is extracted from the software Fusion Risk [2]. This software uses all the monthly data of zero-coupon swap rates to construct the yield curve. However, some points, as the 1m, 6m, 1y, 10y, and others, are more reliable points, because there are more trades on these points and then their prices are closer to reality. The reliable points will be stored and the remaining points will be computed using either a square interpolation or a cubic interpolation.

For every t that goes from 0y to 40y with a monthly step, the logic works as follows:

- If the zero-coupon rate at time t is considered to be a reliable point, then this rate does not change.
- If the zero-coupon rate at time t is not reliable but we can find two reliable points before t and two reliable points after t (if there are more than two, then the two closest points are taken in each case), then the zero-coupon rate at time t is reconstructed using a cubic interpolation.
- If the zero-coupon rate at time t is not reliable but we can find two reliable points before t (if there are more than two, then the two closest points are taken) and one reliable point after t, then the zero-coupon rate at time t is reconstructed using a square interpolation.

So in any case, we can easily reconstruct a more reliable zero-coupon yield curve.

Calculation of the market-implied instantaneous forward rates

Once the zero-coupon yield curve is rebuilt, the forward rate can be computed. Definitions of the forward rate are already given in Chapter 2.

All over the horizon of simulation, we have a monthly time step, so $\Delta t = t_2 - t_1$ will remain between $\frac{28}{365} \simeq 0.0767$ and $\frac{31}{365} \simeq 0.0849$. The instantaneous forward rate is defined as the limit when Δt goes to zero to the forward rate. As Δt is quite small and close to 0, with a view to simplification, the instantaneous forward rate will be approximated by the forward rate, using the continuously compounded mode.

$\forall i \in [1, 480]$, as we have monthly observations from 0 years to 40 years, the time value in years is $t_i = \frac{i}{12}$. Then, we will have the following approximation

$$F^M(0, t_{i-1}) \simeq \frac{r_{t_i}t_i - r_{t_{i-1}}t_{i-1}}{t_i - t_{i-1}}, \tag{3.3}$$

$$\frac{\partial}{\partial t}F^M(0, t_{i-1}) \simeq \frac{F^M(0, t_i) - F^M(0, t_{i-1})}{t_i - t_{i-1}}. \tag{3.4}$$

The last step to compute $\theta(t)$ for every t over the horizon of simulation is to calibrate the mean-reversion and the volatility parameters.

Calibration of the mean-reversion parameter and of the volatility

There are different ways to calibrate the mean-reversion and the volatility in the Hull-White model. In this thesis, a_{HW} and σ_{HW} will be calibrated using historical Euribor rates.

The following approach is used:

- The short-term rates $r(t)$ are approximated by the one-month Euribor rates.
- The empirical variances of the series $r(t)$ and $r(t+\Delta t)-r(t)$ are computed using any calculation software.
- a_{HW} and σ_{HW} are then determined so that the long-term variances of the two series defined by the model and the empirical variances measured are equal.

The calculation details are described in subsection 2.3.2.

3.1.2. Simulation of the market rates

Once the three parameters a_{HW}, σ_{HW} and θ are calculated, we compare three different Monte-Carlo simulation methods, the first one using an Euler discretization, the second one using an Euler semi-discretization and the last one being an exact simulation. The three methods are explained in [22]. Eventually, as one can see in equations 3.5 and 3.6, we will not need to calculate the θ function but only the market-implied instantaneous forward rates. We could also have tested a Milstein scheme [17], that has a superior strong order of convergence than the Euler scheme, but we will see after that the two schemes - Euler and Milstein, are the same in this particular case.

In the book from Brigo and Mercurio [9], the short-rate is written as a sum of a deterministic term and of a stochastic term:

$$r(t) = x(t) + \alpha(t), \tag{3.5}$$

where

$$\begin{aligned} \alpha(t) &= F^M(0,t) + \frac{\sigma^2}{2a_{HW}}(1-\mathrm{e}^{-2a_{HW}t})^2, && (3.6)\\ \mathrm{d}x(t) &= -a_{HW}x(t)\mathrm{d}t + \sigma_{HW}\mathrm{d}W(t), && (3.7)\\ x(0) &= 0 && (3.8) \end{aligned}$$

Using Ito's lemma, the stochastic term $x(t)$ is given by:

$$\mathrm{d}(e^{a_{HW}t}x(t)) = a_{HW}e^{a_{HW}t}x(t)\mathrm{d}t + e^{a_{HW}t}\mathrm{d}W(t),$$

So for any $s \geq t$,

$$x(s) = \mathrm{e}^{-a_{HW}(s-t)}\,x(t) + \int_t^s \sigma\,\mathrm{e}^{-a_{HW}(s-u)}\,\mathrm{d}W(u). \tag{3.9}$$

From equation (3.9), the distribution of $x(s)$ conditionally to $x(t)$ follows a normal law, where the conditional expectation and variance are given below:

$$\begin{aligned} \mathrm{E}(x(s)|x(t)) &= x(t)\,\mathrm{e}^{-a_{HW}(s-t)}, && (3.10)\\ \mathrm{Var}(x(s)|x(t)) &= \frac{\sigma_{HW}^2}{2a_{HW}}(1-\mathrm{e}^{-2a_{HW}(s-t)}\,. && (3.11) \end{aligned}$$

The process $\alpha(t)$ is deterministic so it can be exactly simulated for the given time grid $t = t_0, \ldots, t_n$, assuming we have $n+1$ data points. As for the stochastic process, we will use three different equations:

1. Full Euler discretization scheme:

$$x(t+\Delta t) = (1 - a_{HW}\Delta t)x(t) + \sigma_{HW}\sqrt{\Delta t}\mathcal{N}(0,1). \tag{3.12}$$

2. Semi Euler discretization scheme:

$$x(t+\Delta t) = \mathrm{e}^{-a_{HW}\Delta t}\, x(t) + \frac{\sigma_{HW}}{\sqrt{2a_{HW}}}\sqrt{1-\mathrm{e}^{-2a\Delta t}}\mathcal{N}(0,1). \tag{3.13}$$

3. Exact simulation:

$$x(t)|x(0) \sim \mathcal{N}(x(0)\,\mathrm{e}^{-a_{HW}t}, \frac{\sigma_{HW}^2}{2a}(1-\mathrm{e}^{-2a_{HW}t})), \tag{3.14}$$

   where $x(0) = 0$.

The yield curve given as input in the market rate model has monthly data from 0 to 40y, and similarly, the historical data of the bank are also given on a monthly basis. Then, immediately, we have to choose a monthly time step, so $\Delta t = \frac{1}{12}$. Regarding the number of paths, we choose $N = 10^4$. Results regarding the simulation run-times and the error are presented in Chapter 5. Given that $r(t) = x(t) + \alpha(t)$, the error term for the first two methods is defined as:

$$\epsilon = \frac{1}{N}\sum_{i=1}^{N} |r_i(T) - \tilde{r}_i(T)|, \tag{3.15}$$

where $\tilde{r}_i(T)$ is the approximation of the short rate, using one of the three methods.

Finally, we can estimate $r(t)$ by computing the expectation of all the simulated paths. Both the 10^4 paths and the average trajectory will served as input data in the deposit volume model.

3.1.3. Summary of the market rate model

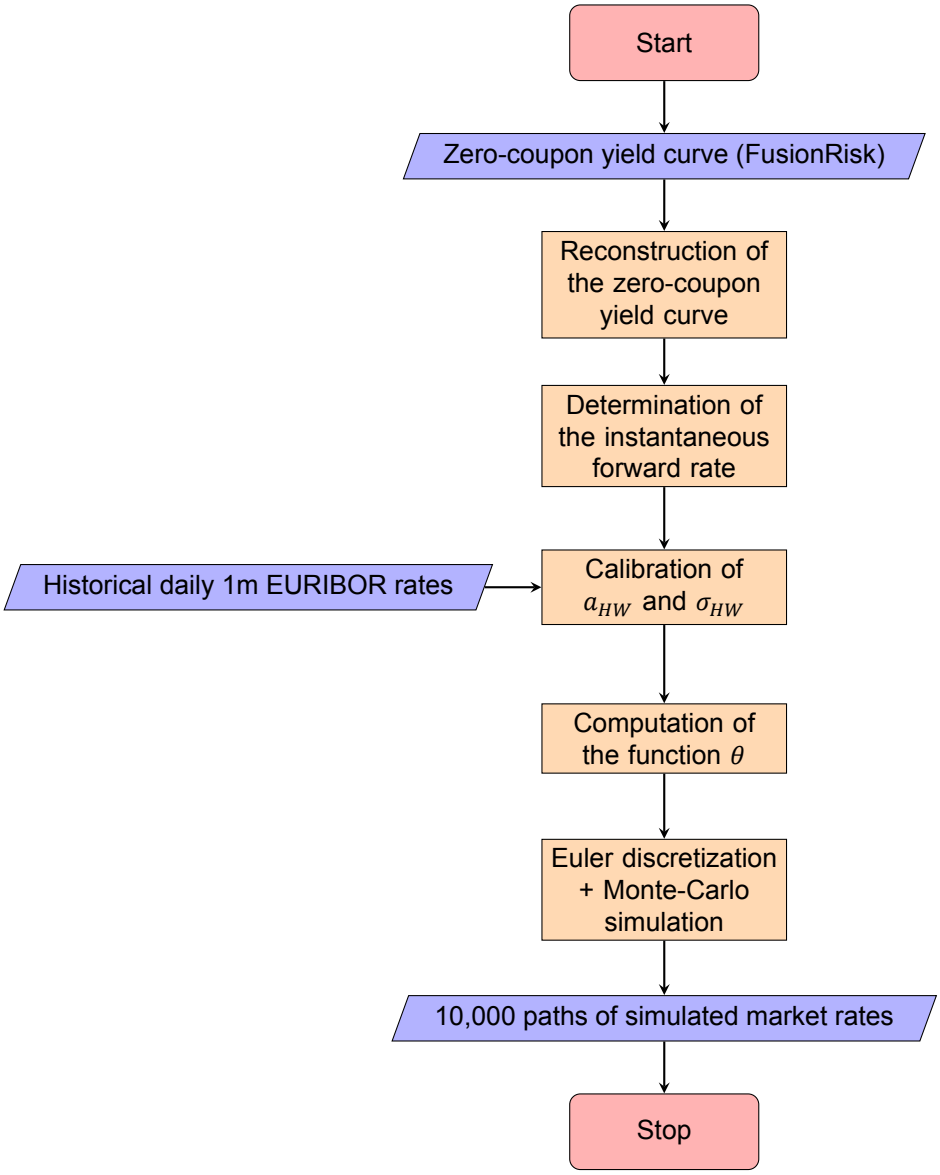

Figure 3.1: Flowchart of the market rate model

3.2. Deposit volume model

3.2.1. Model

The deposit volume model is more complex than the market rate model but is based on a very simple idea: the deposit volume fluctuates around a trend $\tilde{D}(t)$. In other words, the dynamics of deposit volume follow an Ornstein-Uhlenbeck process, where the drift term is not really a drift term but a trend that varies over time. This trend $\tilde{D}(t)$ evolves differently depending on whether we are in state 1 or in state 2. The probability of belonging to one state or another depends on the market rate $r(t)$, that have been computed in section 3.1. State 1 corresponds to a normal-rate environment and state 2 to a low-rate environment. Mathematically speaking, the model can be written as follows:

$$\begin{aligned}
\mathrm{d}D(t) &= k(\tilde{D}(t) - D(t))\mathrm{d}t + \sigma \mathrm{d}W(t), && (3.16)\\
\tilde{D}(t+\Delta t) &= \begin{cases} \tilde{D}(t) + b \cdot \Delta t & \text{if } S_{t;t+\Delta t} = 1, \\ \tilde{D}(t) + b_2 \cdot \Delta t & \text{if } S_{t;t+\Delta t} = 2, \end{cases} && (3.17)\\
\tilde{D}(t_0) &= a + bP_1(t_0) + b_2 P_2(t_0) && (3.18)\\
P(S_{t;t+\Delta t} = 1) &= P_1(t) = \frac{1}{1 + \mathrm{e}^{-d(r(t)-s)}}, && (3.19)\\
P(S_{t;t+\Delta t} = 2) &= P_2(t) = 1 - P(S_{t;t+\Delta t} = 1), && (3.20)
\end{aligned}$$

where k, σ, d and s are deterministic parameters, a, b and b_2 are respectively the intercept, the long-term slope in a normal-rates environment (state 1) and the slope in a low-rates environment (state 2). The equation $S_{t;t+\Delta t} = i$ means that we are state in state i (either 1 or 2) in time period $(t, t + \Delta t)$.

With a view to simplification, we will here calibrate parameters on a monthly scale, and as we have monthly data, immediately $\Delta t = 1$. Δt will be sometimes left out, however, moving from a monthly scale to yearly scale for parameters can be done really quickly.

3.2.2. Calibration of d and s

The probability of being either in state one or two depends on market rates and follows a logit model. Then, to calibrate d and s, we will use historical data of the national central bank. This choice was made because of the loss of homogeneity among customers in the bank during recent years. As a matter of fact, if we choose to calibrate d and s using historical data of accounting balances, then, the impact of recent cohorts will be overestimated. Cohorts after 2013 account for more or less 80% of the number of customer accounts. On the contrary, if we use historical data of average account balances, then the impact of old cohorts will be overestimated. Indeed, cohorts after 2013 "only" represent two thirds of the total accounting balance. This confirms the idea that features of new customers have changed compared to features of "old" customers. So if either case, and because of this specific growth the last few years, we cannot use historical data of the bank. The historical data of the national central bank are then well appropriate to calibrate the model in a more general way.

