

FACULTY OF MATHEMATICS AND PHYSICS Charles University

MASTER THESIS

Bc. Kseniya Kuzminskaya

Acceleration of calculations in life insurance

Department of Probability and Mathematical Statistics

Supervisor of the master thesis: RNDr. Janeček Martin, Ph.D. Study programme: Mathematics Study branch: Financial and Insurance Mathematics

Prague 2018

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In Prague date 20.7.2018

signature of the author

Title: Acceleration of calculations in life insurance

Author: Bc. Kseniya Kuzminskaya

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Janeček Martin, Ph.D., Department of Statistics and Probability, VŠE

Abstract:

One of the major issue for life insurance companies is proper and consistent valuation of liabilities. This thesis introduces the standard estimation methods used in practice and discussed the alternative methods, which might help to speed up these calculations. It studies two possible methods of acceleration of calculations in life insurance: analytic function and cluster analysis. The outcome of these work is comparison of discussed methods applied on generated life insurance portfolio. All methods were applied on two possible insurance products. Comparison of the results is based on the calculation precision and time needed to process the liabilities of the insurance company's portfolio.

Keywords: life insurance, stochastic interest rate scenarios, cash flow calculation, cluster analysis, analytic function

I would like to thank my supervisor RNDr. Janeček Martin, Ph.D. for his valuable advice during my studies and dedication to helping me succeed. Thank you to my consultants RNDr. Branda Martin, Ph.D. and Dr. Seabstiano Vitali, Ph.D. for inspiration and knowledge.

Contents

In	Introduction		
1	Valu 1.1 1.2 1.3 1.4 1.5	uation of Liabilities in Insurance CompaniesMain principles of valuationOptions and Guarantees in Life InsuranceFair Value approachMonte Carlo simulationsAssumptions used in best estimate valuation1.5.1Non-financial assumptions1.5.2Financial AssumptionsProjection of cash flows in insurance company1.6.1Definition of cash flows in projection	3 3 4 5 6 8 8 9
2	Acc 2.1 2.2	eleration techniquesAnalytic function2.1.1Death Benefit as a summation of fund value and sum assured2.1.2Death Benefit as Maximum of Fund Value and Sum AssuredCluster analysis2.2.1Data Preparation2.2.2Distance Measures2.2.3Clustering methods	 14 15 19 25 25 25 27
3	Inte 3.1	erest rate scenarios Models of interest rates	30 30
4	Imp 4.1 4.2 4.3	Dementation Interest rate models	32 35 35 37 39 40 43
Co	onclu	ision	47
Bi	bliog	graphy	49
\mathbf{Li}	st of	Tables	51
Aj	ppen	dix	53

Introduction

Proper valuation of liabilities in insurance companies is the key, not only from the view of good risk management of insurance company, but also it is strongly emphasized by the supervisor. Under Solvency II legislative, which came into effect at the beginning of 2016, they paid specific attention to the correct valuation of liabilities which is included in internal and external reporting.

The standard technique for valuating the liabilities is risk-neutral Monte Carlo valuation, where the development of liabilities is simulated for a large number of investment return scenarios. Currently, insurance companies consume a significant amount of time to process liabilities and cash flows. This thesis discusses the techniques of accelerating the valuation of these processes. The aim of this work is to introduce and test two possible approaches how to calculate the future cash flow of the company faster under many interest rate scenarios, with reasonable error.

This thesis will show the results of the implementation of standard cash flow calculations compared to the usage of analytic function and cluster analysis. We will simulate the interest rate scenarios using a Hull-White model. We want to show the result comparison between the standard calculation technique, analytic function and cluster analysis. We will simulate the rates and a sample of life insurance company's portfolio to show the different time achievements between each technique.

The thesis is consisted of four chapters. The first chapter introduces us the main principles of valuation of cash flows in insurance companies. We get acquainted with the basic insurance principle and with the typical future projection of cash inflows and outflows in life insurance company. It explaines the processes of calculation of best estimate liability. In this part we also discuss the assupption used in calculation of best estimate of liabilities.

Second chapter insroduces the acceleration techniques tested in this thesis. It presents the method of analytic function and cluster analysis. In this chapter we are explaining the usage of analytic function based on two types of insurance products and introducing two methods of cluster analysis .

Third chapter introduces the theory for simulation of interest rates. It briefly describes the Hull-White interest rate model.

Final forth chapter introduces the implementation of previous chapters in practice. It compares the obtained results for two types of insurance profucts using each of discussed methods. Comparison is based on the accuracy of the result and time needed to process the liabilities.

This thesis was implemented in Wolfram Mathematica Software.

1. Valuation of Liabilities in Insurance Companies

1.1 Main principles of valuation

Life insurance is a form of insurance in which a person makes payments to an insurance company, in return for a sum of money to be paid to him/her after a period of time, or to his/her family if he/she dies or survives.

Usually, life insurance companies offer two main types of products, one is *traditional*, that provides guaranteed financnial coverage, and *unit-link* product. In the scope of this thesis, we will consider unit-link products. Unit-link insurance combines elements of term life insurance with an investment savings option. It "links" the policyholder's benefit to some financial index or fund. Premiums within the unit-link insurance policy are broken down by the insurance and saving component. Savings premiums are continuously invested in the underlying assets and the benefits might be different according to its performance. To attract more people to invest in such products, the insurance companies usually offer a *guaranteed minimum interest rate*, which is also defined as *technical interest rate* (TIR).

1.2 Options and Guarantees in Life Insurance

Holders of unit-link policies have accounts in insurance companies, which is termed as *fund value*. The insurance companies invest money received from policyholders and in exchange, every year they credit policyholder's accounts with some investment return. This investment return is a random element and it heavily depends on market performance. In the case of low investment return, the contract requires insurance companies to pay the guaranteed interest rate from their resources. If the investment return is higher than guranteed interest rate, the difference comes to a so called *profit share*. Part of profit share is turned to provision that will cover the future undesirable investment performance. Another part of profit share is paid to policyholders as a profit.

The guranteed return or technical interest rate might be different for each type of product. Within this work, we will assume the same technical interest rate for all products and policyholders. We will assume technical interest rate to be equal to 2.1%.

We can write the final payoff of investment return that would be assigned to poliholder's account as follows:

$$InvReturnPayoff = TIR + \max\{0, i - TIR\},\$$

where i stands for market return of investment.

Such a defined financial instrument is called *financial option* in insurance. We can compare it with formula of payoff for call option used in finance. Policyholders can profit in future situations where the investment return is higher than the

guaranteed interest rate. Figure 1.1 shows the policyholders' payoff of investment return.



Figure 1.1: Investment return of financial option in life insurance

1.3 Fair Value approach

A proper and consistent valuation of options and guarantees is vitally important for insurance companies, not only from good risk managment point of view, but also for internal and supervisor reports. One of the main examples of such reports is Solvency II legislative. It defines the amount of capital that European insurance companies must hold to be solvent.

Solvency II defines the fair value of insurance company's liabilities as a sum of a *best estimate liability* (BEL) and a *risk margin* ([1] Art.77):

$$FV_0 = BEL_0 + RM, (1.1)$$

where

 FV_0 fair value of liabilities at present time, BEL_0 best estimate liability at present time, RM risk margin.

Risk margin from the formula 1.1 can be calculated for example, within the Solvency II legislative as defined in Article 37 in [2]. Further, we will focus mainly on calculation of best estimate liability. Valuation of options and guarantees is included in the value of BEL.

Under Solvency II, the best estimate liability is defined as "probability weighted average of future cash-flows, taking account of the time value of money (expected present value of future cash-flows), using the relevant risk-free interest rate term structure" ([1] Art.77). Formula 1.2 shows the present value of future cash flow projection at the moment of valuation (time 0).

$$PVCF_0 = \sum_{t=1}^{\infty} \frac{CF_t}{(1+rfr_t)^t},$$
(1.2)

where

$PVCF_0$	present value of future cash flows at present time
CF_t	cash flow at the end of the projection year t ,
rfr_t	risk free interest rate related to projection year t ,
t	projection year.

Projection of future cash flows is calculated for all policies that are in-force at the moment of valuation. We don't consider any future new business of the insurance company. So, cash flows are projected till the maximum length of all policy periods: $T = \max(n_1, \ldots, n_J)$, where J is the total number of policies in-force, and n_1, \ldots, n_J are periods of each policy. In formula 1.2 we assume the total years of projection T to be infinity.

Present value of cash flows projection in formula 1.2 equals to the sum of all discounted future cash flow calculated in risk-neutral world. In a risk-neutral world, each individual is indifferent to risk and therefore expects to gain a return for all investment categories equal to the risk-free interest rate [3].

1.4 Monte Carlo simulations

The standard technique for valuating options and guarantees, and also for best estimate liability is *Monte Carlo* simualations. Monte Carlo algorithm relies on repeated random samplings to obtain numerical results. For example, to obtain the best estimate of liabilities, we need to calculate many random cash flow projections then calculate present value and average the obtained values. Random cash flows projection refers to random possible scenarios of risk-free interest rate. We can write the estimation of best estimate as follows:

$$BEL_0 = \mathbb{E}\{PVCF_0\}$$

We estimate the expected value of cash flows at the present time as an average of present values of cash flows under the interest rate scenario $s = \{1, \ldots, S\}$

$$\mathbb{E}\{PVCF_0\} = \overline{PVCF}_{0,S} = \frac{1}{S}\sum_{s=1}^{S}PVCF_{0,s}$$

where S is total number of scenarios, and $PVCF_{0,s}$ means the present value of cash flows at present time calculated under the interest rate scenario s.

The aim is to make so many scenarios S that the difference between the average of present value of cash flow for S and S + 1 scenarios is insignificant.

$$S: \lim_{S \to \infty} (\overline{PVCF}_{0,S+1} - \overline{PVCF}_{0,S}) = 0.$$

Usually insurance companies make from 500 to 1000 scenarios of interest rates to obtain the stable result.

Figure 1.2 shows the idea of best estimate liability calcualtion using the method of Monte Carlo.



Figure 1.2: BEL calculation using Monte Carlo simulations

Gray line shown in Figure 1.2 is an average of all S scenarios used in simulation and is equal to BEL_0 . The red line shows the average of exactly $s = \{1, \ldots, S\}$ scenarios and is calculated by formula:

$$BEL_{0,s} = \overline{PVCF}_{0,s} = \frac{1}{s} \sum_{i=1}^{s} PVCF_{0,i}, \quad s = 1, \dots, S$$

To perform an estimation of best estimate in this way, with large amount of scenarios, takes an extreme amount of calculation time for insurance companies. However, some advanced computers could help and reduce the calculation time, but it is not always possible to perform such calculations within reasonable time. In this thesis, we want to show two of many possible techniques that might help to accelerate the cash-flow calculation for life insurance companies.

1.5 Assumptions used in best estimate valuation

1.5.1 Non-financial assumptions

In this section we will discuss the issues for the assumptions used in calculation of liabilities. All assumptions used for the cash flow projection are to be on the best estimate level, which is understood to be their expected value.

Mortality

Underlying expected mortality assumptions are based on mortality tables, which are statistical tables of expected annual mortality rates.

The insurance companies should use their past years of experiences when creating mortality assumptions. Mortality experience tables might be split according to sex and age of the insured person, as well as smoker status, type of policy, etc [4]. Information gathered on an individual is taken into account for insurance calculations. A selected mortality table includes mortality data on individuals who have recently purchased life insurance. These individuals tend to have lower mortality rates than individuals who are already insured, due to the fact that they have most likely just passed certain medical exams required to obtain insurance. We will assume such adjustments in mortality tables (expected mortality) for our cash-flow model defined below.

Lapses

Lapses are the cancellation of the policy and it can be an important component for the pricing of long term cash flows. The policyholder is allowed to cancel his policy at any time. As well as mortality experience, the companies should take into account their recent and reliable experience of lapse development.

In case of policy lapse, the insurance company returns the fund value deducted by some surrender fee to the policyholder.

Lapses analysis is usually built according to the policy year of insurance, type of product or calendar year of the policy inception [4].

Commissions

Commissions are usually based upon the size of the policy the agent is selling (means the size of annual premiums) and by the type of product.

Within this thesis, we will use two forms of commission payments: initial and renewal. In case of regular premium payments we will use initial and renewal commissions payments. In case of single premium payment, there will be only one commission payment.

Usually, the initial commissions payment is a payment that is equal to a percentage of the total annual premium that will be made to the policy during the first policy year.

A renewal commission (in the case of regular premium payments) is a commission paid for a specific number of years after the first policy year. The number of years that a renewal is paid vary between the companies, but frequently it is a significant number of years.

There can be claw back commission, which allows companies to return some amount of money back from the agents due to the withdrawal by the insurer of the policy agreement. It is usually concerned with the initial commissions during the first years.

Expenses

General and administrative expenses typically refer to policies, regardless of whether the company produces or sells anything. Examples of expenses can be product advertisement, salaries, building rent etc.

The expenses can be split into initial and renewal expenses just like commissions.

Initial commissions usually exist during the closure of the life insurance contract. Examples of such expenses can be the initial medical treatments or product advertisement. The renewal expenses exist during the life time of the policy. Possible examples of such expenses can be building rent or salaries to insurance company's employees.

The expenses can be a fixed amount or calculated as a percentage of the sum assured or the premium. The expenses increase over time due to inflation. We will consider the increase in expenses as well.

1.5.2 Financial Assumptions

In order to calculate the liabilities of cash flows in risk-neutral world, the risk-free rate is used. Risk-free rate is the theoretical rate of return of an investment with zero risk. Often for such a rate, the return yield of government bonds is used.

Usually, there are two rates used in cash flow projection: one for discount of cash flows and another for investment return on assets (based on which the profit share is distributed). Within this work we will assume that the insurance company invests in risk-free assets, and for further calculation, we will use one rate for evaluation of the policyholder's fund value and for discounting of cash flows.

1.6 Projection of cash flows in insurance company

In this section we will define the cash flow projection for life insurance company.

The main idea of any cash flow valuation is the simple principle of income amounts minus outcome amounts at the end of the year. Let's assume that an insurance company has J policies in their portfolio. The income for insurance company is premium that it gets from policyholders. We assume the insurance company collects the premium $P_t^{(j)}$ from some policy $j = \{1, \ldots, J\}$ at the beginning of the projection year t. Further, we assume the insurance company pays the commissions to agents $C_t^{(j)}$ and expenses $E_t^{(j)}$, both at the beginning of the year t for some policy j. And finally, we assume the insurance company pays the benefits in case of jth policyholder's death resp. maturity at the end of the year t. We will denote these amount as $Dths_t^{(j)}$ resp. $Mat_t^{(j)}$. Also we assume that in case of lapse the company pays the agreed surrender amount $Surr_t^{(j)}$ at the end of the year t for some policy j. All these cash flow amounts are derived including the probabilities to be happen every year.

Formula 1.3 shows the cash flows projection of one policy.

$$CF_t^{(j)} = (P_{t-1}^{(j)} - C_{t-1}^{(j)} - E_{t-1}^{(j)})(1 + rfr_t) - Dths_t^{(j)} - Mat_t^{(j)} - Surr_t^{(j)}, \quad (1.3)$$

To valuate the cash flow projection, we will total the projections for all policies in the portfolio. Formula 1.4 shows the valuation of annual cash flows of all policies used in calculation by the end of the year t [4].

$$CF_{t} = \sum_{j=1}^{J} CF_{t}^{(j)} =$$

$$= \sum_{j=1}^{J} (P_{t-1}^{(j)} - C_{t-1}^{(j)} - E_{t-1}^{(j)})(1 + rfr_{t}) - Dths_{t}^{(j)} - Mat_{t}^{(j)} - Surr_{t}^{(j)} =$$

$$= (P_{t-1} - C_{t-1} - E_{t-1})(1 + rfr_{t}) - Dths_{t} - Mat_{t} - Surr_{t},$$
(1.4)

where

CF_t	cash flow at the end of the projection year t ,
P_{t-1}	probability-weighted premium income at the beginning of the year t ,
C_{t-1}	probability-weighted commisions paid to agents at the beginning
	of the year t .

 E_{t-1} probability-weighted expenses paid at the beginning of the year t,

 rfr_t risk free interest rate related to projection year t,

- $Dths_t$ outflow representing the death benefit assumed to be paid at the end of the projection year,
- Mat_t outflow representing the maturity benefits assumed to be paid at the end of projection year,
- $Surr_t$ outflow representing surrenders assumed to be paid at the end of the projection year.

1.6.1 Definition of cash flows in projection

It is important to note the difference between the projection and policy year in our cash flow calculation. Projection year is a year of our future projection, that means the projection year 0 is a moment of calculation and the total year of projection was defined in Formula 1.2 as infinity. Policy year, which we will denote as τ , is the year of policy existing. Depending on the year of one policy existence, the amounts of cash flows might differ. For example, a policy with single premium payment has premium inflow paid for the first policy year only. It is equal to zero, if the policy year doesn't equal to projection year. The size of commissions and expenses outcome also depends on the one's policy year. The policy year of some contract equals to projection year, when one signs the insurance contract at the year of valuation (before the valuation date). The cash flow CF_t is then defined as the sum of all probability-weighted cash flows for all policies in-force $j = 1, \ldots, J$ that occure in a projection year τ .

Every projection year there are probabilities of all cash flows to happen. For futher defining of cash flows probability we will use the following notation ([4], [5]):

x	age of a policy holder at the projection year t				
\mathbf{T}_x	the remaining number of life years at the age x				
q_x	probability that a person who is alive at the age of x will die				
	before the age $x + 1$				
	$q_x = \mathcal{P}(\mathbf{T}_x \le 1)$				
p_x	probability that a person who is alive at the age of x will be alive				
	at the age of $x + 1$				
	$p_x = \mathcal{P}(\mathbf{T}_x > 1)$				
$_t p_x$	probability that a person who is alive at the age of x will be alive				
	at the age of $x + t$				
	$_{t}p_{x} = \mathcal{P}(\mathbf{T}_{x} > t)$				
$q_{x,\tau}^{exp}$	expected mortality; adjusted probability that a person who is				
,	alive at the age of x will die before the age $x + 1$				
	$q_{x,\tau}^{exp} = coef_{\tau} \cdot q_x$, where				
	$coef_{\tau}$ is mortality adjustmet depending on the policy year τ				
$wthd_t$	probability of lapse at the policy year t				
ℓ_t	expected number of policies in-force at the end of the projection				
	year t				
	$\ell_t = \ell_{t-1} - d_t - m_t - w_t$				
d_t	expected number of deaths at the end of the projection year t				
	$d_t = \ell_{t-1} \cdot q_{x,\tau}^{exp}$				
m_t	expected number of maturities at the end of the projection year t				
	$m = \int 0,$ if $\tau < n$				
	$m_t = \int \ell_{t-1} - d_t - w_t$, if $\tau = n$				
w_t	expected number of lapses at the end of the year t				
	$w_t = (\ell_{t-1} - d_t) \cdot wthd_t$				

Insurance companies might offer a large variety of unit-link products. Depending on the type of policyholder's contract, the outflow payments in case of death or maturity defined in Formula 1.4 might be different. Also products can differ depending on frequencey or method of premium payments.

We will consider two types of unit-link insurance products according to their death benefit. For the first product, the death benefit will be a value of sum assured plus the policyholder's fund value at the end of the year of the occurance. Sum assured is amount paid by the contract in case of event occurence, in our case, it is a death of policyholder. For the second type of product, we assume that the death benefit is the maximum value of sum assured and policyholder's fund value within the year of payment. Later on, we will use the notation SA for sum assured amount, and CV_t will denote the fund or capital value at the year t.

Further, we will define the cash flows per one policy without probablity assumption

Premium

As we mentioned in section 1.1 the premium in unit-link is divided into saving and risk component. Also the gross premium, that the insurance companies get from the policyholders, contains the amount for coverage of administrative expenses $\alpha^{(j)}, \beta^{(j)}, \gamma^{(j)}$ (see [5]):

$$prem_t^{(j)} = SP_t^{(j)} + RP_t^{(j)} + \alpha^{(j)} + \beta^{(j)} + \gamma^{(j)},$$

where $SP_t^{(j)}$ is saving part of premium of some policy j at the year t; $RP_t^{(j)}$ is risk part of premium of some policy j at the year t; and $prem_t$ is premium per one policy at the year t.

Risk part of the premium is intended to cover the risk of death. We wil define the risk premium as follows:

$$RP_t^{(j)} = \frac{SAR_{p,t}^{(j)} \cdot q_x^{(j)}}{1 + TIR},$$

where $SAR_{p,t}^{(j)}$ is sum at risk per one policy and is equal to the difference between the death benefit and fund value $SAR_{p,t}^{(j)} = Benefit_{p,t}^{(j)} - CV_t^{(j)}$, and p means a type of product p = 1, 2.

For the first type of product, where the death benefit is paid as a summation of fund value and sum assured, it is equal to $SAR_{1,t}^{(j)} = SA^{(j)}$. For the type of products, where the death benefit is a maximum amount of sum assured and policyholder's fund value, it is equal to $SAR_{2,t}^{(j)} = [SA^{(j)} - CV_t^{(j)}]_+$.