To calibrate d and s, we will do the following:

- We use historical data of the national central bank and the historical one-month Euribor rates, from January 2010 to May 2020.
- We perform, for different values of d and s, a linear regression between the trend of the time series T_{STL} (representing historical data of the accounting balances) and the cumulative sum of $P_1(t)$ and $P_2(t)$. The trend T_{STL} can be obtained with a STL (Seasonal and Trend decomposition using Loess) decomposition, that is a robust method for decomposing time series into three components: the trend part, the seasonal part, and a residual term. We will use here a additive STL decomposition, where three components are summed up.

  $$T_{STL}(t) = a + b \cdot \sum_{i=1}^{t} P_1(t) + b_2 \cdot \sum_{i=1}^{t} P_2(t) + \epsilon(t), \quad (3.21)$$

  where T_{STL} is the trend component of the STL decomposition of the time series and $\epsilon(t)$ is the error term.
- Given a, b and b_2, we compute $\tilde{D}(t)$ as

  $$\begin{aligned}
  \tilde{D}(t_{i+1}) &= \tilde{D}(t_i) + b \cdot P_1(t_{i+1}) + b_2 \cdot P_2(t_{i+1}) && (3.22)\\
  \tilde{D}(t_0) &= a + bP_1(t_0) + b_2 P_2(t_0) && (3.23)
  \end{aligned}$$

- Then, we can compute the sum of squared residuals RSS as

  $$RSS = \sum_{i=0}^{n} \left(T_{STL}(t_i) - \tilde{D}(t_i)\right)^2. \quad (3.24)$$

- Eventually, the parameters d and s used in the model will be the ones that minimized the squared error RSS.

3.2.3. Calibration of k and σ

The dynamics of deposit volume follow a kind of Ornstein-Uhlenbeck process. Under the risk-neutral measure **Q**, an Ornstein-Uhlenbeck process $y(t)$ is defined by the following stochastic differential equation:

$$\mathrm{d}y(t) = k(\mu - y(t))\mathrm{d}t + \sigma \mathrm{d}W(t), \tag{3.25}$$

where k, μ and σ are constant parameters and $W(t)$ is a Wiener process.

In our case, we do not have a constant μ. Instead, we have a deterministic trend $\tilde{D}(t)$:

$$\mathrm{d}D(t) = k(\tilde{D}(t) - D(t))\mathrm{d}t + \sigma \mathrm{d}W(t). \tag{3.26}$$

Using an Euler discretization, we have

$$\begin{aligned} D(t+\Delta t) - D(t) &= k(\tilde{D}(t) - D(t))\Delta t + \sigma\sqrt{\Delta t}\mathcal{N}(0,1) \\ D(t+\Delta t) &= (1-k\Delta t)D(t) + k\Delta t\tilde{D}(t) + \sigma\sqrt{\Delta t}\mathcal{N}(0,1), \end{aligned} \tag{3.27}$$

We have monthly observations, and as we calibrate the parameters on a monthly scale, the time step Δt is equal to 1.

$$D(t+1) = (1-k)D(t) + k\tilde{D}(t) + \sigma\mathcal{N}(0,1). \tag{3.28}$$

Two different methods, an ordinary least squares method and a maximum likelihood estimation, will be used to calibrate the two parameters based on historical data of the bank. The historical data that will be considered here is the average accounting balances of the bank, without the intra-month volatility (denoted as $D(t)$). Given those data and in order to facilitate the calibration, the trend $\tilde{D}(t)$ will be assimilated to the trend component of the STL decomposition of the time series $D(t)$.

Maximum Likelihood Estimation

The maximum likelihood estimation is a statistical method to estimate parameters. This method uses the assumption that the estimated parameters are the ones that maximize the probability of obtaining the observed data set.

Given (3.28), the conditional distribution of $D(t+1)|(D(t), \tilde{D}(t), k, \sigma)$ is a normal one, where the expectation and the variance are given below:

$$\begin{cases} \mathrm{E}\Big(D(t+1)|(D(t), \tilde{D}(t), k, \sigma)\Big) = (1-k)D(t) + k\tilde{D}(t), \\ \mathrm{Var}\Big(D(t+1)|(D(t), \tilde{D}(t), k, \sigma)\Big) = \sigma^2 \end{cases} \tag{3.29}$$

The conditional probability density function is then given by

$$f\Big(D(t+1)|(D(t), \tilde{D}(t), k, \sigma)\Big) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{\Big(D(t+1) - ((1-k)D(t) + k\tilde{D}(t))\Big)^2}{2\sigma^2} \right) \tag{3.30}$$

The log-likelihood function L is easily derived using the equation (3.30):

$$\begin{aligned} L &= \ln \prod_{i=1}^{n} f(D(i)|D(i-1), \tilde{D}(i-1), k, \sigma) \\ &= \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(D(i) - ((1-k)D(i-1) + k\tilde{D}(i-1))\right)^2}{2\sigma^2}\right)\right) \\ &= n \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \sum_{i=1}^{n} \frac{\left(D(i) - ((1-k)D(i-1) + k\tilde{D}(i-1))\right)^2}{2\sigma^2} \\ &= -\frac{n}{2} \ln\left(2\pi\sigma^2\right) - \sum_{i=1}^{n} \frac{\left(D(i) - ((1-k)D(i-1) + k\tilde{D}(i-1))\right)^2}{2\sigma^2}. \end{aligned} \tag{3.31}$$

The first partial derivatives with respect to k and σ are given by:

$$\frac{\partial L}{\partial k} = -\frac{1}{\sigma^2} \sum_{i=1}^{n} \left(D(i-1) - \tilde{D}(i-1)\right)\left(D(i) - ((1-k)D(i-1) + k\tilde{D}(i-1))\right), \tag{3.32}$$

$$\frac{\partial L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} \left(D(i) - ((1-k)D(i-1) + k\tilde{D}(i-1))\right)^2. \tag{3.33}$$

Setting the two partial derivatives to 0, the estimated parameters are:

$$\begin{cases} k &= \dfrac{\sum_{i=1}^{n}(D(i) - D(i-1))(\tilde{D}(i-1) - D(i-1))}{\sum_{i=1}^{n}(\tilde{D}(i-1) - D(i-1))^2}, \\ \sigma^2 &= \dfrac{1}{n} \sum_{i=1}^{n} \left(D(i) - (1-k)D(i-1) - k\tilde{D}(i-1)\right)^2. \end{cases} \tag{3.34}$$

Least Squares Estimation

The Least Squares method is a widely used method in statistics for estimating the true value of some parameters. It minimizes the sum of the squares of the differences between the observations and the values to be found.

The equation (3.28) can be also written as follows:

$$D(t+1) = \beta D(t) + \gamma \tilde{D}(t) + \sigma \mathcal{N}(0,1), \tag{3.35}$$

where $\beta = 1 - k$ and $\gamma = k$.

At every time step i, the residuals R_i are given by

$$R_i = D(i+1) - \beta D(i) - \gamma \tilde{D}(i). \tag{3.36}$$

The sum RSS of squared residuals is equal to

$$\begin{aligned} RSS &= \sum_{i=1}^{n} R_i^2 \\ &= \sum_{i=1}^{n} D(i+1)^2 + \sum_{i=1}^{n} (\beta D(i) + \gamma \tilde{D}(i))^2 - 2 \sum_{i=1}^{n} D_{i+1}(\beta D(i) + \gamma \tilde{D}(i)). \end{aligned} \tag{3.37}$$

The first partial derivatives with respect to β and γ are

$$\frac{\partial(RSS)}{\partial\beta} = 2\sum_{i=1}^{n} D(i)(\beta D(i) + \gamma\tilde{D}(i)) - 2\sum_{i=1}^{n} D(i)D(i+1), \tag{3.38}$$

$$\frac{\partial(RSS)}{\partial\gamma} = 2\sum_{i=1}^{n} \tilde{D}(i)(\beta D(i) + \gamma\tilde{D}(i)) - 2\sum_{i=1}^{n} \tilde{D}(i)D(i+1). \tag{3.39}$$

Setting the two partial derivatives to 0, we can estimate both β and γ and check if we have indeed $\beta = 1 - \gamma$.

$$\beta = \frac{\sum_{i=1}^{n} D(i)(D(i+1) - \gamma\tilde{D}(i))}{\sum_{i=1}^{n} D(i)^2} \tag{3.40}$$

$$\gamma = \frac{\sum_{i=1}^{n} \tilde{D}(i)(D(i+1) - \beta D(i))}{\sum_{i=1}^{n} \tilde{D}(i)^2} \tag{3.41}$$

We have a system of two equations with two unknowns, so β and γ can be easily estimated.

If we have indeed $\beta \simeq 1 - \gamma$, then we can set $k = \gamma$. As for σ, this parameter can be directly estimated by computing the standard deviation of the residuals.

3.2.4. Calibration of a, b and b_2

We now want to calibrate a, b and b_2, where the following equations hold:

$$\tilde{D}(t + \Delta t) = \begin{cases} \tilde{D}(t) + b \cdot \Delta t & \text{if } S_{t;t+\Delta t} = 1, \\ \tilde{D}(t) + b_2 \cdot \Delta t & \text{if } S_{t;t+\Delta t} = 2, \end{cases} \tag{3.42}$$

$$\tilde{D}(t_0) = a + bP_1(t_0) + b_2P_2(t_0), \tag{3.43}$$

$$\tag{3.44}$$

where $\tilde{D}$ represents the trend followed by the deposit volume, and a, b and b_2 are respectively the intercept, the growth coefficient in state 1, and the growth coefficient in state 2.

The three parameters a, b and b_2 are first estimated by annual cohorts. The annual cohort of year Y includes all the customers that activated their accounts in year Y. However, the most recent cohorts do not have any history in a normal rates environment. We will then take a different approach, and only estimate the parameters for the ”old” cohorts (from 1991 to 2013) and not for the ”new” ones (from 2014 to 2020). As a matter of fact, the probability of being in state 1 (normal rates) at time t, $P_1(t)$, is almost always equal to 0 for the new cohorts, so we cannot estimate any growth coefficient in a normal-rate environment.

Starting from historical data of the average accounting balances for a given cohort, we compute the stable part of the average accounting balances AAB^S and then, with a STL decomposition, we get the trend component of this time series AAB^S. We perform now a linear regression between the trend component (that represents $\tilde{D}(t)$) and the cumulative sums of $P_1(t)$ and $P_2(t)$.

$$\tilde{D}(t) = a + b\sum_{i=0}^{t} P_1(t) + b_2\sum_{i=0}^{t} P_2(t) + \epsilon_{\tilde{D}}(t), \tag{3.45}$$

where $\tilde{D}(t)$ is the trend component of the stable part of the average accounting balances and $\epsilon_{\tilde{D}}(t)$ is an error term associated to this linear regression.

Then, for a given cohort, a, b and b_2 will be respectively the intercept, the long-term growth coefficient in state 1, and the growth coefficient in state 2. The parameters b and b_2 selected are such that

$$b = \sum_{i=1991}^{2013} b_i \cdot \omega_i \tag{3.46}$$

$$b_2 = \sum_{i=1991}^{2013} b_{2i} \cdot \omega_i, \tag{3.47}$$

where b_i and b_{2i} are the estimated parameters for cohort i, $\omega i = \frac{N_i^{CA}}{\sum_{j=1991}^{2013} N_j^{CA}}$ is the weight given to the cohort i and N_i^{CA} is the number of accounts for the cohort i.

As for the parameter a, it is slightly more difficult. a actually represents the non-optional stable part of the average accounting balance at the start of the simulation. In other words, the parameter a is independent of the level of interest rates. We do not also consider the optional stable part when initializing this parameter because the low-rate environment responsible for the term is assumed to be temporary in any way, even though this has been the case for a few years already. Given a cohort i, the non-optimal stable part of the accounting balance at time t, $AB_i^{Non-O}(t)$, assumes that we remain in state 1 during all the calibration period and can be defined as follows

$$AB_i^{Non-O}(t) = (a_i + b_i N_i^P(t)) \cdot N_i^{CA}(t), \tag{3.48}$$

where $N_i^P(t)$ is the number of period between the start of the calibration period and the start of the simulation t , and $N_i^{CA}(t)$ is the number of customer accounts for the cohort i at time t.

However, this parameter $AB_i^{Non-O}(t)$ can only be computed for the old cohorts. The general non-optional stable part of accounting balance at time t $AB^{Non-O}(t)$ can still be estimated, using a cross multiplication:

$$AB^{Non-O}(t) = \sum_{i=1991}^{2013} AB_i^{Non-O}(t) \frac{\sum_{i=1991}^{2020} AB_i(t)}{\sum_{i=1991}^{2013} AB_i(t)} \tag{3.49}$$

Eventually, as a represents the non-optional stable part of the average accounting balance, it will be estimated as

$$a = \frac{AB^{Non-O}(t)}{N^{CA}(t)} \tag{3.50}$$

3.2.5. Computation of the stable and volatile parts

Intuitively, the accounting balance always contains a volatile part, that will disappear almost instantly (because of loan repayment for instance) and a stable part, that will take several years to disappear. The distinction between the stable and the volatile parts is of capital importance in demand deposits modeling. To make this distinction, banks must follow the BIS Basel III ([4] and [5]) and the ECB IRRBB (Interest Rate Risk in the Banking Book) [6] recommendations for managing liquidity risk. But banks are also encouraged to develop ”their own methodologies for capital allocation and liquidity buffers based on their risk appetite” [1]. Complete explanations will not be detailed here, however, broadly speaking, it can be summarized as follows. At time t, the accounting balance $AB(t)$ is actually the sum of a stable part, and a volatile part, where the stable part can be itself divided in two more terms:

$$AB(t) = AB^{Vol}(t) + AB^S(t), \tag{3.51}$$

$$AB(t) = AB^{Vol}(t) + AB^{Non-O}(t) + AB^O(t). \tag{3.52}$$

The volatile part contains a constant intra-month volatility factor Vol_{IM} and a seasonality term $S(t)$, that vary from month to month. It is defined as follows by the bank:

$$AB^{Vol}(t) = AB(t) - Vol_{IM}\frac{AB(t)}{S(t)}. \tag{3.53}$$

The non-optional stable part AB^{Non-O} has been defined above (3.49). It is the stable part we would have if we were still in a normal-rate environment. The optional stable part AB^{O} captures the excess of deposits-taking because of the context of extremely low or negative rates.

One important indicator for the bank is the percentage of optional component in the stable balance, that the bank aims to minimize. This indicator will be denoted as $O_{\%}$ and can be computed at the simulation start date t_{Sim} as follows:

$$\begin{aligned} O_{\%} &= \frac{AB^{O}(t_{Sim})}{AB^{S}(t_{Sim})}, \\ O_{\%} &= \frac{AB^{S}(t_{Sim}) - AB^{Non-O}(t_{Sim})}{AB^{S}(t_{Sim})}, \end{aligned} \tag{3.54}$$

where t_{Sim} is the simulation start time.

3.2.6. Model Simulation

Once all the parameters are calibrated, we can, using data of a unique month that represents the start of the simulation, simulate thousands of paths over the horizon of simulation (here it is 40 years). The simulation will be done in three steps, that are the following:

1. Simulation of the dynamics of the average accounting balance using the stochastic differential equation (3.16).
2. Application of the method explained in definition 15 on every path to move from a dynamic simulation to a static simulation.
3. Application of the linear adjustment 15y on every path, transition from average accounting balance to accounting balance and inclusion of the volatile part between the first two months.

The first step is a classic Monte-Carlo simulation using an Euler discretization. A full Euler discretization scheme has been chosen for its simplicity. A partial Euler discretization scheme could also be used, as presented in the market rate model, but this will not influence the final trajectory. Steps 2 and 3 will have more influence on it, as we can see in Chapter 5. As already explained, the volatile part of accounting balance (or average accounting balance) should not be taken into account when simulating the deposit volume model. This volatile part is volatile by definition and will disappear almost instantly, so there is no sense of considering this part under a horizon of 40 years. It will then be taken into account at the last step, by subtracting the volatile part between the first two months. To start the simulation, D_0 and $\tilde{D}_0$ will be initialized at the stable part of the average accounting balance at the simulation date. Once the Monte-Carlo simulation is done, the second step consists of applying an adjustment on every path to move from a dynamic scenario to a static scenario. The growth coefficients b and b_2 are regressed in a dynamic context, so the Monte-Carlo simulation gives a dynamic result. However, the goal of demand deposits modeling is to answer the question: " How fast the customer accounts constituting the average accounting balance will disappear ?" Two methods can do so and are explained in section 2.5. However, the method 2 will be used here, as the method 1 requires additional analysis on the attrition rate. The third step will force all the paths to reach 0 before or at 15 years. This adjustment is actually a conservative measure that can be justified by different arguments. Firstly, hedging available instruments are typically much less liquid after the term 15 years. Also, the historical period used to calibrate the parameters only lasts 10 years, so the forecasts with a forty-year horizon might be too uncertain.

3.2.7. Summary of the deposit volume model

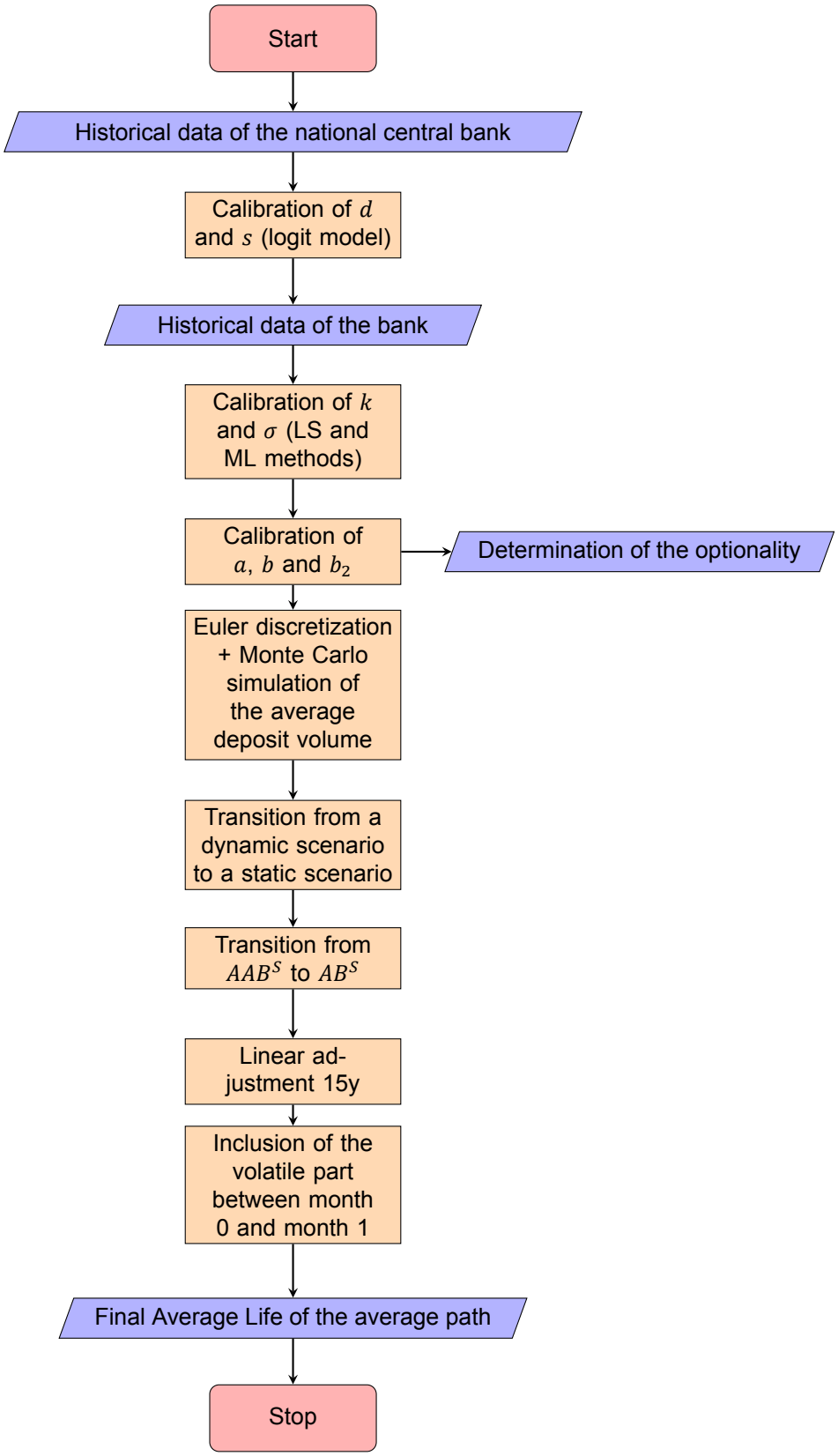

Figure 3.2: Flowchart of the deposit volume model

4

Customer segmentation

4.1. Context

The bank benefits from a rapid growth despite a declining banking sector. The number of clients has multiplied by three in four years. Moreover, new clients do not have the same characteristics as the old ones. One noticeable difference is based on the age. As a matter of fact, half of the new clients have less than thirty years. In addition, the business model of the bank has evolved. At the beginning, the bank was mainly for elderly and rich clients. Nowadays, new clients are drawn by low-cost banking fees. Then, it is usually younger clients with less wealth that come as new clients in the bank.

As seen in the previous chapters, deposit volume modeling is a two-state model, where the deposit volume follows a trend that depends on the one-month Euribor rates. The first state corresponds to standard rates environment, where in state two, rates are unusually low.

Since 2015, Euribor rates are non-positive. That raises the issue of how to treat clients of the bank that came after 2013 and that have mostly only experienced a low-rate environment. It is all the more important that those clients account for more than 80% of the bank's portfolio in terms of accounts number and almost two thirds of the bank's accounting balance.

This chapter aims to segment customers following one or several consistent explanatory variables. This customer segmentation could then help to predict how recent cohorts will behave if rates increase again.

4.2. Choice of the explanatory variable(s)

Every customer is unique. Assume that a customer has $1,000€$ on her demand deposit account. Then, in a very simplified manner, this customer can either spend a portion of the $1,000€$, keep this money on her account, or put a portion of this money on a saving account (usually remunerated). But her future behavior can be influenced by a significant number of variables, from her individual characteristics to her relationship with the bank and her interest in economic environment. Identifying those variables and picking the most influencing ones can be a hard task.

There are plenty of variables that can segment customers, depending on what we want to segment. The challenge here is to find groups of customers that will behave more or less the same, in terms of fluctuations of accounting balance, no matter in which state we are. Intuitively, some important explanatory variables that can segment customers efficiently can be the age, the gender, the wealth level when entering into a relationship with the bank, the financial knowledge, and macroeconomic factors. Some factors are even dependent on others, which complicates the task of segmenting customers. In what follows, we will quickly present several factors that can affect clients' behavior.

- Age

  Customers' behavior is clearly influenced by age. A client has different needs depending on

her age. While she will be more likely to make a loan in her thirties, she will be more likely to save money in her forties and then to spend her money in retirement. Also, the age influence people's tendency to have bank accounts in several banks. The behavior of a customer having one banking institution and another one having more banking institutions is clearly not the same.

- Wealth level 4 months after the activation date

  The wealth level when entering into a relationship with the bank is the accounting balance a customer will have 4 months after becoming a customer of the bank. As a matter of fact, the months following her account activation, a customer may have a strong growth of her accounting balance. Reasons for this include that the new customer will transfer her money, her salary deposit, and so on. This time period is quantified at 4 months but this choice is partly arbitrary.

- Income

  The income influences both the ability to save and the number of banks customer have and use. Intuitively, people with high income tend to save more than people that earn a low income. But the saving horizon is also influenced by income. High-income customers will save on a longer horizon than low-income customers. As for the number of banking institutions, low-income people will be more likely to have only one bank while high-income people will usually have several banks. [7]

- Financial knowledge

  Intuitively, people with financial knowledge tend to pay attention to the daily financial conditions. They turn out to be logically more arbitrageur.

- Macroeconomic factors

  Macroeconomic factors, such as the unemployment rate, the inflation, and so on, influence interest rates, and interest rates influence customers' behaviors. As a matter of fact, if the saving account interest rates tend to be high, people will be more likely to put money on their savings account and leave less in their demand deposits accounts, this account being not remunerated by the bank.

Eventually, many factors can influence clients' behavior. However, some of them cannot be obtained with reliability or cannot be obtained at all from the bank database. The age and the wealth level when entering into a relationship with the bank appear to be particularly consistent and are readily available. The wealth level when entering into a relationship with the bank is fixed 4 months after the account activation and remains constant over time, while the age will evolve at each time step. Then, those two variables, one remaining constant and the other changing over time, will effectively segment customers and will be incorporated each month into the deposit volume model without adding too much complexity.

However, we cannot consider the age and the accounting balance 4 months after becoming a customer of the bank for each customer. The data extraction and the data processing would be too long and too complex. Therefore, we need to create a number x of categories of age and a number y of categories of wealth level, such that each couple (Age ; wealth level) accounts for more or less $\frac{1}{xy}$ of the total number of accounts. Arbitrarily, we set $x = 10$ and $y = 10$, so that we have 100 categories in total. This makes it possible to be sufficiently accurate without being too complex. The determination of the 10 categories of age and the 10 wealth level is an optimisation problem, that can be resolved using a solver function on Excel. This step will not be detailed here, as the Solver Add-in on Excel did all the work. In what follows, the 10 categories of age and the 10 categories of wealth level will be numbered from 1 to 10 in ascending order in both case. The couple $"02\_04"$ will then represent the customers in the 2^{nd} category of age and in the 4^{th} wealth level. For confidentiality reasons, we will not explicit the categories of age and of wealth.

4.3. Data processing

There is quite an upstream work of processing the data before the cluster analysis. The data extracted from the internal software contains 7 different columns that gives information about the birth date, the date, the year of activation, the semester of activation, the wealth level, the number of accounts, and

the accounting balance. Two customers will then be grouped together if they have the same birth date (YYYYMM), the same year and semester of activation and the same wealth level. The data processing involves several steps.