We will get the saving part of policyholder j's premium by deducting the risk component and expenses $\alpha^{(j)}, \beta^{(j)}, \gamma^{(j)}$ (see [5]) from the premium:

$$SP_t^{(j)} = prem_t^{(j)} - \alpha^{(j)} - \beta^{(j)} - \gamma^{(j)} - RP_t^{(j)}.$$

We can distinguish the insurance products according to premium payment method and frequencies. Insurance products might be with single or regular payments. For simplicity we assume that regular premium payments are on a yearly basis only.

Commissions

We distinguish the initial commission at the first year of policy existence, and renewal commissions, which is paid regularly during the whole policy period.

$$\begin{split} \tau &= 1: comm_t^{(j)} = SA^{(j)} \cdot InitCommSA\%^{(j)} + prem_t^{(j)} \cdot InitCommP\%^{(j)}, \\ \tau &\geq 1: comm_t^{(j)} = SA^{(j)} \cdot RenCommSA\%^{(j)} + prem_t^{(j)} \cdot RenCommP\%^{(j)}, \end{split}$$

where

$comm_t^{(j)}$	is commission per one policy at the year t ,
$InitCommSA\%^{(j)}$	initial commission per one policy as a percentage of a
	sum assured,
$InitCommP\%^{(j)}$	initial commission per one policy as a percentage of a
	premium,
$RenCommSA\%^{(j)}$	renewal commission per one policy as a percentage of
	sum assured,
$renCommP\%^{(j)}$	renewal commission per one policy as a percentage of
	premium,
$SA^{(j)}$	sum assured per one policy,
$prem_t^{(j)}$	premium per one policy at the year t .

Expenses

We distinguish the initial and renewal expenses, and we also assume the increase of expenses in time at least due to inflation. In our example, we assume the expenses outflows to be calculated as follows:

$$\begin{aligned} \tau &= 1: exp_t^{(j)} = InitFixExp^{(j)} + InitExpP\%^{(j)} \cdot prem_t^{(j)} + \\ &+ RenFixExp^{(j)}(1 + Infl)^{(t-1)} + RenExpP\%^{(j)} \cdot prem_t^{(j)}; \\ \tau &\geq 1: exp_t^{(j)} = RenFixExp^{(j)}(1 + Infl)^{(t-1)} + RenExpP\%^{(j)} \cdot prem_t^{(j)}; \end{aligned}$$

where

$exp_t^{(j)}$	total expenses per one policy at the year t ;
$InitFixExp^{(j)}$	fix amount of initial expenses per one policy;
$InitExpP\%^{(j)}$	initial expenses as a percentage of the policyholder's
	premium;
$RenFixExp^{(j)}$	fix amount of renewal expenses per one policy;
$RenExpP\%^{(j)}$	renewal expenses as a percentage of the policyholder's
	premium;
Infl	expense inflation

Capital Value

The savings part of the premium is continuously invested in by insurance company and is becoming a policyholder's fund value. We assume that in the first policy year, the policyholder's fund value is zero $(CV_t^{(j)} = 0)$. For year $t \ge 1$ it is equal to

$$CV_t^{(j)} = (CV_{t-1}^{(j)} + SP_t^{(j)}) \cdot (1 + max\{i_t, TIR\}),$$
(1.5)

where i_t stands for the return of investment in year t.

As we agreed before in Section 1.5.2 we assume for simplicity the investment return and risk-free rate used for discount to be equal. Further, in the text we will use the notation i as possible scenarios of risk-free interest rate and return from company's investments.

Surrender payment

We assume in our example the existence of surrender period. Before that period it is forbidden to cancel the contract and surrender payment is equal to zero. After the surrender period, in case of lapse, the insurance company pays the policyholder's fund value decreased by surrender charge.

$$\tau \leq SurrPeriod : Surr_t^{(j)} = 0,$$

$$\tau > SurrPeriod : Surr_t^{(j)} = CV_t^{(j)} \cdot (1 - fee_t^{(j)}),$$

where

$fee_t^{(j)}$	surrender charge (as a percentage from fund value) per one
	policy applied when the surrender is paid assumed to be at
	the end of the year t ,
SurrPeriod	surrender period.

Benefit payments

As we mentioned before we consider two types of products with aggregate amount of sum assured and capital value

$$Benefit_{1,t}^{(j)} = SA^{(j)} + CV_t^{(j)}$$

and as a maximum of these two values

$$Benefit_{2,t}^{(j)} = \max(CV_t^{(j)}, SA^{(j)})$$

For both type of products, we assume the benefit in case of maturity to be equal to the policyholder's fund value $CV_t^{(j)}$.

Adding the probability weight

Table 1.1 shows the defined probability-weighted cash-flows from formula 1.4 for some policy j.

Projection			
year			
1	Premium	Income	$P_0^{(j)} = \ell_0^{(j)} \cdot prem_0^{(j)}$
	Comm. & Exp.	Outcome	$(E_0^{(j)} + C_0^{(j)}) = \ell_0^{(j)} \cdot (comm_0^{(j)} + exp_0^{(j)})$
	Death Benefit	Outcome	$Dths_1^{(j)} = d_1^{(j)} \cdot (Benefit_{i,t}^{(j)})$
	Maturity Benefit	Outcome	$Mat_1^{(j)} = m_1^{(j)} \cdot \mathrm{CV}_1^{(j)}$
	Surrender	Outcome	$Surr_1^{(j)} = w_1^{(j)} \cdot CV_1^{(j)}(1 - fee_1^{(j)})$
÷			:
÷			÷
t	Premium	Income	$P_{t-1}^{(j)} = \ell_{t-1}^{(j)} \cdot prem_{t-1}^{(j)}$
	Comm. & Exp.	Outcome	$(E_{t-1}^{(j)} + C_{t-1}^{(j)}) = \ell_{t-1}^{(j)} \cdot (comm_{t-1}^{(j)} + exp_{t-1}^{(j)})$
	Death Benefit	Outcome	$Dths_t^{(j)} = d_t^{(j)} \cdot (Benefit_{i,t}^{(j)})$
	Maturity Benefit	Outcome	$Mat_n^{(j)} = m_t^{(j)} \cdot \mathrm{CV}_t^{(j)}$
	Surrender	Outcome	$Surr_t^{(j)} = w_t^{(j)} \cdot CV_t^{(j)}(1 - fee_t^{(j)})$
÷			:
:			:

Table 1.1: Cash flow of Endowment policy

2. Acceleration techniques

In the section above, we have defined the cash flows of the insurance company. The total cash flows is calculated as projection of each policy in the portfolio and then it is summarized. Usually, to calculate the present value of insurance company's cash flows by policy-by-policy approach take an extreme amount of time.

In this section we will discuss two possible methods to speedup cash flows calculation: the analytic function and cluster analysis. The method of analytic function is based on the partition of the formula of cash flow projection into parts that depend and do not depend on the return of investment. We will discuss two types of possible insurance products according to their death benefit:

- 1. Death benefit is paid as an amount of sum assured and fund value;
- 2. Death benefit is paid as maximum amount of sum assured and fund value.

The method of cluster analysis relies on reduction of number of policies needed to be processed; this method doesn't depend on the type of product. We will describe two possible techniques of cluster analysis and we will also try to speed up the calculation for two types of products as in analytic function.

2.1 Analytic function

In the previous chapter we have defined the basic cash flow model for life insurance company (Formula 1.4 and Table 1.1). Furhermore, we won't distinguish if the cash-flow was in the beginning or at the end of year t. It is a common practice for insurance companies on the market. We will assume that all cash flows occur at the end of the year. Formula 2.1 shows the cash flow for all policies in life insurance company's portfolio.

$$CF_t = \sum_{j=1}^{J} P_t^{(j)} - C_t^{(j)} - E_t^{(j)} - Dths_t^{(j)} - Mat_t^{(j)} - Surr_t^{(j)}, \qquad (2.1)$$

where

J	total number of policies in-force used in projection,
CF_t	cash flow at the end of the policy year t for the whole life insurance portfolio,
$P_t^{(j)}$	probability-weighted premium income of some policy j paid at policy year t ,
$C_t^{(j)}$	probability-weighted commissions of some policy j paid at
()	poncy year t,
$E_t^{(j)}$	probability-weighted expenses of some policy j paid at policy
	year t ,
$Dths_t^{(j)}$	outflow representing the death benefit for some policy j
	assumed to be paid at the policy year t ,
$Mat_t^{(j)}$	outflow representing the maturity benefit for some policy j
	assumed to be paid at policy year t ,
$Surr_t^{(j)}$	outflow representing surrender for some policy j assumed

to be paid at the the policy year t.

Our aim is to divide the formula 2.1 into parts that depend on the value of investment return and doesn't depend on it. The part that does not depend on the investment income is the same for every scenario, and can be easily summarized for all policies in our portfolio. The part that does depend on investment return we can calculate separetely with aggregated values for the whole portfolio. If we are able to separate the cash flows projection and find such a way, then the calculation won't be policy-by-policy but for aggregated portfolio. Such an approach might speed up the calculation compared to standard policy-by-policy approach.

We want to split formula 2.1 in the following form [6]:

$$CF_t = \sum_{j:\forall policies} fix CF_t^{(j)} + \sum_q Coef_t^q \cdot f_t^{(q)}(i_1^s, i_2^s, \dots, i_t^s),$$
(2.2)

where

$fix CF_t^{(j)}$	part of cash flow for some policy j , that doesn't depend on
	interest rate i_t^s and can be calculated from the one run of
	the full model,
s	number of interest rate scenario $s \in \{1, \ldots, S\}$
$Coef_t^{(q)}$	coefficients relevant at the time t , that is determined for the whole
	portfolio
$f_t^{(q)}$	is a function of investment return on assets at the time t that is
	common for all policies,
q	number of $Coef_t$ and f_t pair, typically more than 1.