1. We start by adding five new columns that are respectively the age, the age category (from ”01” to ”10”), the accounting balance one month after (M+1) and the number of customer accounts one month after (M+1) and a new variable ”AC_WC” that concatenates the age category and the wealth category.
2. We remove the rows where we cannot find any data for the month after.
3. We create a pivot table that groups data by the ”Date” and the ”AC_WC”, and that sums all the accounting balance and the number of customers accounts at the date and at the month after.
4. We compute for every row of the pivot stable the stable part of the average accounting balance for the date M, $AAB^S(M)$, and for the month after $M+1$, $AAB^S(M+1)$.
5. We compute for every row the growth as: $AAB^S(M+1) - AAB^S(M)$.
6. We weight the number of customer accounts per date. For each date, the sum of the weights should be equal to 1.
7. We create two data frame that will respectively contain: the data of the pivot table in state 1, representing a normal-rate environment, and the data of the pivot table in state 2, representing a low-rate environment.
8. In both data frames, we normalize the weighted coefficients.
9. For each data frame, we create a new pivot table that groups data by the variable ”AC_WC”, and that computes the weighted mean and the weighted standard deviation of the growths. Each data frame will contain 100 row, as we have 100 possible values for the concatenation ”AC_WC”.
10. Finally, we will join the two pivot tables in only one.

4.4. Clustering

When the data has been proceed, we can use clustering algorithms to segment customer. Both the k-means algorithm and the hierarchical clustering will be tested in this thesis. The k-means algorithm implies that the optimal number of clusters has already been determined. To determine the optimal number of clusters for the k-means algorithm, we can use either a Elbow method or a gap statistic method. As for the hierarchical clustering, we will also the gap statistic method to pick the optimal number of clusters. All the theoretical explanations were given in section 2.4. The clustering is done on the 100 ”AC_WC” data points, under 4 variables: the weighted mean in state 1, the weighted standard deviation in state 1, the weighted mean in state 2 and the weighted standard deviation in state 2.

4.5. Integration of the customer segmentation in the model

The stage of the clustering, either using the K-means method or the hierarchical clustering, is rather simple once the data are processed. After this step, we obtain, for each pair ”AC_WC” the cluster to which it belongs. This data is added to the final data frame obtained at the last step of the data processing. Then, the challenge is to integrate the cluster analysis in the deposit volume model. The most intuitive way to include the clustering in this model is to modify the parameters b and b_2 that represent the long-term growth in state 1 and the growth in state 2. Assuming that we have p number of clusters, instead of a unique b and a unique b_2, we will have p coefficients b and p coefficients b_2, such that b_j and b_{2j} represent respectively the average long-term growth coefficient in state 1 for cluster j and the growth coefficient in state 2 for cluster j. For each cluster j from 1 to p, b_j and b_{2j} can be directly from the data frame as we already have for each pair ”AC_WC”, the associated growth coefficients in state 1 and in state 2. However, some adjustments need to be done before starting the deposit volume model simulation. They can be summarized in the following steps:

1. We start by extracting data for the simulation start date t_{Sim}. For every row, the following information is provided: the date (YYYYMM), the date of birth (YYYYMM), the activation year, the activation month, the wealth level, the accounting balance and the number of customers accounts. For every row, we compute the date 4 months after the activation date, named t_{M+4}, that will be used later in step 5.

2. We concatenate the date of birth, the age group and the wealth level in the format: "YYYYMM_AC_WC".

3. We create a pivot table that groups customers by the concatenation variable and by the activation year, and that sums all the accounting balance, the number of customer accounts and the stable part of the average accounting balance.

4. Then, we join this data frame with the clustering data frame to obtain, for every row, the associated cluster (from 1 to p).

5. The idea is now to use a backward-looking approach and then a forward-looking approach to guess the stable average accounting balance of each row if we were still in a normal-rate environment. To do so, we go back in time for every row until the date where we moved from state 1 to state 2. Let us denote by t_{CS} the date of the change of state, in the format "YYYYMM". At every time step, and for every row, we need to modify the age, the age group, the date, the associated cluster, the growth coefficients and finally the stable average accounting balance. Assuming that a specific row belongs to a cluster i at time t, the stable average accounting balance is then defined as follows:
$$AAB^S_{tmp}(t) = AAB^S_{tmp}(t+1) - b_{2i}, \tag{4.1}$$
where b_{2i} is the growth coefficient of cluster i in a context of low rates and the characters "tmp" are put in subscript to specify that it is a temporary adjustment. If a row has its four-month activation date t_{M+4} before the date of the change of state, then we will stop at the date t_{M+4}.

6. After looking back, we use a forward-looking approach to estimate the evolution in case of normal rates. Then, for every row, we will start either at the t_{CS} date or at the t_{M+4} date and at every time step until the simulation start date t_{Sim}, we modify the age, the age group, the date, the associated cluster, the growth coefficients and the stable average accounting balance. The last variable, at time $t+1$ for a specific row belonging to a cluster i is defined as follows:
$$AAB^S_{tmp}(t+1) = AAB^S_{tmp}(t) + b_i, \tag{4.2}$$
where b_i is the growth coefficient of cluster i in state 1.

7. After the backward-looking and the forward-looking approaches, we obtain a data frame with the following information for each row: the concatenation parameter, the activation year, the accounting balance, the number of customer accounts, the stable part of the average accounting balance and the reconstructed stable part of the average accounting balance. The last two variables can be transformed as accounting balance by multiplying by the number of customer accounts. The reconstructed stable part of the accounting balance actually represents the non-optional stable part of the accounting stable.

8. We can now create a final pivot table where we group the data only by the concatenation parameter, and where we have 5 other columns: the number of customer accounts, the accounting balance, the non-optional stable part of the accounting balance, the optional-stable part of the accounting balance, and the volatile part of the accounting balance. We recall that we have the following equations, at every time t:
$$AB^S(t) = Vol_{IM}\frac{AB(t)}{S} \tag{4.3}$$
$$AB^{Vol}(t) = AB(t) - AB^S(t) \tag{4.4}$$
$$AB^O(t) = AB^S(t) - AB^{Non-O}(t), \tag{4.5}$$
where S is the seasonal coefficient of t_{Sim}, Vol_{IM} is the intra-month volatility and AB and AB^{Non-O} are already determined in the previous steps. The three unknowns can then be easily computed

for each row. From this data frame, we can compute the percentage of optional component in the stable balance $O_{\%}$ as in Chapter 3:

$$O_{\%} = \frac{\sum_{i=1}^{R} AB_i^O(t_{Sim})}{\sum_{i=1}^{R} AB_i^S(t_{Sim})}, \tag{4.6}$$

where R is the number of rows in the final data frame. In other words, R is the number of different combinations that we have for the date of the birth, the age group and the wealth group.

9. Once we have this final data frame, we only need to set a, b and b_2. As the age is a dynamic variable, a customer belonging to a cluster i at time t can belong to a cluster j at time $t+1$. So the number of customers into each cluster evolves at each time step, and that will influence b and b_2. Those two parameters will not be constant anymore but will vary each month. We start by creating a matrix $Evol_{Clust}$, that will contain the clusters' evolution, of size $R \times N_T$, where N_T is the number of time observation. At every time step, the pair "AC_WC" of each row will evolve so the associated cluster to this row may evolve as well.

10. Then, we create a vector Ω of length R that will contain the weighted coefficients of the number of customer accounts at D_{Sim}. The vector will remain constant over time as we will not consider any attrition. For any i between 1 and R, we have:

$$\Omega_i = \frac{N_i^{CA}}{\sum_{j=1}^{R} N_j^{CA}}, \tag{4.7}$$

where N_i^{CA} is the number of customer accounts of row i.

11. At time t, the two growth coefficients $b(t)$ and $b_2(t)$ are defined as:

$$b(t) = \sum_{i=1}^{p} b_i \Big(\sum_{\substack{j=1,\\ j\in i}}^{R} \Omega_j \Big) \tag{4.8}$$

$$b_2(t) = \sum_{i=1}^{p} b_{2i} \Big(\sum_{\substack{j=1,\\ j\in i}}^{R} \Omega_j \Big) \tag{4.9}$$

To avoid to include dynamic elements that are not justified from a business point of view and therefore not quantified, we will make a final adjustment:

$$b^S(t) = \min(0, b(t)) \tag{4.10}$$

$$b_2^S(t) = b2(t) + b^S(t) - b(t), \tag{4.11}$$

where $b^S(t)$ and $b_2^S(t)$ will be the final parameters used for the model simulation.

12. Eventually, we perform the simulation as in section 3.2 and the same adjustments afterwards (simulation in a static scenario and a fifteen-year linear adjustment).

5 Results

This chapter presents the results for both the market rate model and for the deposit volume model. These results are achieved with the software R. However, the input data are confidential and have then been rescaled. As for the market rate model, only one proposal is made. But, as for the deposit volume model, three proposals are made. The first one is the basic model, without considering the loss of homogeneity among customers. Proposals two and three introduce in each case two explanatory variables in the model that are respectively the age and the wealth level 4 months after the account activation. Those two variables are used in the clustering to form similar groups of customers. The difference between proposals two and three is the cluster analysis method. The clustering is then integrated in the deposit volume model, that is based on an Ornstein-Uhlenbeck process. Important outputs of demand deposits modeling are the average life and the optionality.

5.1. Market rate model

The market rate model is based on a one-factor Hull-White model, and the parameters are calibrated using historical data, the one-month Euribor rates. However, the time-limit to be taken into account needs to be determined and is quite arbitrary. We will choose here to take 7 years of data, but before 2010. The choice is justified by the fact that the last decade, the one-month Euribor rates were not volatile enough (see figure 2.1). The calibration period is then from 2003-01-01 to 2009-12-31. So using the calibration method described in Chapter 2, the estimations for the speed of mean-reversion and the volatility are given in the table 5.1 below.

a_{HW}	5.60 %
σ_{HW}	0.39 %

Table 5.1: Estimated parameters for the single factor Hull-White model

Once a_{HW} and σ_{HW} are estimated, we can construct the instantaneous forward rate yield curve using the zero-coupon yield curve, compute $\alpha(t)$ for every t over the horizon of simulation, where $r(t) = \alpha(t) + x(t)$. As $\alpha(t)$ is deterministic and $x(t)$ is a stochastic process, we will only perform a Monte-Carlo simulation on 10 000 paths of $x(t)$, given three different simulation schemes and then add the deterministic function $\alpha(t)$ to every path. The three simulation schemes compared here are the complete Euler discretization scheme, the partial Euler discretization scheme and the exact simulation. In the figure 5.2 below, we compare the simulation run-times with 10 000 paths for the three methods, under the same conditions.

Simulation scheme	Run-time in seconds
Full discretization	22.39
Partial discretization	24.44
Exact simulation	34.11

Table 5.2: Comparison of the simulation run-times

Similarly, we can compare the error for the full discretization and the partial discretization schemes, for different seeds.

	Full discretization	Partial discretization
Seed 1	0.02148489	0.02148009
Seed 2	0.02153032	0.02152547
Seed 3	0.02176855	0.02176373

Table 5.3: Comparison of the absolute error at time $T = t_n$

The run-times of the three Monte-Carlo simulation schemes are of the same order, the exact simulation being a bit longer than the two others. However, none of the above methods have a discriminatory run-time. As for the absolute error, the values are extremely close at each seed. In this very particular case, the Euler scheme and the Milstein scheme [17], defined in equation (5.1) below, are equivalent. Given the stochastic differential equation $\mathrm{d}x(t) = -a_{HW}x(t)\mathrm{d}t + \sigma_{HW}\mathrm{d}W(t)$, the Milstein scheme is defined as follows:

$$\begin{aligned} x(t+\Delta t) &= x(t) - a_{HW}x(t)\Delta t + \sigma\sqrt{\Delta t}\mathcal{N}(0,1) + \frac{1}{2}\sigma \underbrace{\frac{\partial\sigma}{\partial x(t)}}_{=0} \cdot(\Delta t(\mathcal{N}(0,1))^2 - \Delta t), \\ x(t+\Delta t) &= (1 - a_{HW}\Delta t)x(t) + \sigma\sqrt{\Delta t}\mathcal{N}(0,1). \end{aligned} \tag{5.1}$$

So as the two discretization schemes are equivalent, then, under the Euler method, the estimated short rate converges strongly to the short rate with order 1.

Eventually, a sample of simulated rate paths and the average path, using the exact Monte-Carlo simulation, are shown in figure 5.1:

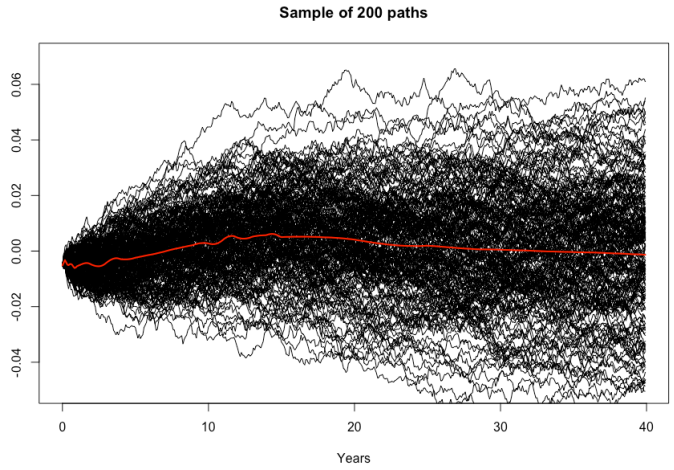

Figure 5.1: Sample of 200 paths simulated with a one-factor Hull-White model

5.2. Deposit volume model

The second part of demand deposits modeling consists of deposit volume modeling. This model is based on an Orstein-Uhlenbeck process, but the drift-term is not constant. Actually, the drift-term corresponds to a trend $\tilde{D}_t$ that is a two-state function and that depends on the market rates. State one represents a normal-rate environment, and state two represents a low-rate environment. Three proposals are made for this second part of demand deposits modeling. Proposal 1 is rather restrictive and simulates all the cohorts without really considering the characteristics of the bank. Proposals 2 and 3 include a cluster analysis, either using k-means algorithm or a hierarchical clustering. Even though the three proposals are quite different, the calibration step of some variables is common in the three proposals.

5.2.1. Calibration of some parameters in the deposit volume model

We recall that the deposit volume model, in its basic form and without the cluster analysis, is:

$$\mathrm{d}D_t = k(\tilde{D}_t - D_t)\mathrm{d}t + \sigma \mathrm{d}W(t), \tag{5.2}$$

$$\tilde{D}_{t+\Delta t} = \begin{cases} \tilde{D}_t + b * \Delta t & \text{if } S_{t;t+\Delta t} = 1, \\ \tilde{D}_t + b_2 * \Delta t & \text{if } S_{t;t+\Delta t} = 2, \end{cases} \tag{5.3}$$

$$P(S_{t;t+\Delta t} = 1) = P_1(t) = \frac{1}{1 + \mathrm{e}^{-d(r_t - s)}}, \tag{5.4}$$

$$P(S_{t;t+\Delta t} = 2) = P_2(t) = 1 - P(E_{t;t+\Delta t} = 1). \tag{5.5}$$

(5.6)

Then, the calibration of the parameters d, s, k and σ is common in the three proposals, and the method used is detailed in Chapter 3. The parameters of the logistic model are calibrated using historical data of accounting balances of the national central bank. The estimated parameters, that are used to determine the probability of being in either state, are:

d	20 000
s	0.0007

Table 5.4: Estimated parameters for the logistic model

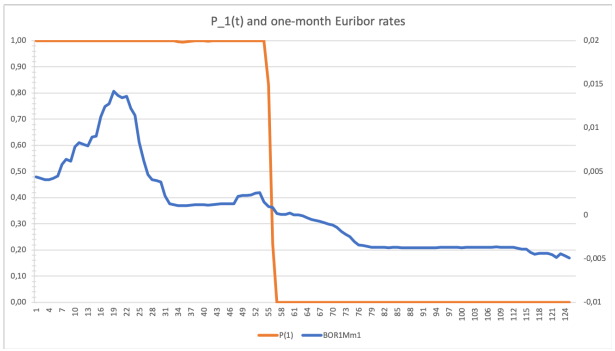

Figure 5.2: Probability of being in state 1 and the one-month Euribor rates

In the graphic 5.2, the probability of being in a normal rates environment and the one-month Euribor rates are shown in two different scales. Roughly, we can see that as soon as the Euribor rates are below $s = 0.0007$, then we switch almost instantly from state 1, to state 2, that represents a low-rate environment. So once the probability of being in state 1 has been determined, the deposit volume model can be assimilated to a regime switching model, where the growth coefficient at each step is b, if $r(t) > s$ or b_2, if $r(t) < s$ In the very particular case of $r(t) = s$, then the growth coefficient is the mean of b and b_2.

The next step, after the calibration of the logistic model, is to calibrate k and σ in the deposit volume equation (5.2). We use historical data of the stable part of average accounting balances. Both the maximum likelihood estimation and the least squares method are tested. The results are given in the table 5.5

	Maximum Likelihood	Least Squares
k	0.950	0.942
σ	149.32	148.87

Table 5.5: Estimated parameters for the dynamics of deposit volume

The two approaches give very similar results for both k and σ. However, the maximum likelihood is to be preferred as it is asymptotically optimal when estimating unknown parameters. In other words, when the size n of the sample is large enough (usually large means 30), the maximum likelihood method guarantees to perform better than any other estimation method, and than the least squares method in particular. In our case, we have around 120 time observations. In what follows, the estimated k and σ will be the ones found with the maximum likelihood method.

5.2.2. Proposal 1

Proposal 1 corresponds to the foundations of the model, that have been explained in Chapter 3. The parameters a, b and b_2 remain to be calibrated. b and b_2 are respectively estimated as the weighted mean of b_i and b_{2i} for every old cohort i. The last parameter a roughly represents the average accounting balance that would have been reached after the calibration period if we were still in a normal-rate environment. Details about the method used to calculate this parameter are given in Chapter 3. The results are given in the table 5.6:

a	1850
b	- 5.10
b_2	32.4

Table 5.6: Estimation parameters for a, b and b_2

In other words, it means that in a normal-rate environment, the average accounting balance will decline by $5€$ per month while it will increase by $32€$ in a low-rate environment. It is fairly intuitive as when rates are quite high, customers will be more likely to move money from their demand deposits account to a saving account, usually remunerated by the bank.

The next step after the parameters' calibration is to simulate the deposit volume model. An exact Monte-Carlo simulation is not considered here, as the regime-switching model adds too much complexity. We arbitrary chose to use a full Euler discretization but this choice does not have a significant impact on the final paths. Indeed, after this first dynamic simulation, we do a static adjustment and we force every path to reach 0 before or at 15 years. Those two adjustments will have much more impacts on the final paths. The graphics obtained at each step are presented in figures 5.3, 5.4 and 5.5.

From those three graphs, we can see the effects of the two adjustments done in the deposit volume model: first a transition from a dynamic scenario to a static scenario and then the fifteen-year linear adjustment as well as the volatile-part inclusion. As a matter of fact, in figure 5.3, we reach a plateau after 20 years, and then the average accounting balance remains around 1800. The average life in this case is around 23 y (see 5.7) but this is quite unrealistic from a practical point of view. As a matter of fact, the bank must comply with the IRRBB principles [6], that caps the average maturity of NMDs at 5 years. The transition to a static vision achieves more or less the ”zero” goal at 40 years, but the bank aims at a zero-goal at 15 years, for several reasons explained earlier (see subsection 2.6.1). Also, a significant difference between the first two steps and the third step is the evolution between the first two months. As a matter of fact, the volatile-part inclusion leads to a huge decrease between the first two months, as the volatile part disappear almost immediately.

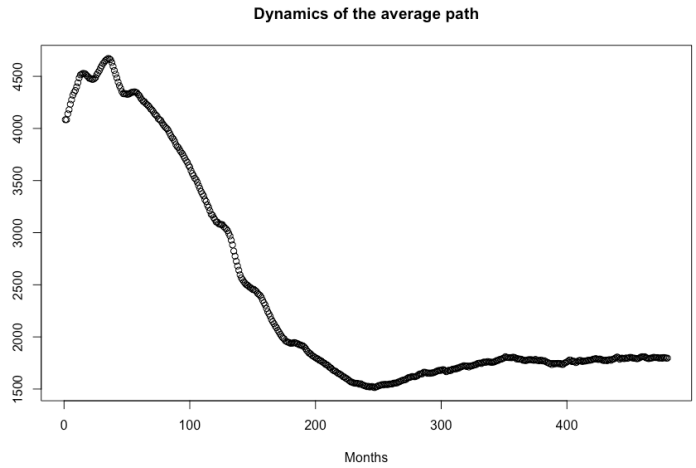

Figure 5.3: Dynamics of the average path

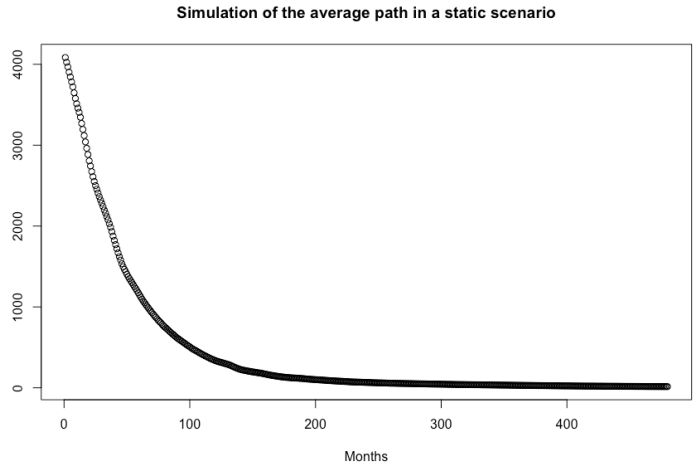

Figure 5.4: Simulation of the average path in a static scenario

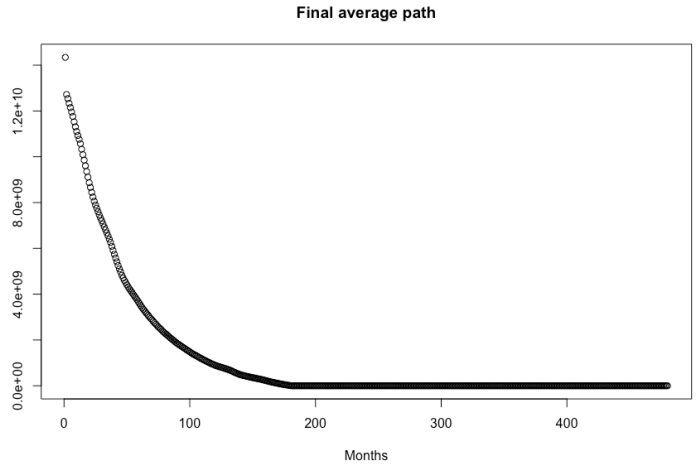

Figure 5.5: Final average path

The average life for each step and the optionality can also be computed and are given in table 5.7:

Dynamic scenario (stable part)	23.62 y
Static scenario (stable part)	4.25 y
Final average life	3.45 y
Optionality	58.5 %

Table 5.7: Outputs of the model - proposal 1

The average life at each step of the simulation are shown in table 5.7, however, only the average life of the final path are actually used in practice. Conclusions may be drawn from the final average life and from the optionality. We can observe a high optionality of 58.5% and a final average life that could be improved. Overall, proposal 1 assumes that the new cohorts will behave as a weighted average of old cohorts but this may be too conservative. However, it provides a good approach in demand deposits modeling, in a basis case. Proposals 2 and 3 will include both the age and the wealth category at four months in the deposit volume model. This captures better the characteristics of new customers, that are quite different from old ones.

5.2.3. Proposal 2

Proposal 2 includes a cluster analysis using a hierarchical algorithm in the deposit volume model. This cluster analysis forms similar groups of customers, that will behave identically in a context of normal rates and in a context of low rates. The hierarchical clustering can be done either with a top down approach, or with a bottom up approach. As for the splitting or merging criterion, we will use the Ward's method [13] under the Euclidean distance. All the methods to calculate the similarity have pros and cons, however, the Ward's method performs well in separating clusters if there is noise between clusters. We tried both the divisive hierarchical clustering, also named as DIANA, and the agglomerative hierarchical clustering, also named as AGNES. In practice, the agglomerative hierarchical clustering is often used. The differences as for the outputs of the model between DIANA and AGNES, for a similar number of clusters, are extremely small. So we will here present the agglomerative cluster analysis and the results of the divisive cluster analysis are left in Appendix B.1. The dendrogram for the agglomerative clustering is shown in figure 5.6. It gives the height at which any two objects will be joined together. Choosing 3 or 4 clusters seems to be a good suggestion, however, this needs to be supported by the gap statistic method, shown in figure 5.7. Given the method described in subsection 2.4.3, the optimal number of clusters here is 3. Once the clustering is done, internal clustering validation measures are quite useful to check quality of the clustering. Those internal measures are explicitly given in [23] and are presented in Appendix B.2.

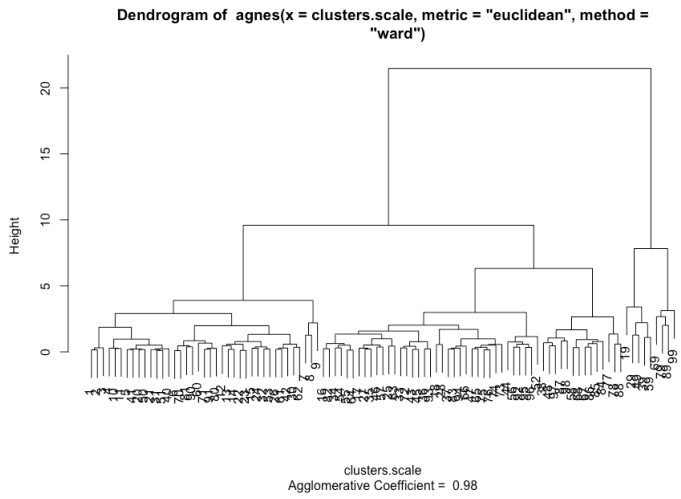

Figure 5.6: Dendrogram under the agglomerative cluster analysis

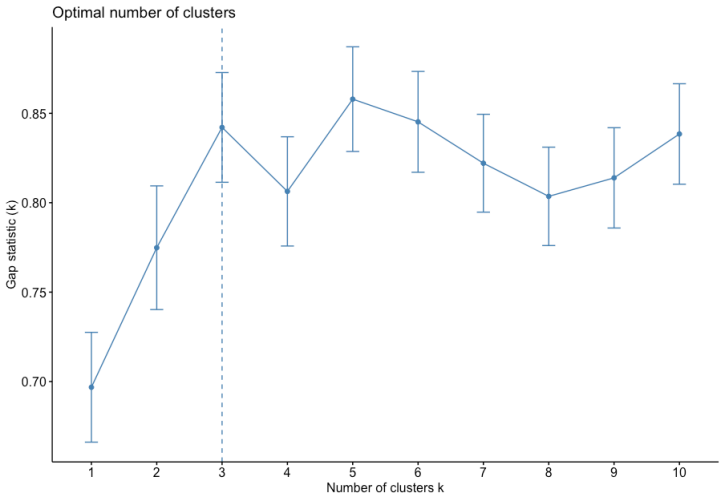

Figure 5.7: Gap statistic method under the agglomerative cluster analysis

The next step once the clusters are formed is to simulate the deposit volume. As precised in Chapter 4, the parameters b and b_2, that are respectively the growth coefficient in state 1 and the growth coefficient in state 2 are not constant anymore. At each time step, a customer can move from one cluster to another, so the number of accounts in each cluster will evolve as well, and finally, the growth coefficients too. The different steps of this simulation, first a dynamic simulation and then with some adjustments are shown in figures 5.8, 5.9 and 5.10. Similar conclusions to proposal 1 can be drawn. The effects of the final adjustments become even more visible here, as the dynamic simulation of the deposit volume model seem to increase indefinitely.