2.1.1 Death Benefit as a summation of fund value and sum assured

We will start with the first product, where the death benefit is paid to the policyholder as a summation of sum assured defined by contract and policyholder's accumulated fund value at the year of payment. Let's start with calculation of the cash flow from one contract during one year as defined in Formula 1.4 and in the Table 1.1. For this type of product, the death benefit for some policy j is defined as $(SA^{(j)}+CV_t^{(j)})$, where t is a year of payment. In a case of maturity, the insurance compamy pays the policyholder's capital value at the year of payment $t (CV_t^{(j)})$. In the case of lapse at the year t, the company pays the policyholder the fund value deducted by the surrender fee $(CV_t^{(j)}(1 - fee_t^{(j)}))$. So, we have:

$$\begin{split} CF_{1}^{(j)} &= l_{0}^{(j)}(prem_{1}^{(j)} - comm_{1}^{(j)} - exp_{1}^{(j)}) - CV_{1}^{(j)}(1 - fee_{1}^{(j)})w_{1}^{(j)} - \\ &\quad - (CV_{1}^{(j)} + SA^{(j)})d_{1}^{(j)} - CV_{1}^{(j)}m_{1}^{(j)} = \\ &= l_{0}^{(j)}(prem_{1}^{(j)} - comm_{1}^{(j)} - exp_{1}^{(j)}) - \\ &\quad - CV_{1}^{(j)}[w_{1}^{(j)}(1 - fee_{1}^{(j)}) + d_{1}^{(j)} + m_{1}^{(j)}] - SA^{(j)} \cdot d_{1}^{(j)} = \\ &= l_{0}^{(j)}(prem_{1}^{(j)} - comm_{1}^{(j)} - exp_{1}^{(j)}) - SA^{(j)} \cdot d_{1}^{(j)} - \\ &\quad - (CV_{0}^{(j)} + SP_{1}^{(j)})(1 + i_{1}^{s})[w_{1}(1 - fee_{1}^{(j)}) + d_{1} + m_{1}] - \\ &= l_{0}^{(j)}(prem_{1}^{(j)} - comm_{1}^{(j)} - exp_{1}^{(j)}) - SA^{(j)} \cdot d_{1}^{(j)} - \\ &\quad - CV_{0}^{(j)}[w_{1}^{(j)}(1 - fee_{1}^{(j)}) + d_{1}^{(j)} + m_{1}^{(j)}](1 + i_{1}^{s}) - \\ &\quad - SP_{1}^{(j)}[w_{1}^{(j)}(1 - fee_{1}^{(j)}) + d_{1}^{(j)} + m_{1}^{(j)}](1 + i_{1}^{s}), \end{split}$$

where

 $SP_t^{(j)}$ denotes the saving part of premium of some policy j, that increased the fund value,

- $SP_t^{(j)} = prem_t^{(j)} \alpha^{(j)} \beta^{(j)} \gamma^{(j)} RP_t^{(j)},$ α, β, γ denote the expenses used for calculation of premium [5]:
- $\alpha^{(j)}$ initial expenses on policy j,
- $\beta^{(j)}$ regular administrative expenses on policy j,
- $\gamma^{(j)}$ collecting expenses on policy j;
- denotes the risk part of premium of some policy j, $PR_{t}^{(j)}$
- return on investment under the scenario $s = \{1, \ldots, S\}$ related to i_t^s the policy year t.

Using the same logics, we can continue with the value of cash flow for one policy in the second year:

$$\begin{split} CF_2^{(j)} &= \dots = l_1^{(j)} (prem_2^{(j)} - comm_2^{(j)} - exp_2^{(j)}) - SA^{(j)} \cdot d_2^{(j)} - \\ &\quad - (CV_0^{(j)} + SP_1^{(j)}) [w_2^{(j)}(1 - fee_2^{(j)}) + d_2^{(j)} + m_2^{(j)}](1 + i_1^s)(1 + i_2^s) - \\ &\quad - SP_2^{(j)} [w_2^{(j)}(1 - fee_2^{(j)}) + d_2^{(j)} + m_2^{(j)}](1 + i_2^s) = \\ &= l_1^{(j)} (prem_2^{(j)} - comm_2^{(j)} - exp_2^{(j)}) - SA^{(j)} \cdot d_2^{(j)} - \\ &\quad - CV_0^{(j)} [w_2^{(j)}(1 - fee_2^{(j)}) + d_2^{(j)} + m_2^{(j)}](1 + i_1^s)(1 + i_2^s) - \\ &\quad - SP_1^{(j)} [w_2^{(j)}(1 - fee_2^{(j)}) + d_2^{(j)} + m_2^{(j)}](1 + i_1^s)(1 + i_2^s) - \\ &\quad - SP_2^{(j)} [w_2^{(j)}(1 - fee_2^{(j)}) + d_2^{(j)} + m_2^{(j)}](1 + i_1^s)(1 + i_2^s) - \\ &\quad - SP_2^{(j)} [w_2^{(j)}(1 - fee_2^{(j)}) + d_2^{(j)} + m_2^{(j)}](1 + i_2^s). \end{split}$$

And for the third year we will have:

$$\begin{split} CF_3^{(j)} &= \dots = l_2^{(j)} (prem_3^{(j)} - comm_3^{(j)} - exp_3^{(j)}) - SA^{(j)} \cdot d_3^{(j)} - \\ &\quad - (CV_0^{(j)} + SP_1^{(j)}) [w_3^{(j)}(1 - fee_3^{(j)}) + d_3^{(j)} + m_3^{(j)}](1 + i_1^s)(1 + i_2^s)(1 + i_3^s) - \\ &\quad - SP_2^{(j)} [w_3^{(j)}(1 - fee_3^{(j)}) + d_3^{(j)} + m_3^{(j)}](1 + i_3^s) - \\ &\quad - SP_3^{(j)} [w_3^{(j)}(1 - fee_3^{(j)}) + d_3^{(j)} + m_3^{(j)}](1 + i_3^s) \\ &= l_2^{(j)} (prem_3^{(j)} - comm_3^{(j)} - exp_3^{(j)}) - SA^{(j)} \cdot d_3^{(j)} - \\ &\quad - CV_0^{(j)} [w_3^{(j)}(1 - fee_3^{(j)}) + d_3^{(j)} + m_3^{(j)}](1 + i_1^s)(1 + i_2^s)(1 + i_3^s) - \\ &\quad - SP_1^{(j)} [w_3^{(j)}(1 - fee_3^{(j)}) + d_3^{(j)} + m_3^{(j)}](1 + i_1^s)(1 + i_2^s)(1 + i_3^s) - \\ &\quad - SP_2^{(j)} [w_3^{(j)}(1 - fee_3^{(j)}) + d_3^{(j)} + m_3^{(j)}](1 + i_2^s)(1 + i_3^s) - \\ &\quad - SP_2^{(j)} [w_3^{(j)}(1 - fee_3^{(j)}) + d_3^{(j)} + m_3^{(j)}](1 + i_2^s)(1 + i_3^s) - \\ &\quad - SP_3^{(j)} [w_3^{(j)}(1 - fee_3^{(j)}) + d_3^{(j)} + m_3^{(j)}](1 + i_2^s)(1 + i_3^s) - \\ &\quad - SP_3^{(j)} [w_3^{(j)}(1 - fee_3^{(j)}) + d_3^{(j)} + m_3^{(j)}](1 + i_2^s)(1 + i_3^s) - \\ &\quad - SP_3^{(j)} [w_3^{(j)}(1 - fee_3^{(j)}) + d_3^{(j)} + m_3^{(j)}](1 + i_2^s)(1 + i_3^s) - \\ &\quad - SP_3^{(j)} [w_3^{(j)}(1 - fee_3^{(j)}) + d_3^{(j)} + m_3^{(j)}](1 + i_3^s). \end{split}$$

We can derive the parts of Formula 2.2 from the last equation for projection year t = 3 and some policy j. So, we have:

$$\begin{split} fix CF_3^{(j)} =& l_2^{(j)}(prem_3^{(j)} - comm_3^{(j)} - exp_3^{(j)}) - SA^{(j)} \cdot d_3^{(j)};\\ Coef_3^1 =& (CV_0^{(j)} + SP_1^{(j)})[w_3^{(j)}(1 - fee_3^{(j)}) + d_3^{(j)} + m_3^{(j)}];\\ Coef_3^2 =& SP_2^{(j)}[w_3^{(j)}(1 - fee_3^{(j)}) + d_3^{(j)} + m_3^{(j)}];\\ Coef_3^3 =& SP_3^{(j)}[w_3^{(j)}(1 - fee_3^{(j)}) + d_3^{(j)} + m_3^{(j)}];\\ f_3^1 =& (1 + i_1^s)(1 + i_2^s)(1 + i_3^s);\\ f_3^2 =& (1 + i_2^s)(1 + i_3^s);\\ f_3^3 =& (1 + i_3^s) \\q =& 3 \end{split}$$

The cash flow in year t for some policy j is:

$$\begin{split} CF_t^{(j)} &= \dots = l_{t-1}^{(j)}(prem_t^{(j)} - comm_t^{(j)} - exp_t^{(j)}) - SA^{(j)} \cdot d_t^{(j)} - \\ &\quad - (CV_0^{(j)} + SP_1^{(j)})[w_t^{(j)}(1 - fee_t^{(j)}) + d_t^{(j)} + m_t^{(j)}](1 + i_1^s) \dots (1 + i_t^s) - \\ &\quad - SP_2^{(j)}[w_t^{(j)}(1 - fee_3^{(j)}) + d_t^{(j)} + m_t^{(j)}](1 + i_2^s) \dots (1 + i_t^s) - \\ &\quad - \dots - \\ &\quad - SP_t^{(j)}[w_t^{(j)}(1 - fee_3^{(j)}) + d_t^{(j)} + m_t^{(j)}](1 + i_t^s) \\ &= l_{t-1}^{(j)}(prem_t^{(j)} - comm_t^{(j)} - exp_t^{(j)}) - SA^{(j)} \cdot d_t^{(j)} - \\ &\quad - CV_0^{(j)}[w_t^{(j)}(1 - fee_t^{(j)}) + d_t^{(j)} + m_t^{(j)}](1 + i_1^s) \dots (1 + i_t^s) - \\ &\quad - SP_1^{(j)}[w_t^{(j)}(1 - fee_t^{(j)}) + d_t^{(j)} + m_t^{(j)}](1 + i_1^s) \dots (1 + i_t^s) - \\ &\quad - SP_2^{(j)}[w_t^{(j)}(1 - fee_t^{(j)}) + d_t^{(j)} + m_t^{(j)}](1 + i_2^s) \dots (1 + i_t^s) - \\ &\quad - M \\ &\quad - SP_t^{(j)}[w_t^{(j)}(1 - fee_t^{(j)}) + d_t^{(j)} + m_t^{(j)}](1 + i_2^s) \dots (1 + i_t^s) - \\ &\quad - M \\ &\quad - M \\ &\quad - SP_t^{(j)}[w_t^{(j)}(1 - fee_t^{(j)}) + d_t^{(j)} + m_t^{(j)}](1 + i_2^s) \dots (1 + i_t^s) - \\ &\quad - M \\ &\quad -$$