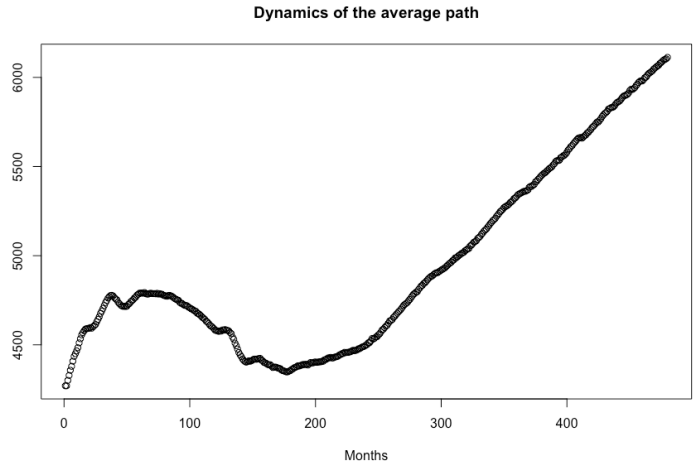

Figure 5.8: Dynamics of the average path - AGNES method with 3 clusters

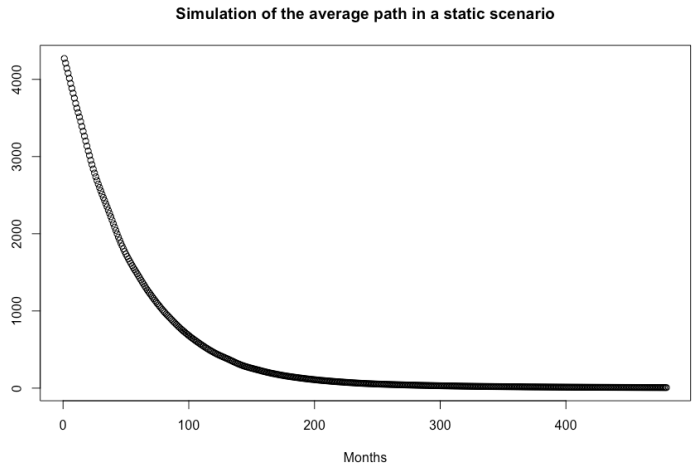

Figure 5.9: Simulation of the average path in a static scenario - AGNES method with 3 clusters

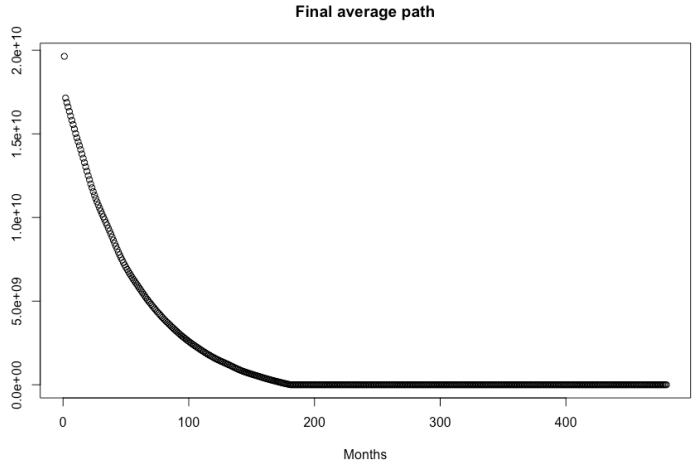

Figure 5.10: Final average path - AGNES method with 3 clusters

Eventually, all the different outputs of the model are given in table 5.8:

Dynamic scenario (stable part)	46.15 y
Static scenario (stable part)	4.65 y
Final average life	3.78 y
Optionality	21.9 %

Table 5.8: Output of the model - proposal 2

Proposal 2 introduces a customer segmentation in the deposit volume model, through a agglomerative cluster analysis. We form three different clusters, and each customer, given his age and his wealth level at four months, will be assigned to a cluster at each time step. This customer segmentation will provide a more realistic simulation of the deposit volume model. Compared to proposal 1, the final average has increased and the optionality has reduced, which constitutes an optimal scenario for any bank. We can also remark that the five-year cap for the average maturity of the NMDs [6] is still respected.

5.2.4. Proposal 3

Proposal 3 also includes a cluster analysis in the deposit volume model, but using a K-means method [12]. This method is described in the subsection 2.4.1. As for the optimal number of clusters, both the Elbow method [15] and the gap statistic method [21] are used, the first one being more a heuristic method. The two graphs are shown in figures 5.11 and 5.12. The figure 5.11 suggests 3 or 4 as an optimal number of cluster. However, this method is quite subjective, and only gives indications. The gap statistic method, presented in figure 5.12, provides more accurate results, and picks 3 as the optimal number of cluster. Similarly to proposal 2, internal cluster validations measures are presented in Appendix B.2.

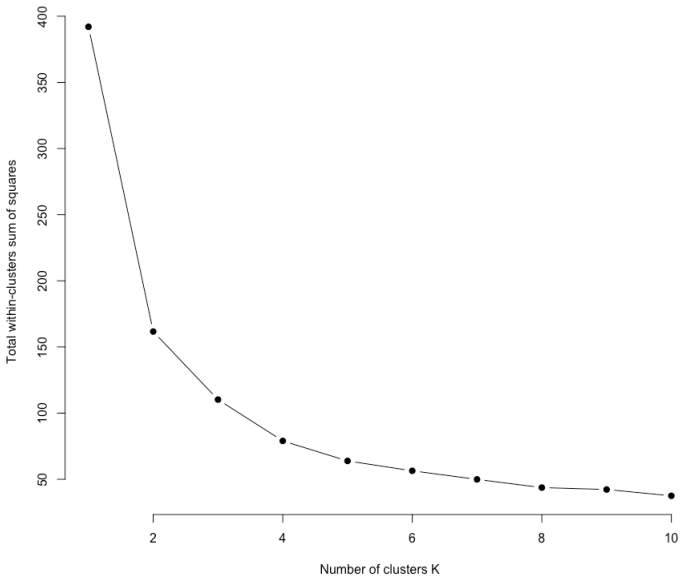

Figure 5.11: Elbow method under the K-means algorithm

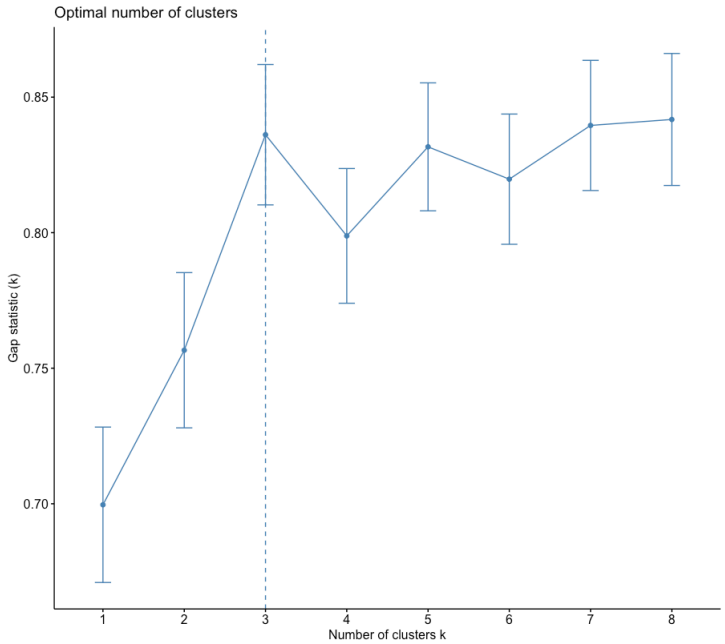

Figure 5.12: Gap statistic method under the K-means cluster analysis

Once the clusters are formed, the dynamics of the deposit volume model are simulated using Monte-Carlo and under an Euler discretization. An exact Monte-Carlo simulation is not an option here, as the stochastic differential equation is too complicated to compute exactly the deposit volume. However, the differences between a full Euler discretization scheme and between a partial Euler discretization schema are really small, and will not influence the final average path, after adjustments. The three graphs are presented in figures 5.13, 5.14 and 5.15. Similar conclusions to proposal 2 can be drawn.

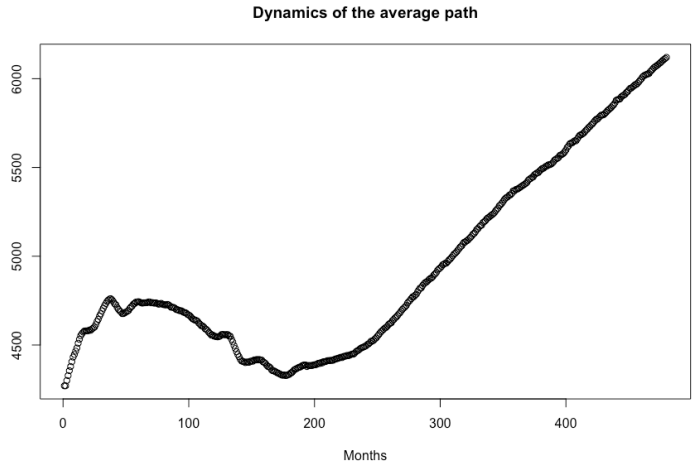

Figure 5.13: Dynamics of the average path - K-means method with 3 clusters

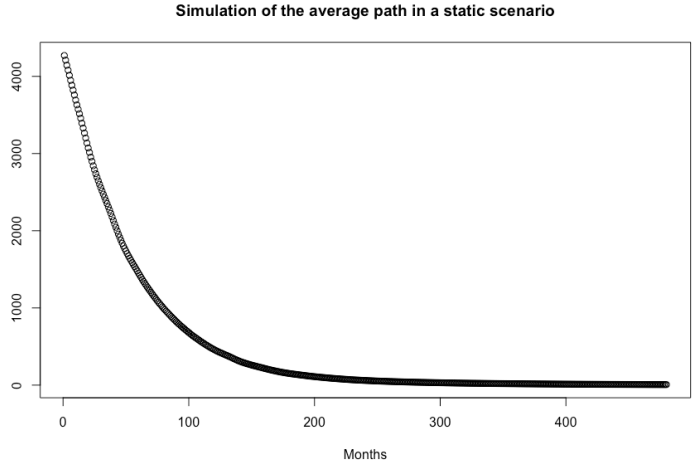

Figure 5.14: Simulation of the average path in a static scenario - K-means method with 3 clusters

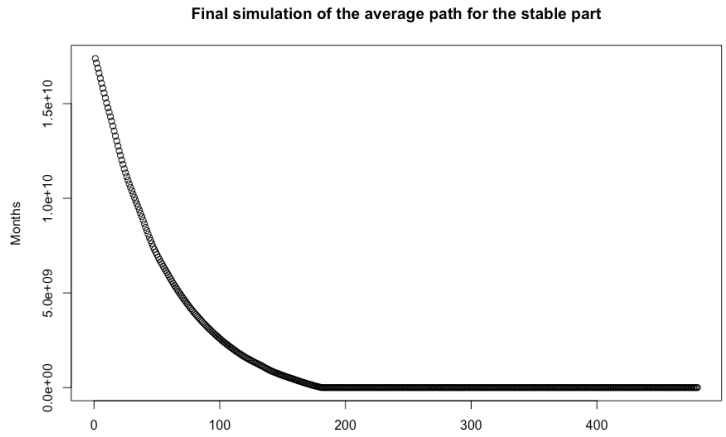

Figure 5.15: Final average path - K-means method with 3 clusters

The different outputs of the model can be easily computed. Results are given in table 5.9. Proposals 2 and 3 give really extremely similar results, in terms of optionality and of final average life. Once again, the five-year cap for the average maturity of the NMDs is respected.

Dynamic scenario (stable part)	45.22 y
Static scenario (stable part)	4.67 y
Final average life	3.79 y
Optionality	22.7 %

Table 5.9: Output of the model - proposal 3

6

Conclusion and recommendations

In this chapter, we present the conclusion of our research and discuss some recommendations for future work.