We can derive parts from Formula 2.2 for some policy j in year t as follows:

$$fix CF_{t}^{(j)} = l_{t-1}^{(j)} (prem_{t}^{(j)} - comm_{t}^{(j)} - exp_{t}^{(j)}) - SA^{(j)} \cdot d_{t}^{(j)},$$

$$Coef_{t}^{1} = (CV_{0}^{(j)} + SP_{1}^{(j)}) [w_{t}^{(j)}(1 - fee_{t}^{(j)}) + d_{t}^{(j)} + m_{t}^{(j)}].$$

$$Coef_{t}^{q \neq 1} = SP_{q}^{(j)} [w_{t}^{(j)}(1 - fee_{t}^{(j)}) + d_{t}^{(j)} + m_{t}^{(j)}],$$

$$f_{t}^{q} = \prod_{k=1}^{q} (1 + i_{k}^{s}), \quad q \in \{1, \dots, t\}, s \in \{1, \dots, S\}$$

$$(2.3)$$

The summarization of all policies in-force J in insurance company's portfolio will be

$$\begin{split} CF_t &= \sum_{j=1}^{J} CF_t^{(j)} = \\ &= \sum_{j=1}^{J} \left(l_t^{(j)}(prem_t^{(j)} - comm_t^{(j)} - exp_t^{(j)}) - SA^{(j)} \cdot d_t^{(j)} \right) - \\ &\quad - \sum_{j=1}^{J} CV_0^{(j)}[w_t^{(j)}(1 - fee_t^{(j)}) + d_t^{(j)} + m_t^{(j)}](1 + i_1^s) \dots (1 + i_t^s) - \\ &\quad - \sum_{j=1}^{J} SP_1^{(j)}[w_t^{(j)}(1 - fee_t^{(j)}) + d_t^{(j)} + m_t^{(j)}](1 + i_2^s) \dots (1 + i_t^s) - \\ &\quad - \sum_{j=1}^{J} SP_2^{(j)}[w_t^{(j)}(1 - fee_t^{(j)}) + d_t^{(j)} + m_t^{(j)}](1 + i_2^s) \dots (1 + i_t^s) - \\ &\quad - \dots - \\ &\quad - \sum_{j=1}^{J} SP_t^{(j)}[w_t^{(j)}(1 - fee_t^{(j)}) + d_t^{(j)} + m_t^{(j)}](1 + i_t^s) = \\ &= \sum_{j=1}^{J} fix CF_t^{(j)} - \\ &\quad - \sum_{j=1}^{J} CV_0^{(j)} \cdot [w_t^{(j)}(1 - fee_t^{(j)}) + d_t^{(j)} + m_t^{(j)}] \cdot \prod_{k=1}^t (1 + i_k^s) - \\ &\quad - \sum_{q=1}^t \sum_{j=1}^J SP_q^{(j)} \cdot [w_t^{(j)}(1 - fee_t^{(j)}) + d_t^{(j)} + m_t^{(j)}] \cdot \prod_{k=q}^t (1 + i_k^s). \end{split}$$

We can note from the Formula 2.4 that the coefficients $Coef_t^q$ defined in 2.3 doesn't depend on investment rate of return and it can be obtained without assumptions about rate of return. Functions f_t^q doesn't depend on the policy, they are the same for all policies and can be extracted from the sum [6].

The calculation process by the analytic function can be summarized as the follows, [7]:

- 1. Run the full model once. This run doesn't depend on investment return;
- 2. Derive the coefficients $fix CF_t$ and $Coef_t^{(q)}$ and save them (usually about 100 ths or more variables based on the product complexity);

3. Take the selected scenario of interest rates $(i_1^s, i_2^s, \ldots, i_t^s)$;

4. Calculate
$$CF_t = \sum_{j=1}^{J} fix CF_t^{(j)} + \sum_k Coef_t^{(q)} \cdot f_t^{(q)}(i_1^s, i_2^s, \dots, i_t^s).$$

The estimation of cash flow by analytic function requires patience and concentration, but the final results of the estimation by analytic function are fast and concluded for our defined product with no significant errors.

2.1.2 Death Benefit as Maximum of Fund Value and Sum Assured

For a product with death benefit as $(SA^{(j)}+CV_t^{(j)})$ and maturity benefit as $CV_t^{(j)}$ the Formula 2.1 can be easily split into two parts: one depends on interest rate and another doesn't. It becomes more complicated when the death benefit is the greatest of these two values $\max(CV_t^{(j)}, SA^{(j)})$. However, we also can try split the cash flow projection into parts as defined in Formula 2.2. Figure 2.1 shows the development of the policyholder *j*'s death benefit. According to the year of payment, the death benefit can be the amount of policyholder's sum assured or fund value.



Figure 2.1: Death Benefit as a maximum value of policyholder's fund value and sum assured

We also can start from derivation of cash flow for one policy. Except the death benefit, all other cash flows are defined in the same way as in the first product. Insurance company pays the capital value $CV_t^{(j)}$ in the year t in case of maturity. In case of lapse, the outflow is $(CV_t^{(j)}(1 - fee_t^{(j)}))$. Cash flow of some policy j in year 1:

$$\begin{split} CF_{1}^{(j)} &= l_{0}^{(j)}(prem_{1}^{(j)} - comm_{1}^{(j)} - exp_{1}^{(j)}) - CV_{1}^{(j)}(1 - fee_{1}^{(j)})w_{1}^{(j)} - \\ &\quad - \max\{CV_{1}^{(j)}; SA^{(j)}\}d_{1}^{(j)} - CV_{1}^{(j)}m_{1}^{(j)} = \\ &= l_{0}^{(j)}(prem_{1}^{(j)} - comm_{1}^{(j)} - exp_{1}^{(j)}) - CV_{1}^{(j)}[w_{1}^{(j)}(1 - fee_{1}^{(j)}) + m_{1}^{(j)}] - \\ &\quad - \max\{CV_{1}^{(j)}; SA^{(j)}\} \cdot d_{1}^{(j)} = \\ &= l_{0}^{(j)}(prem_{1}^{(j)} - comm_{1}^{(j)} - exp_{1}^{(j)}) - \\ &\quad - (CV_{0}^{(j)} + SP_{1}^{(j)})(1 + i_{1}^{s}) \cdot [w_{1}^{(j)}(1 - fee_{1}^{(j)}) + m_{1}^{(j)}] - \\ &\quad - \max\{(CV_{0}^{(j)} + SP_{1}^{(j)})(1 + i_{1}^{s}); SA^{(j)}\} \cdot d_{1}^{(j)} = \\ &= l_{0}^{(j)}(prem_{1}^{(j)} - comm_{1}^{(j)} - exp_{1}^{(j)}) - \\ &\quad - CV_{0}^{(j)}[w_{1}^{(j)}(1 - fee_{1}^{(j)}) + m_{1}^{(j)}](1 + i_{1}^{s}) - \\ &\quad - SP_{1}^{(j)}[w_{1}^{(j)}(1 - fee_{1}^{(j)}) + m_{1}^{(j)}](1 + i_{1}^{s}) - \\ &\quad - \max\{CV_{0}^{(j)}(1 + i_{1}^{s}) + SP_{1}^{(j)}(1 + i_{1}^{s}); SA^{(j)}\} \cdot d_{1}^{(j)} \end{split}$$

For year t the cash flow projection for some policy j and some scenario of investment rate s is:

$$\begin{split} CF_t^{(j)} &= l_t^{(j)}(prem_t^{(j)} - comm_t^{(j)} - exp_t^{(j)}) - \\ &- CV_0^{(j)}[w_t^{(j)}(1 - fee_t^{(j)}) + m_t^{(j)}] \prod_{k=1}^t (1 + i_k^s) - \\ &- SP_1^{(j)}[w_t^{(j)}(1 - fee_t^{(j)}) + m_t^{(j)}] \prod_{k=1}^t (1 + i_k^s) - \\ &- \cdots - \\ &- SP_t^{(j)}[w_t^{(j)}(1 - fee_t^{(j)}) + m_t^{(j)}](1 + i_t^s) - \\ &- \max\{(CV_0^{(j)} + SP_1^{(j)}) \prod_{k=1}^t (1 + i_k^s) + \cdots + SP_t^{(j)}(1 + i_t^s); SA^{(j)}\} \cdot d_t^{(j)} \end{split}$$

In section 1.2 we have discussed an assumption of guaranteed minimum rate of return. It means that the policyholder's fund value will be increased in time by at least guranteed rate of return. This assumption implies the monotone increase of fund value function, that is shown in the Figure 2.1. We can use an intermidiate value theorem, which states that it is a continuous function f, that is defined on an interval [a, b], takes the values f(a) and f(b) at each end of the interval, then it also takes any value between [f(a); f(b)] at some point in the interval [a, b] (see Appendix).

Applying the theorem in our example, we have: the domain interval of the fund function for some policy j is the projection interval $t \in [1, n_j]$, where n_j is a period of policy j. We assume, that the policyholder's fund value in time 1 is less than the value of sum assured $(CV_1^{(j)} < SA^{(j)})$ and in time n_j is greater than the sum assured $(CV_{n_j}^{(j)} > SA^{(j)})$. Then from the intermediate value theorem there exists time $t^{*j} \in [1, n_j]$ where the policyholder's fund value equals to the value of sum assured. So, we can bisect an interval $[1, n_j]$ on two subintervals $[1, t^{*j}]$

and $(t^{*j}; n_j]$. Then in each subinterval we can select the relevant value of death benefit. The cash flow projections in year t of some policy j is:

$$\begin{split} t &\leq t^{*j} : CF_t^{(j)} = l_t^{(j)}(prem_t^{(j)} - comm_t^{(j)} - exp_t^{(j)}) - SA^{(j)}d_t^{(j)} \\ &\quad - CV_0^{(j)}[w_t^{(j)}(1 - fee_t^{(j)}) + m_t^{(j)}] \prod_{k=1}^t (1 + i_k^s) - \\ &\quad - SP_1^{(j)}[w_t^{(j)}(1 - fee_t^{(j)}) + m_t^{(j)}] \prod_{k=1}^t (1 + i_k^s) - \\ &\quad - \cdots - \\ &\quad - SP_t^{(j)}[w_t^{(j)}(1 - fee_t^{(j)}) + m_t^{(j)}](1 + i_t^s) \end{split}$$
(2.5)
$$t > t^{*j} : CF_t^{(j)} = l_t^{(j)}(prem_t^{(j)} - comm_t^{(j)} - exp_t^{(j)}) - \\ &\quad - CV_0^{(j)}[w_t^{(j)}(1 - fee_t^{(j)}) + m_t^{(j)} + d_t^{(j)}] \prod_{k=1}^t (1 + i_k^s) - \\ &\quad - SP_1^{(j)}[w_t^{(j)}(1 - fee_t^{(j)}) + m_t^{(j)} + d_t^{(j)}] \prod_{k=1}^t (1 + i_k^s) - \\ &\quad - \cdots - \\ &\quad - SP_t^{(j)}[w_t^{(j)}(1 - fee_t^{(j)}) + m_t^{(j)} + d_t^{(j)}] \prod_{k=1}^t (1 + i_k^s) - \\ &\quad - \cdots - \\ &\quad - SP_t^{(j)}[w_t^{(j)}(1 - fee_t^{(j)}) + m_t^{(j)} + d_t^{(j)}] \prod_{k=1}^t (1 + i_k^s) - \\ &\quad - \cdots - \\ &\quad - SP_t^{(j)}[w_t^{(j)}(1 - fee_t^{(j)}) + m_t^{(j)} + d_t^{(j)}](1 + i_t^s) \end{split}$$

And also the division on the parts as in Formula 2.2 will be depend on year t of the projection:

$$\begin{split} t &\leq t^{*j} : fix \mathrm{CF}_{t}^{(j)} = l_{t-1}^{(j)}(prem_{t}^{(j)} - comm_{t}^{(j)} - exp_{t}^{(j)}) - SA^{(j)} \cdot d_{t}^{(j)}, \\ &\quad Coef_{t}^{1} = (CV_{0}^{(j)} + SP_{1}^{(j)})[w_{t}^{(j)}(1 - fee_{t}^{(j)}) + m_{t}^{(j)}], \\ &\quad Coef_{t}^{q \neq 1} = SP_{q}^{(j)}[w_{t}^{(j)}(1 - fee_{t}^{(j)}) + m_{t}^{(j)}], \\ &\quad f_{t}^{q} = \prod_{k=1}^{q}(1 + i_{k}^{s}), \quad q \in \{1, \dots, t\}, s \in \{1, \dots, S\} \\ t > t^{*j} : fix \mathrm{CF}_{t}^{(j)} = l_{t-1}^{(j)}(prem_{t}^{(j)} - comm_{t}^{(j)} - exp_{t}^{(j)}), \\ &\quad Coef_{t}^{1} = (CV_{0}^{(j)} + SP_{1}^{(j)})[w_{t}^{(j)}(1 - fee_{t}^{(j)}) + d_{t}^{(j)} + m_{t}^{(j)}], \\ &\quad Coef_{t}^{q \neq 1} = SP_{q}^{(j)}[w_{t}^{(j)}(1 - fee_{t}^{(j)}) + d_{t}^{(j)} + m_{t}^{(j)}], \\ &\quad f_{t}^{q} = \prod_{k=1}^{q}(1 + i_{k}^{s}), \quad q \in \{1, \dots, t\}, s \in \{1, \dots, S\} \end{split}$$

We can note from the Formula 2.5 that the functions of investment f_t^q does not depend on the values of t^{*j} and also can be extracted from the sum. The total amount of cash flows for all policies is then:

$$CF_t = \sum_{j=1}^J CF_t^{(j)}$$

To derive the formula of cash flows for all policies in year t we will define the time t^{*j} for each of the policy used in calculation. Before that time, the death

benefit for policy j is the value of sum assured, and after that time, the benefit is calculated as fund value at the year t. According to the year of death benefit change $(1 < t^{*(1)} < \cdots < t^{*(J)} < \infty)$, we will order the policies in our portfolio $(\{j^{*(1)}, \ldots, j^{*(J)}\})$. We assume that before the first time of change $[1, t^{*(1)}]$ all the policies have a death benefit as the value of sum assured. Then we assume that in interval $(t^{*(1)}; t^{*(2)}]$ there is some policy $j^{*(1)}$ that "changes" the benefit, in interval $(t^{*(2)}; t^{*(3)}]$ there are already two policies $\{j^{*(1)}, j^{*(2)}\}$ with death benefit as a fund value, etc. Finally, after the last time of change $(t^{*(J)}, \infty)$ all policies in the portfolio have the death benefit as a value of policyholder's fund.

According to our assumptions we have:

$$t \in [1, t^{*(1)}] : CF_t = \sum_{j=1}^{J} CF_t^{(j)} =$$

$$= \sum_{j=1}^{J} \left(l_t^{(j)}(prem_t^{(j)} - comm_t^{(j)} - exp_t^{(j)}) - SA^{(j)}d_t^{(j)} \right)$$

$$- \sum_{j=1}^{J} CV_0^{(j)}[w_t^{(j)}(1 - fee_t^{(j)}) + m_t^{(j)}] \prod_{k=1}^{t} (1 + i_k^s) -$$

$$- \sum_{j=1}^{J} SP_1^{(j)}[w_t^{(j)}(1 - fee_t^{(j)}) + m_t^{(j)}] \prod_{k=1}^{t} (1 + i_k^s) -$$

$$- \cdots -$$

$$- \sum_{j=1}^{J} SP_t^{(j)}[w_t^{(j)}(1 - fee_t^{(j)}) + m_t^{(j)}](1 + i_t^s),$$

where we can easily defined the parts of Formula 2.2 as follows:

$$fix CF_{t}^{(j)} = l_{t-1}^{(j)} (prem_{t}^{(j)} - comm_{t}^{(j)} - exp_{t}^{(j)}) - SA^{(j)} \cdot d_{t}^{(j)},$$

$$Coef_{t}^{1} = (CV_{0}^{(j)} + SP_{1}^{(j)})[w_{t}^{(j)}(1 - fee_{t}^{(j)}) + m_{t}^{(j)}].$$

$$Coef_{t}^{q\neq 1} = SP_{q}^{(j)}[w_{t}^{(j)}(1 - fee_{t}^{(j)}) + d_{t}^{(j)} + m_{t}^{(j)}],$$

$$f_{t}^{q} = \prod_{k=1}^{q} (1 + i_{k}^{s}), \quad q \in \{1, \dots, t\}, s \in \{1, \dots, S\}$$

$$(2.7)$$

After the first time $t^{*(1)}$ and before the the second time $t^{*(2)}$ the sum of cash flows is:

$$\begin{split} t \in [t^{*(1)}; t^{*(2)}] : & CF_t = \sum_{j=1}^J CF_t^{(j)} = \\ & = \sum_{j=1}^J l_t^{(j)} (prem_t^{(j)} - comm_t^{(j)} - exp_t^{(j)}) - \sum_{j=2}^J SA^{(j)} d_t^{(j)} \\ & - \left(\sum_{j=1}^J CV_0^{(j)} [w_t^{(j)}(1 - fee_t^{(j)}) + m_t^{(j)}] + CV_0^{(1)} d_t^{(1)}\right) \prod_{k=1}^t (1 + i_k^s) - \\ & - \left(\sum_{j=1}^J SP_1^{(j)} [w_t^{(j)}(1 - fee_t^{(j)}) + m_t^{(j)}] + SP_1^{(1)} d_t^{(1)}\right) \prod_{k=1}^t (1 + i_k^s) - \\ & - \cdots - \\ & - \left(\sum_{j=1}^J SP_t^{(j)} [w_t^{(j)}(1 - fee_t^{(j)}) + m_t^{(j)}] + SP_t^{(1)} d_t^{(1)}\right) (1 + i_t^s), \end{split}$$

with partitiation

$$\begin{split} \sum_{j=1}^{J} fix \operatorname{CF}_{t}^{(j)} &= \sum_{j=1}^{J} l_{t-1}^{(j)} (prem_{t}^{(j)} - comm_{t}^{(j)} - exp_{t}^{(j)}) - \sum_{j=2}^{J} SA^{(j)} \cdot d_{t}^{(j)}, \\ \sum_{j=1}^{J} Coef_{t}^{1} &= \sum_{j=1}^{J} (CV_{0}^{(j)} + SP_{1}^{(j)}) [w_{t}^{(j)}(1 - fee_{t}^{(j)}) + m_{t}^{(j)}] + (CV_{0}^{(1)} + SP_{1}^{(1)}) d_{t}^{(1)}, \\ \sum_{j=1}^{J} Coef_{t}^{q \neq 1} &= \sum_{j=1}^{J} SP_{q}^{(j)} [w_{t}^{(j)}(1 - fee_{t}^{(j)}) + d_{t}^{(j)} + m_{t}^{(j)}] + SP_{q}^{(1)} d_{t}^{(1)}, \\ f_{t}^{q} &= \prod_{k=1}^{q} (1 + i_{k}^{s}), \quad q \in \{1, \dots, t\}, s \in \{1, \dots, S\} \end{split}$$

$$(2.8)$$

After the time $t^{*(2)}$ there are already two policies, that have a death benefit as a capital value. The sum of total cash flows in this time interval is:

$$\begin{split} t \in [t^{*(2)}; t^{*(3)}] &: CF_t = \sum_{j=1}^J CF_t^{(j)} = \\ &= \sum_{j=1}^J l_t^{(j)} (prem_t^{(j)} - comm_t^{(j)} - exp_t^{(j)}) - \sum_{j=3}^J SA^{(j)} d_t^{(j)} \\ &- \left(\sum_{j=1}^J CV_0^{(j)} [w_t^{(j)} (1 - fee_t^{(j)}) + m_t^{(j)}] + \sum_{j=1}^2 CV_0^{(j)} d_t^{(j)}\right) \prod_{k=1}^t (1 + i_k^s) - \\ &- \left(\sum_{j=1}^J SP_1^{(j)} [w_t^{(j)} (1 - fee_t^{(j)}) + m_t^{(j)}] + \sum_{j=1}^2 SP_1^{(j)} d_t^{(j)}\right) \prod_{k=1}^t (1 + i_k^s) - \\ &- \dots - \\ &- \left(\sum_{j=1}^J SP_t^{(j)} [w_t^{(j)} (1 - fee_t^{(j)}) + m_t^{(j)}] + \sum_{j=1}^2 SP_t^{(j)} d_t^{(j)}\right) (1 + i_t^s), \end{split}$$

with separation on the parts as:

$$\begin{split} \sum_{j=1}^{J} fix \operatorname{CF}_{t}^{(j)} &= \sum_{j=1}^{J} l_{t-1}^{(j)} (prem_{t}^{(j)} - comm_{t}^{(j)} - exp_{t}^{(j)}) - \sum_{j=3}^{J} SA^{(j)} \cdot d_{t}^{(j)}, \\ \sum_{j=1}^{J} Coef_{t}^{1} &= \sum_{j=1}^{J} (CV_{0}^{(j)} + SP_{1}^{(j)}) [w_{t}^{(j)} (1 - fee_{t}^{(j)}) + m_{t}^{(j)}] + \\ &+ \sum_{j=1}^{2} (CV_{0}^{(j)} + SP_{1}^{(j)}) d_{t}^{(j)}, \end{split}$$
(2.9)
$$\begin{split} \sum_{j=1}^{J} Coef_{t}^{q \neq 1} &= \sum_{j=1}^{J} SP_{q}^{(j)} [w_{t}^{(j)} (1 - fee_{t}^{(j)}) + m_{t}^{(j)}] + \sum_{j=1}^{2} SP_{q}^{(j)} d_{t}^{(j)}, \\ f_{t}^{q} &= \prod_{k=1}^{q} (1 + i_{k}^{s}), \quad q \in \{1, \dots, t\}, s \in \{1, \dots, S\} \end{split}$$

And finally, we have that all policies in the portolio have after time of last policy change the death benefit as value of policyholder's fund.