6.1. Conclusion

In this thesis, we studied demand deposits modeling at a European bank. The academic literature on demand deposits models is quite rich in references [14] [16] [20]. However, those models were introduced at least 10 years ago, and the last few years have experienced extremely low and even negative interest rates that were not really envisaged by all the models. Moreover, the bank has its specificities that distinguishes it from other banks on the market. For instance, the deposit volume remain more or less stable year after year for the large majority of banks. This is clearly not the case for the bank, that benefits from a rapid growth in recent years. Simultaneously, the customers' characteristics have changed, and the new customers are quite different from the old ones, usually younger and less wealthy. The parallel between the number of accounts and the deposit volume for the recent customers clearly illustrates the differences between old customers and new ones. As the matter of fact, customers that arrived after 2013 account for more than 80 % of the bank portfolio in terms of number of accounts. However, they only account for two thirds of the bank portfolio in terms of deposit volume. All theses specificities made demand deposits modeling rather important at the bank, and most of the models presented in the literature cannot be applied to this bank. Overall, the goal is this thesis was to **model demand deposits of the bank while taking its features into account**.

Demand deposits modeling is a three-step model, where the steps are respectively the market rates, the deposit volumes and the deposit rates. The deposit rates being not remunerated by the bank, only the first two steps will be considered in this thesis. As for the market rate model, a single factor Hull-White model has been chosen for several reasons. This model allows interest rates to be negative, which is crucial in current conditions. The one-factor Hull-White model also has the advantage of fitting perfectly the initial term structure of interest rates. A two-factor Hull-White model has not been tried, as we aimed at modeling market rates without adding further complexities, this model not being the main focus of the thesis. To calibrate the parameters, the $\theta(t)$ function is determined from the initial yield curve that describes the current term structure of interest rates. Regarding the mean-reverting and the volatility parameters, they are calibrated using historical data of the one-month Euribor rates. Three different Monte-Carlo schemes are compared for the simulation. The first one is a full Euler discretization, the second one is a partial Euler discretization and the last one is an exact simulation. Details and results about the three schemes are given in subsection 3.1.2 and in section 5.1. In terms of accuracy and efficiency, the full and the partial discretization schemes give extremely similar results. The exact simulation has a longer run-time compared to the other two, however, this run-time is not really a selective criterion. Eventually, the exact Monte-Carlo simulation will be chosen for the rest of the thesis, and the market rate model will be used as input data in the deposit volume model.

The next step after the market rate model is to model the deposit volume. The model proposed in

this thesis is an extension of the Ornstein-Uhlenbeck (O-U) process, where the Brownian motion is independent from the Brownian motion in the market rate model. The drift term μ in the O-U process will be replaced by a trend $\tilde{D}(t)$, that definitely depends on the rates. Depending on the rates' level, we are either in a normal-rate environment, named as state 1, or in a low-rate environment, named as state 2. The slope between two consecutive months, as we have monthly data, will be b if we are in state 1 and b_2 if we are in state 2. Finally, the probability of being either in state 1 or in state 2 follows a logistic model and is a function of the market rate. The parameters' calibration is done through linear regression or through a maximum likelihood method. A full Euler discretization scheme is then used for the Monte-Carlo simulation. A partial Euler discretization scheme has also been tried, but has no influence on the final outcomes. Indeed, final adjustments are done after the Monte-Carlo simulation, in order to move in a static simulation and to force the deposit volume to reach zero before or at fifteen years. The arguments justifying those two adjustments are presented in section 2.5 and in subsection 2.6.1.

Because of the loss of homogeneity among its customers in recent years, the bank aimed at include several explanatory variables in the deposit volume model to better model customers' behavior. Two explanatory variables, respectively the age and the wealth level at four months, are chosen for several reasons. They are readily available, they clearly influence customers' behavior and they can be integrated in the model without adding too much complexity. A cluster analysis is conducted with the age category and the wealth level of a customer, with the following variables: the weighted mean of the growth between two consecutive months in state 1, the weighted mean of the growth between two consecutive months in state 2, the weighted standard deviation of the growth between two consecutive months in state 1 and the weighted standard deviation of the growth between two consecutive months in state 2. This cluster analysis, using either the k-means method or the hierarchical clustering, highlights 3 or 4 clusters. The clusters are then integrated into the deposit volume, through the growth coefficients in a normal-rate environment and in a low-rate environment.

Three proposals are made for demand deposits modeling. The first one is a basic one, that takes into account rates' conjecture but not the customers' features. Proposals 2 and 3, by including a cluster analysis and by being an extension of proposal 1, take into account both rates' conjecture and the customers' features. Proposal 2 uses a hierarchical clustering method while proposal 3 includes a k-means algorithm. In Chapter 5, the final average life and the optionality of the three proposals are compared. The best-case scenario for a bank corresponds to a high average life associated with a low optionality. Overall, proposals 2 and 3 are definitely optimal for the bank, compared to proposal 1. As we can see in subsections 5.2.3 and 5.2.4, the method used for the cluster analysis will have little impact on the final outcomes. However, the Silhouette index, presented in B.2, is an internal cluster validation measures and shows that the clustering is of better quality with the K-means approach. The global Silhouette index is higher with the K-means approach, and more specifically, all the data points have a positive Silhouette index with this clustering method. On the contrary, in proposal 2, that uses a agglomerative hierarchical clustering, some points have a negative Silhouette index, meaning that these points were misclassified.

6.2. Recommendations

In this thesis, we proposed two different methods, assuming proposals 2 and 3 can be summarized in one proposal as they give extremely similar results, to model demand deposits: a first one that takes into account the context of low rates that has hit over the last few years, and a second one that tries to take into account both the context of low rates and the bank's features. Not surprisingly, the results with the second method (proposals 2 and 3) are more optimal for the bank. However, some points deserve further attention.

In section 2.5, two approaches are presented to move from a dynamic scenario, to a static scenario. We chose to use the easiest one, that only focuses on decreases in deposit volumes, but this approach might be too drastic and less justifiable from a theoretical point of view. The other one focuses on decreases in deposit volumes and on the attrition rates, and is more conservative. Additional analysis on the attrition rates could be done to use this approach instead of the one currently used.

Moreover, the customer segmentation in 10 age categories and 10 wealth levels was quite arbitrary.

We chose 10 categories for both explanatory variables so that the volume of data in the model is not unmanageable. However, the method used for the composition of the segments was limited by the internal data retrieval software. Indeed, we had to make a first choice by selecting about 80 wealth categories in the software, and then reduced those 80 categories to 10 using the Solver Add-in on Excel. As this software is about to change in the bank, further analysis could be conducted, especially on the wealthier customers. Similarly, other explanatory variables could be tested in the cluster analysis. This point also depends on the internal software used by the bank, and on the variables that are available.

A

Proofs for Vasicek model

A.1. Zero-coupon bond price for a Vasicek model

We recall proposition 1:

Proposition 4. *Zero-coupon bond price. The zero-coupon bond price at time t with maturity T can be written as:*

$$P(t,T) = A(t,T)\,\mathrm{e}^{-B(t,T)r(t)}, \tag{A.1}$$

where

$$\begin{cases} B(t,T) &= \frac{1-\mathrm{e}^{-k(T-t)}}{k} \\ A(t,T) &= \exp\Big[\big(b-\frac{\sigma^2}{2k^2}\big)\big(B(t,T)-T+t\big)-\frac{\sigma^2}{4k}B^2(t,T)\Big], \end{cases} \tag{A.2}$$

Proof. Using the Feymann-Kac theorem and under the risk-neutral measure **Q**, the zero-coupon bond price at time t with maturity T is defined as:

$$P(t,T) = \mathrm{E}\big(\mathrm{e}^{-\int_t^T r_s \mathrm{d}s}\,|\mathcal{F}_t\big) \tag{A.3}$$

From equation (2.9), we get that:

$$\int_t^T r_s \mathrm{d}s = \int_t^T \big(r_t\,\mathrm{e}^{-k(s-t)} + b(1-\mathrm{e}^{-k(s-t)}) + \int_t^s \sigma\,\mathrm{e}^{-k(s-u)}\,\mathrm{d}W_u\big)\mathrm{d}s \tag{A.4}$$

$$= \frac{r_t}{k}(1-\mathrm{e}^{-k(T-t)}) + b(T-t) - \frac{b}{k}(1-\mathrm{e}^{-k(T-t)}) \tag{A.5}$$

$$+ \int_t^T \Big(\int_u^T \mathrm{e}^{-k(s-u)}\,\mathrm{d}s\Big)\mathrm{d}W_u, \text{ by interchanging the two integrals} \tag{A.6}$$

$$= \frac{r_t - b}{k}(1-\mathrm{e}^{-k(T-t)}) + b(T-t) + \int_t^T \frac{\sigma}{k}(1-\mathrm{e}^{-k(T-u)}\,\mathrm{d}W_u. \tag{A.7}$$

Assume that $P(t,T)$ can be written as:

$$P(t,T) = A(t,T)\exp(-B(t,T)r(t)), \tag{A.8}$$

where $A(t,T)$ and $B(t,T)$ are two functions to be determined.

We start by setting

$$B(t,T) = \frac{1}{k}\big(1-\mathrm{e}^{-k(T-t)}\big) \tag{A.9}$$

Then,

$$
\begin{aligned}
P(t,T) &= \mathrm{E}\Big(\exp\big(-B(t,T)r(t) + b(B(t,T)-T+t) + \int_t^T \sigma B(u,T)\mathrm{d}W_u|\mathcal{F}_u\big)\Big) && \text{(A.10)}\\
&= \exp\big(-B(t,T)r(t)\big)\exp\big(b(B(t,T)-T+t)\big)\exp\Big(\frac{1}{2}\int_t^T \sigma^2 B^2(u,T)\mathrm{d}u\Big) && \text{(A.11)}\\
&= \exp\big(-B(t,T)r(t)\big)\exp\big(b(B(t,T)-T+t) + \frac{\sigma^2}{k^2}(T-t-B(t,T)) - \frac{\sigma^2}{4k}B^2(t,T)\big) && \text{(A.12)}\\
&\quad \text{, from equation (2.23)} && \text{(A.13)}
\end{aligned}
$$

Eventually, we have

$$
P(t,T) = A(t,T)\exp\big(-B(t,T)r(t)\big), \tag{A.14}
$$

where

$$
\begin{cases} A(t,T) &= \exp\big(\big(b-\frac{\sigma^2}{k^2}\big)(B(t,T)-T+t) - \frac{\sigma^2}{4k}B^2(t,T)\big), \\ B(t,T) &= \frac{1}{k}\big(1-\mathrm{e}^{-k(T-t)}\big). \end{cases} \tag{A.15}
$$

□

A.2. Zero-coupon rates for a Vasicek model

Proposition 5. *Zero-coupon rates. The zero-coupon rate at time t is maturity T is of the following form*

$$
R(t,T) = \big(b - \frac{\sigma^2}{2k^2}\big) + \frac{1-\mathrm{e}^{-k(T-t)}}{k(T-t)}\big(r_t - \big(b-\frac{\sigma^2}{2k^2}\big)\big) + \frac{\sigma^2}{4k^3}(1-\mathrm{e}^{-k(T-t)})^2. \tag{A.16}
$$

Proof. This proof is very simple. From proposition 1, we can compute the zero-coupon rates as follows

$$
\begin{aligned}
R(t,T) &= -\frac{\ln P(t,T)}{T-t} && \text{(A.17)}\\
&= -\frac{\ln A(t,T)}{T-t} + B(t,T)\frac{r_t}{T-t} && \text{(A.18)}\\
&= -\frac{1}{T-t}\Big(b - \frac{\sigma^2}{2k^2}\big(B(t,T)-T+t\big) + \frac{1}{T-t}\frac{\sigma^2}{4k}B^2(t,T) + B(t,T)\frac{r_t}{T-t} && \text{(A.19)}\\
&= \frac{\sigma^2}{2k^2} + \frac{1-\mathrm{e}^{-k(T-t)}}{k(T-t)}\Big(r_t - \big(b-\frac{\sigma^2}{2k^2}\big)\Big) + \frac{\sigma^2}{4k^3}(1-\mathrm{e}^{-k(T-t)})^2. && \text{(A.20)}
\end{aligned}
$$

□

B

Additional results

B.1. Proposal 2 with a divisive cluster analysis

In chapter 5, only the results with a agglomerative cluster analysis were presented in proposal 2, as the agglomerative cluster analysis is much more used than the divisive cluster analysis in practice. The dendrogram in figure B.1 shows how the data points could be split under the divisive approach.

Choosing three clusters seems to be a good choice in the DIANA method, as the height between three clusters and four clusters is maximized. However, it will also mean that we will have one cluster of 4 points, one cluster of 5 points and one cluster with all the remaining points. From a practical point of view, it is not really an option for the bank. Then, we will choose to divide the data in 4 clusters with the DIANA method.