$$\begin{aligned} t \in [t^{*(J)}, \infty] : CF_t &= \sum_{j=1}^J CF_t^{(j)} = \\ &= \sum_{j=1}^J l_t^{(j)}(prem_t^{(j)} - comm_t^{(j)} - exp_t^{(j)}) \\ &- \sum_{j=1}^J CV_0^{(j)}[w_t^{(j)}(1 - fee_t^{(j)}) + m_t^{(j)} + d_t^{(j)}] \prod_{k=1}^t (1 + i_k^s) - \\ &- \sum_{j=1}^J SP_1^{(j)}[w_t^{(j)}(1 - fee_t^{(j)}) + m_t^{(j)} + d_t^{(j)}] \prod_{k=1}^t (1 + i_k^s) - \\ &- \cdots - \\ &- \sum_{j=1}^J SP_t^{(j)}[w_t^{(j)}(1 - fee_t^{(j)}) + m_t^{(j)} + d_t^{(j)}](1 + i_t^s) \end{aligned}$$

Parts from the Formula 2.2 are defined as follows:

$$\begin{aligned} fix \mathrm{CF}_{t}^{(j)} &= l_{t-1}^{(j)}(prem_{t}^{(j)} - comm_{t}^{(j)} - exp_{t}^{(j)}) \cdot d_{t}^{(j)}, \\ Coef_{t}^{1} &= (CV_{0}^{(j)} + SP_{1}^{(j)})[w_{t}^{(j)}(1 - fee_{t}^{(j)}) + m_{t}^{(j)} + d_{t}^{(j)}]. \\ Coef_{t}^{q \neq 1} &= SP_{q}^{(j)}[w_{t}^{(j)}(1 - fee_{t}^{(j)}) + d_{t}^{(j)} + m_{t}^{(j)} + d_{t}^{(j)}], \\ f_{t}^{q} &= \prod_{k=1}^{q} (1 + i_{k}^{s}), \quad q \in \{1, \dots, t\}, s \in \{1, \dots, S\} \end{aligned}$$

$$(2.10)$$

Of course, there might be already some policies, that at beginning of the projection have the fund value greater than sum assured. The total amount of all policies in-force J is deducted then by this amount of policies. For these policies, the outflows of death benefit is calculated as a policyholder's fund value.

Here we have shown the capabilities of the analytic function using two main insurance products. It also can be easily adjusted for other insurance products, where the death benefit is paid as fund value or sum assured only. Or there might be products with different maturity benefits.

2.2 Cluster analysis

2.2.1 Data Preparation

In this section we will focus on a method of cluster analysis.

The purpose of clustering is to allocate observations of variables into homogenous and distinct groups ("clusters"). That means that observations are similar to each other within the group and different from observation in other groups [8].

Generally, in data clustering we work with the data matrix as shown below [9], where the row stands for an observation and each columns represents a variable:

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_{1}^{T} \\ \mathbf{x}_{2}^{T} \\ \vdots \\ \mathbf{x}_{J}^{T} \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,L} \\ x_{2,1} & x_{2,2} & \dots & x_{2,L} \\ \vdots & \vdots & \ddots & \vdots \\ x_{J,1} & x_{J,2} & \dots & x_{J,L} \end{pmatrix}$$

To make our variables more comparable, it is possible to scale them, so they will have zero mean and the variance equals to one [10]:

$$z_{j,l} = \frac{x_{j,l} - \bar{x}_l}{\sigma_l}, \quad j = 1, \dots, J; l = 1, \dots, L,$$

where \bar{x}_l is the sample mean, and σ_l is the sample standard deviation of *l*th attribute (variable).

2.2.2 Distance Measures

Generally, it is not so easy to define 'cluster' in formal way [11], it has been used in an essentially intuitive character. There are combination of acceptable criterias and requirements that help to understand the common sense of clusters, for example:

- 1. Share the same or closely related properties;
- 2. Show small mutual distances;
- 3. Have "contacts" or "relations" with at least one other object in the group; or
- 4. To be clearly distinguishable from the complements, i.e. rest of the objects in the data set.

We don't have any assumptions about the distibution of the underlying data. Using the cluster analysis, we are able to form groups of related observations. Clustering techniques can be used for any data set. All that is needed is a measure of how far one element in the set is from another element, using the function that gives the distance between the elements. The larger the distance value is, the more dissimilar the pair of observations are, and vice versa. The distance function d satisfies the follows [12]:

- $d(x_i, x_i) = 0$,
- $d(x_i, x_j) \ge 0$,
- $d(x_i, x_j) = d(x_j, x_i).$

where x_i is *i*th element of the dataset.

There are many methods to calculate the distance between the elements of (x_i, x_j) . In cluster analysis the choice of distance function is a important step, which will impact the clustering results [10]. The most common distance function in practical are:

1. Eucledean distance:

$$d_{euc}(x_i, x_j) = \sqrt{\sum_{l=1}^{L} (x_{i,l} - x_{j,l})^2};$$

2. Manhattan distance:

$$d_{manh}(x_i, x_j) = \sum_{l=1}^{L} |x_{i,l} - x_{j,l}|.$$

The generalized form of Euclidean and Manhattan distance is Minkowski distance, which is defined as follows [9]:

$$d_{mink}(x_i, x_j) = \left(\sum_{l=1}^{L} |x_{i,l} - x_{j,l}|^p\right)^{1/p}$$

In case of p = 2, the Minkowski distance is equal to Euclidean, and in case of p = 1 it becomes to Manhattan distance.

In our example we will create clusters of homogenous policies according to the individual value of future cash flows:

$$PVCF_0^{perpolicy} \approx PVCF_0^{MP}.$$
 (2.11)

So the present values within each cluster are as close as possible.

The vector of present value of cash flow according to which we will create our groups of policies is one-dimensional. In our example, the distance between the policies $i, j \in \{1, ..., J\}$ will be the absolute value of difference between their present values of projected cash flows:

$$d(PVCF_{0}^{(i)}, PVCF_{0}^{(j)}) = \left| PVCF_{0}^{(i)} - PVCF_{0}^{(j)} \right|$$

2.2.3 Clustering methods

Clustering algorithms can be broadly divided into two categories:

• Partitional clustering

A partitioning method constructs K groups, that satisfy the requirements [13]:

- Each group contains minimum one object;
- Each object belongs to maximum one group.

Conditions mentioned above imply: $K \leq J$, where K is a number of groups, and J is a number of observations. It means that there are at most as many groups as there are observations. It is important to note that the number of cluster K is given by users. Requirements of K are discussed below.

Generally, the algorithm tries to find a "good" partition, that means that the objects of the same cluster should be close to each other, whereas objects of different clusters should be as far as possible [13].

• Hierarchial clustering

Hierarchial clustering is an alternative approach of clustering. Compared to the partitioning clustering, it does not require to specify the number of clusters [10]. Hierarchial clustering groups the observations into a sequence of nested clusters, the result is a tree-based representation of the objects.

Within this thesis we will focus on the partitioning clustering method, that will help to group our insurance policies into K disjoint subsets. All the policies that belong to the same cluster can be characterized with a group representer using the scale as a number of policies in the cluster. So, we are able to reduce the number of policies J to the given number of clusters K and by that to reduce the total time of calculation.

To group the policies into clusters we will use two commonly used algorithms of partitioning clustering method [10]:

• K-means clustering

It is the most commonly used algorithm for partitioning a given data set into a set of K clusters. In this method, each cluster is represented by the means of the policies that belong to the cluster. K-means clustering algorithm is more sensitive to outliers and anamalous observations.

• K-medoids clustering or PAM (Partitioning Around Medoids).

In K-medoids algorithm, each cluster is represented by one of the objects in the cluster, which are called medoids. The algorithm is less sensitive to outliers compared to k-means.

K-means algorithm

The basic idea of k-means clustering is to define the clusters by minimizing the total within-cluster variation. This is defined as the sum of squared Euclidean distances between the observations and the corresponding centroid [10]:

$$W(S_k) = \sum_{x_i \in S_k} (x_i - C_k)^2,$$

where x_i is an observation that belongs to the cluster S_k , and C_k is the geometric centroid of the data points in S_k .

An observation x_i is assigned to the cluster S_k , if the sum of squares distance of the observation x_i to the cluster center C_k is minimum.

The total within-cluster variation is defined as follow:

$$\sum_{k=1}^{K} W(C_k) = \sum_{k=1}^{K} \sum_{x_i \in S_k} (x_i - C_k)^2.$$
(2.12)

The K-means clustering algorithm can be described in the following steps ([14], [15]):

- 1. Specify the number of clusters and the elements of each cluster. It can be chosen arbitarily or deliberatly.
- 2. Calculate each cluster's centroid, and the distances between each observation and each centroid. If the observation is nearer to the centroid of a cluster than the one to which it currently belongs, re-assign it to the nearest cluster;
- 3. Repeat step number 2 until all observations are the nearest to the centroid of the cluster to which it belongs;
- 4. If the number of clusters cannot be specified with confidence in advance, repeat steps 1 to 3 with a different number of clusters and evaluate the results.

The big disadvantage of such method is that it depends on the order choice, which is used for grouping and this can cause different cluster results each time.

K-medoids algorithm

The idea of K-medoids partioning algorithm is to select K representative objects in the data set. The corresponding K clusters are found by assigning each remaining object to the nearest representative object, that is called the medoid of the cluster. To be exact, the average distance of the medoid to all other observations in the same cluster is being minimized [13]. Medoid in the cluster is then the most centrally located point.

The K-means algorithm minimizes the the average squared distance, so-called centroid. Compared to K-means algorithm, the K-medoids algorithm is less sensitive to noise and outliers, because it uses medoids (representative object) as cluster center.

K-medoids clustering algorithm can be briefly described in the following steps [10]:

- 1. Select K objects to become medoids;
- 2. Calculate the distance matrix between the objects;
- 3. Assign every object to its closest medoid;
- 4. For each cluster, search if any of the objects of the cluster decreases the average dissimilarity coefficient; if it does, select the entity that decreases this coefficient the most as the medoid for this cluster;
- 5. If at least one medoid has changed go to step 3, else end the algorithm.

Like in K-means clustering, the k-medoids algorithm also requires to specify the number of clusters to be generated.

Determining the optimal number of clusters in the given dataset is one of of the fundamental issue in partitioning clustering. Unfortunetely, there is no definite answer for that issue. The optimal number of clusers depends on the used method and parameters, it is somehow subjective. There exists some direct and statistical methods that help to determine the number of clusters, such as silhouette analysis or gap statistic. One of the possible option can be a quick rule of thumb $K = \left\lfloor \sqrt{\frac{N}{2}} \right\rfloor$. Our strategy of determining the optimal number of clusters needed to process and the calculated errors. The error term will be defined as a difference between the value of cash flows calculated by policy-by-policy approach and by cluster analysis using exactly K clusters.

3. Interest rate scenarios

3.1 Models of interest rates

To simulate the scenarios of investment return, we will use the Hull-White model, which is extension of the Vasecek model. The simulation of interest rate scenarios is not the main topic of this thesis, however it's very important step in further calculations. Instead of using the Hull-White model, there can be used any other model of interest rate.