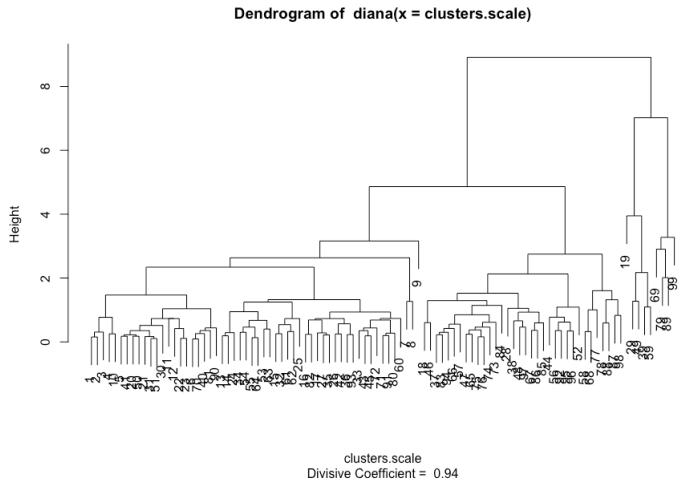

Figure B.1: Dendrogram under the agglomerative cluster analysis

Then, the different steps of the simulation of the deposits volume model are shown in figures B.2, B.3 and B.4.

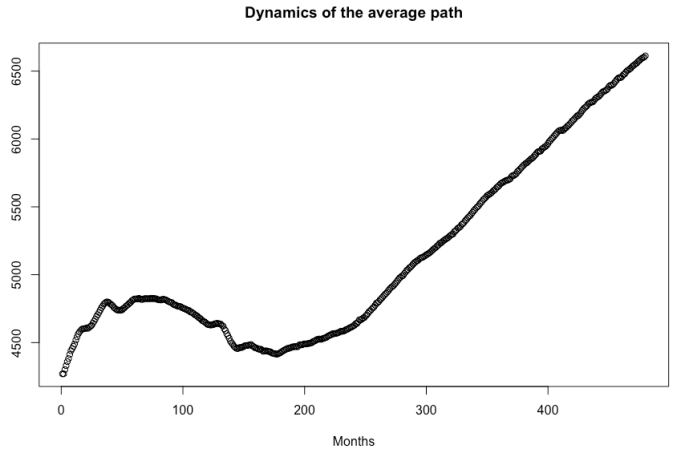

Figure B.2: Dynamics of the average path - DIANA method with 4 clusters

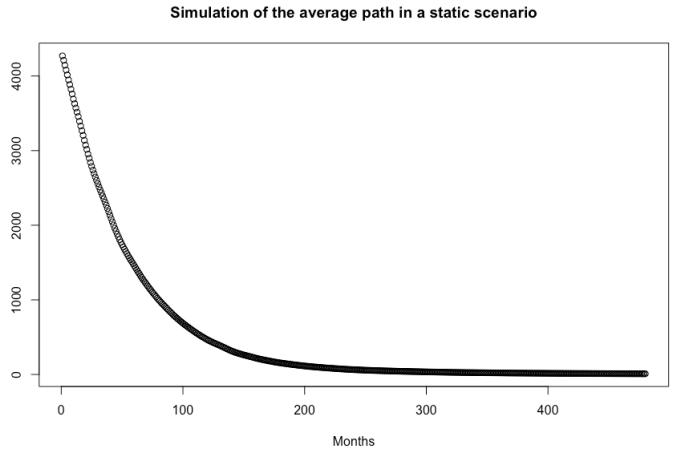

Figure B.3: Simulation of the average path in a static scenario - DIANA method with 4 clusters

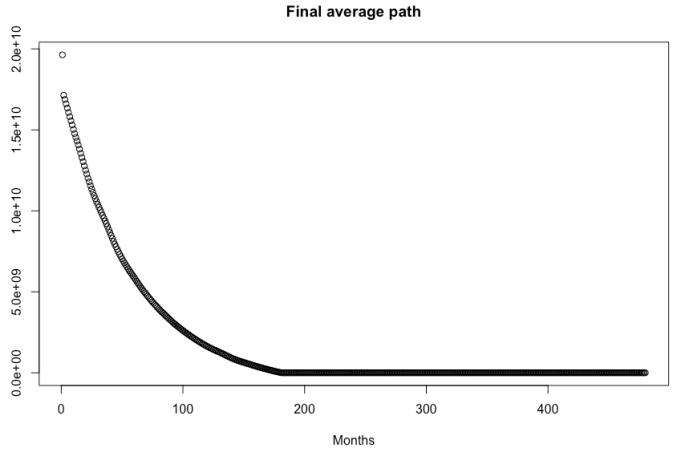

Figure B.4: Final average path - DIANA method with 4 clusters

The average life and the buffer are then given in table B.1. The results are indeed extremely close to the ones obtained with the agglomerative clustering and presented in table 5.8.

Dynamic scenario (stable part)	47.64 y
Static scenario (stable part)	4.70 y
Final average life	3.20 y
Buffer	23.5 %

Table B.1: Output of the model - DIANA method with 4 clusters

B.2. Internal cluster validation measures

Cluster validation measures are critically important in a cluster analysis, as clustering is an unsupervised learning algorithm. External and internal validation measures can be used to validate the goodness of partitions after a cluster analysis. External measures use external information but assume that we already have verified and true information about the ”optimal” clusters. In our case, we do not have verified external information so we will stick to internal validation measures. The internal measures usually aim to verify the compactness and the separation of the cluster analysis. The compactness measures ”how close are the objects within the same cluster”, and the separation measures ”how well-separated a cluster is from other clusters” [23]. Five different aspects are used to investigate the validation properties of an internal validation measures: monotonicity, noise, density, subclusters and skewed distribution. The internal validation measure used here, the Silhouette index, performs well in four out of five aspects. The Silhouette index can be defined for every data point X_i as:

$$S(X_i) = \frac{b(X_i) - a(X_i)}{\max(a(X_i), b(X_i))}, \tag{B.1}$$

where

$$a(X_i) = \frac{1}{n_i - 1} \sum_{X_j \in C_i, X_j \neq X_i} d(\mathbf{X_i}, \mathbf{X_j}), \tag{B.2}$$

$$b(X_i) = \min_{k, k \neq i} \Big[\frac{1}{n_k} \sum_{X_j \in C_k} d(\mathbf{X_i}, \mathbf{X_j}) \Big]. \tag{B.3}$$

In other words, if a point X_i belongs to a cluster C_i, then $a(X_i)$ will be the average distance of X_i with all the points in the same cluster, and $b(X_i)$ will be the average distance of X_i with all the points in the closest cluster C_j.

This value is always between -1 and 1 and should be interpreted as follows. The closer to 1 a Silhouette index will be, the better. If a Silhouette index is below 0 for a point, then it means that this point has been put in the wrong cluster.

We can also compute the Average Silhouette width for a cluster, or more generally for the clustering. The results for the three cluster analysis, the agglomerative algorithm with 3 clusters, the k-means method for 3 clusters and the divisive algorithm with 4 clusters are given in figures B.5, B.6 and B.7. As we can see, the clustering is of better quality with the k-means approach and with the divisive cluster algorithm. On the contrary, for the agglomerative hierarchical clustering, the Silhouette index for some points is below 0, which means that these points were put on the wrong cluster.

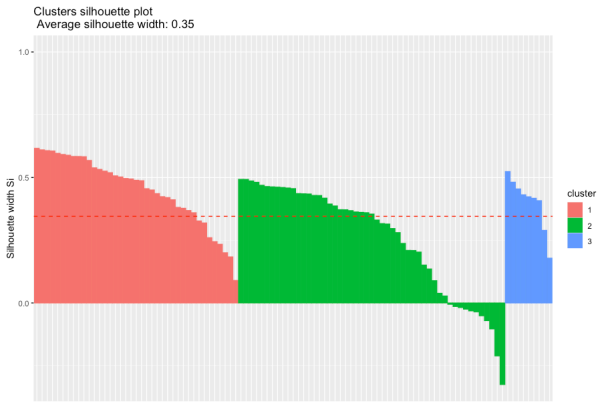

Figure B.5: Clusters Silhouette index plot - AGNES method with 3 clusters

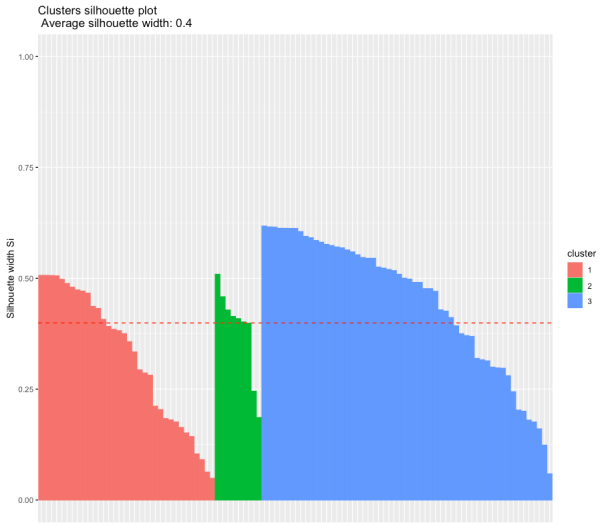

Figure B.6: Clusters Silhouette index plot - K-means method with 3 clusters

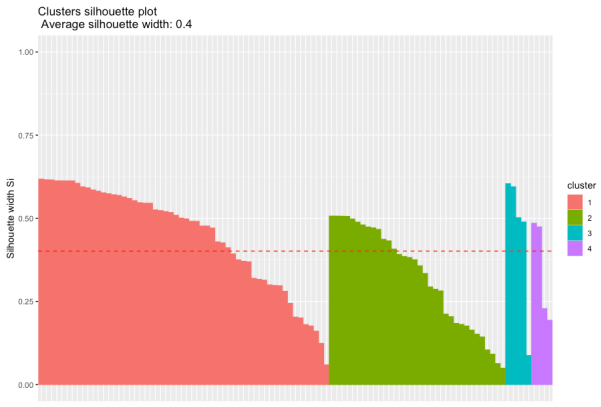

Figure B.7: Clusters Silhouette index plot - DIANA method with 4 clusters

Bibliography

[1] Horovitz A. and Szimayer A. Modeling non-maturing demand deposits: on the determination of the threshold of separation between volatile and stable deposit volumes. Conference paper, May 2020.

[2] Apache Software Foundation. Finastra. URL `https://www.finastra.com/solutions/treasury-capital-markets/risk-compliance/fusion-risk`.

[3] BCBS. *Principles for Sound Liquidity Risk Management and Supervision*. BIS, 2008. ISBN 92-9197-767-5.

[4] BCBS. *Basel III: The Liquidity Coverage Ratio and liquidity risk monitoring tools*. BIS, 2013. ISBN 92-9197- 912-0.

[5] BCBS. *Basel III: The Net Stable Funding Ratio*. BIS, 2014. ISBN 978-92-9131-960-2.

[6] BCBS. *Standards : Interest rate risk in the banking book*. BIS, 2016. ISBN 978-92-9197-498-6.

[7] Baumann C., Burton S., and Elliott G. Predicting consumer behavior in retail banking. *Journal of Business and Management*, 13(1):79–96, 2007.

[8] Cox J. C., Ingersoll J. E., and S. A. Ross. A theory of the term structure of interest rates. *Econometrica*, 53(2):385–407, 1985.

[9] Brigo D. and Mercurio F. *Interest Rate Models - Theory and Practice*. Springer Finance, 2 edition, 2006. ISBN 978-3-540-34604-3.

[10] Heath D., Jarrow R., and Morton A. Bond pricing and the term structure of interest rates: A discrete time approximation. *Journal of Financial and Quantitative Analysis*, 25(4):419–440, 1990.

[11] Hull J. and White A. Princing interest-rate-derivative securities. *The Review of Financial Studies*, 3(4):573–592, 1990.

[12] MacQueen J. Some methods for classification and analysis of multivariate observations. *Proc. Fifth Berkeley Symp. on Math. Statist. and Prob.*, 1:281–297, 1967.

[13] Ward J. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 301(58):236–244, 1963.

[14] Nyström K. On deposit volumes and the valuation of non-maturing liabilities. *Journal of Economic Dynamics and Control*, 32:709–756, 2008.

[15] Thorndike R. L. Who belongs in the family. *Psychometrika*, 18(4):267–276, 1953.

[16] Kalkbrener M. and Willing J. Risk management of non-maturing liabilities. *Journal of Banking and Finance*, 28:1547–1568, 2004.

[17] Mil'shtejn G. N. Approximate integration of stochastic differential equations. *Theory of Probability & Its Applications*, 19(3):557–562, 1995.

[18] Vasicek O. An equilibrium characterization of the term structure. *Journal of Financial Economics*, 5:177–188, 1977.

[19] Cai R., Dai. C, Hao Z., Tung. A. K. H., and Zhang Z. A general framework of hierarchical clustering and its applications. *Information Sciences*, 272:29–48, 2014.

[20] Jarrow R. and van Deventer D. The arbitrage-free valuation and hedging of demand deposits and credit card loans. *Journal of Banking and Finance*, 22:249–272, 1998.

[21] Tibshirani R., Walther G., and Hastie T. Estimating the number of clusters in a data set via the gap statistic. *J. R. Statist. Soc. B*, 63(2):411–423, 2001.

[22] Ostrovski V. Efficient and exact simulation of the hull-white model. SSRN Electronic Journal, 2013.

[23] Gao X., Li Z., Liu Y., Xiong H., and Wu J. Understanding of internal clustering validation measures. Conference paper, December 2010.

[24] Ho T. S. Y and Lee S.-B. Term structure movements and pricing interest rate contigent claim. *The Journal of Finance*, 41(5):1011–1029, 1986.