In this chapter we will present the short description of the Hull-Wite model and introduce the basic formulas needed for simulation of interest rates. We can refer to the work [16], which introduces the practical aspects of interest rate models and also describes all steps needed for parameter estimates.

The Hull-White model is a short rate model. In general, it has the following dynamics ([16], [17]):

$$dr(\theta) = [\theta(t) - \alpha r(t)]dt + \sigma dW(t),$$

where α and σ are positive constants, W_t is a Wiener process which is defined as [17]:

- 1. $W_0 = 0;$
- 2. W has continuous paths a.s.;
- 3. For any $0 = t_0 < t_1 < \cdots < t_m$ the increments $W(t_1) W(t_0), \ldots, W(t_m) W(t_{m-1})$ are independent;

4.
$$W(t+u) - W(t) \sim N(0,u)$$
.

and

$$\theta(t) = \frac{\partial f^M(0,t)}{\partial t} + \alpha f^M(0,t) + \frac{\sigma^2}{2\alpha} (1 - e^{-2\alpha t})$$

Function $f^{M}(0,t)$ stands for market instantenous forward rate.

The equition 3.1 can be integrated so as to yield [17]:

$$\begin{aligned} r(t) = r(s)e^{\alpha(t-s)} + \int_s^t e^{-\alpha(t-u)}\theta(u)du + \sigma \int_s^t e^{-\alpha(t-u)}\mathrm{d}W(u) = \\ = r(s)e^{-\alpha(t-s)} + \varphi(t) - \varphi(s)e^{-\alpha(t-s)} + \sigma \int_s^t e^{-\alpha(t-u)}\mathrm{d}W(u), \end{aligned}$$

where

$$\varphi(t) = f^M(0,t) + \frac{\sigma^2}{2\alpha^2}(1 - e^{-\alpha t})^2.$$

r(t) conditional on \mathcal{F}_0 (the beginning of the simulation) is normally distributed with mean and variance given respectively by ([16],[17])

$$\mathbb{E}[r(t)] = f^{M}(0,t) + \frac{\sigma^{2}}{2\alpha^{2}}(1-e^{-\alpha t})^{2}$$

$$Var[r(t)] = \frac{\sigma^{2}}{2\alpha}[1-e^{-2\alpha t}]$$
(3.1)

There is limited number of bonds with limited range of maturities. However, the models of instantaneous rates require continuous function of time to maturity. Therefore, we will need a suitable model to extrapolate and interpolate the yield to maturity. One of the most popular models is Nelson-Siegel, or the extended version Nelson-Siegel-Svensson. The Nelson-Siegel-Svensson model assumes that the yield curve can be described with the following function:

$$R^{M}(0,T) = \beta_{0} + \beta_{1} \frac{1 - \exp\{-\frac{T}{\gamma_{1}}\}}{\frac{T}{\gamma_{1}}} + \beta_{2} \left(\frac{1 - \exp\{-\frac{T}{\gamma_{1}}\}}{\frac{T}{\gamma_{1}}} - \exp\{-\frac{T}{\gamma_{1}}\}\right) + \beta_{3} \left(\frac{1 - \exp\{-\frac{T}{\gamma_{2}}\}}{\frac{T}{\gamma_{2}}} - \exp\{-\frac{T}{\gamma_{2}}\}\right),$$
(3.2)

where $\beta_0, \beta_1, \beta_2, \beta_3, \gamma_1$ and γ_2 are constant parameters.

The instantaneous forward interest rate is:

$$f^{M}(0,T) = \beta_{0} + \beta_{1}e^{-\frac{T}{\gamma_{1}}} + \beta_{2}T\frac{e^{-\frac{T}{\gamma_{1}}}}{\gamma_{1}} + \beta_{3}T\frac{e^{-\frac{T}{\gamma_{2}}}}{\gamma_{2}}$$

The derivation of the function of instantenous forward rate with respect to T is equal to:

$$\frac{\partial f^M(0,T)}{\partial T} = -\frac{\beta_1}{\gamma_1} e^{-\frac{T}{\gamma_1}} + \beta_2 \left(\frac{e^{-\frac{T}{\gamma_1}}}{\gamma_1} - T \frac{e^{-\frac{T}{\gamma_1}}}{\gamma_1^2} \right) + \beta_3 \left(\frac{e^{-\frac{T}{\gamma_2}}}{\gamma_2} - T \frac{e^{-\frac{T}{\gamma_2}}}{\gamma_2^2} \right).$$

Using the Euler approximation, we can write the discretized equation of 3.1 in the following way [16]:

$$r(t + \Delta t) = r(t) + [\theta(t) - \alpha r(t)]\Delta t + \sigma \Delta W(t) =$$

= $(1 - \alpha \Delta t)r(t) + \Delta t\theta(t) + \sqrt{\Delta t}\sigma N(0, 1);$ (3.3)
 $r(0) = f^M(0, 0),$

where N(0,1) means the random value of standard normal distribution.

The parameters α , σ in Hull-White model can be estimated using the suitable interest rate derivative. The estimation process is called calibration. This is a multi-dimensional optimization task that is trying to find a combination of the parameters, such that the modelled prices fits the best to the market prices of selected derivatives. Calibration using the Swoptions is introduced in work [16].

4. Implementation

In this section we will apply in practice the theoretical methods discussed in previous chapters.

As the first step, we will simulate scenarios of interest rates that will be used for investment return and for discounting. Usage of interest rate scenarios as a return on investment and a discount is quite common practice on insurance market. In our example, we will use 50 scenarios of interest rates. In practice, insurance company might use much more interest rate scenarios, but for the purpose of showing the effectivness of the used method it will be enough.

Then, we will focus on calculation of cash-flows using three possible method for two types of products used in our thesis. These three methods are:

- Policy-by-policy (standard) method;
- Analytic function;
- Clustering method.

And finally, we will compare the results and the final time needed for calculations.

The total calculations are processed in Wolfram Mathematica software. It is used in many scientific, mathematical and computing fields.

4.1 Interest rate models

We will simulate the possible interest rate scenarios based on risk-free rate that is recommended by EIOPA for caclulation of provision under Solvency II legislative. The risk-free curve is spot and is taken for Czech republic as at December 2017 with no volatility adjustment. The number of observation is 150, the data is taken from [18].



Figure 4.1: EIOPA Risk-Free Rate as at 31.12.2017

The estimated parameters of Nelson-Siegel-Svensson model for EIOPA yields, as at the Formula 3.2 are:

β_0	0.0420
β_1	-0.0345
β_2	-0.0180
β_3	-0.0617
γ_1	1.2360
γ_2	6.7211

Figure 4.2 shows the Nelson-Siegel-Svensson function used on EIOPA spot rates.



Figure 4.2: Nelson-Siegel-Svensson model based on EIOPA spot rate

We obtained the estimates of α and σ from the Company Tools4F, Michal Hakala. Parameters are estimated from the actual market data. The Company has the all required data and software needed for parameter estimation.

 $\begin{array}{ccc} \alpha & 0.1413 \\ \sigma & 0.0167 \end{array}$

We will simulate the interest rate scenarios as in the Formula 3.3 with the initial condition r(0,0) = 0.0075. Figure 4.3 shows the simulation of 50 scenarios of future interest rates using the Hull-White model with estimated parameters. Expected value shown in the Figure 4.3 is calculated using the Formula 3.1



Figure 4.3: Simulation of interest rate scenarios

The relationship between spot and forward rates is the following [4]:

$$_{1}f_{t} = \frac{(1+s_{t})^{t}}{(1+s_{t-1})^{t-1}} - 1,$$

where

 $_{1}f_{t}$ forward rate for the *t*th year,

 s_t zero rate for t years.

4.2 Assumptions

4.2.1 Policy data

Unfortunetly, we didn't have any suitable dataset to use in this thesis for comparison of calculation time. So, we've decided to generate the dataset that would be suitable for this work. We will generate 2000 policies in-force in a life insurance company. We assume that we have only two type of contracts according to the frequency of premium payment. The first type of contract is with regular (annual) premium payments and the second type are contracts with single premium payment. We define a variable *Sex* as equals to zero for male policyholders and equals to one for female policyholders. Further, we assume that in our dataset we have 20% of all policyholders at age of 15 - 25 years old, 20% at age 25 - 35 years old, 30% at age 35 - 45 years old, 15% at age 45 - 55 years old, and final 15% at age 55 - 65 years old. We assume that the insurance company offers products with 8 possible amounts of sum assured (80 TCZK, 100 TCZK, 120 TCZK, 150 TCZK, 180 TCZK, 200 TCZK, 220 TCZK, 250 TCZK). The summary of our generated variables is shown in the Table 4.1.

Variable	Categories	Number of observations
Policy Type	1	994
(1 - regular, 2 - single)	2	1006
Sex	0	1013
(0 - male, 1 - female)	1	987
Age	(15; 25)	400
(in years)	(25; 35)	400
	(35; 45)	600
	(45; 55)	300
	(55; 65)	300
Sum Assured	80 000	260
(in CZK)	100 000	236
	120000	264
	150 000	275
	180000	241
	200 000	238
	220 000	256
	250 000	230
Policy period	5	305
(in years)	10	362
	15	375
	20	288
	25	250
	30	162
	35	113
	40	90
	45	55

Table 4.1: The summary of simulated dataset variables

Figure 4.4 shows the histogram of entry age of policyholders in our generted portfolio.



Figure 4.4: Histogram of entry age

4.2.2 Other Assumptions

For assumptions of mortality we will use the mortality tables from the Czech Statistical Office from year 2016. The mortality data can be found at [19]. We will use the mortality experience coefficients depending on the policy year that is shown in the Table 4.2. The expected mortality in policy year t is $q_{x,\tau}^{exp} = coef_{\tau} \cdot q_x$. We assume that the new insurers have lower mortality rates than the individuals who are already insured.

Assumptions of lapses are shown in the Table 4.3. We assume that the probabilities of lapses are higher during the first five policy years.

Policy	Policy	type	Policy	Policy	type
year	Regular	Single	year	Regular	Single
1	0.30	0.30	1	0.20	0.15
2	0.40	0.40	2	0.15	0.10
3	0.50	0.50	3	0.18	0.13
4	0.60	0.60	4	0.15	0.10
≥ 5	0.70	0.70	5	0.12	0.07
			≥ 6	0.08	0.03

Table 4.2: Mortality experience Table 4.3: Lapses assumptions

We will assume that the technical interest rate is equal to 2,1%. All cash flows are valuated to the date 1.1.2017.

All other assumptions that were used in calculation of our example can be found in the Table 4.4.

Assumptions	Policy	Policy type		
	Regular	Single		
$\alpha\%$ from SA	3.00%	3.00%		
$\alpha\%$ from Premium	25.0%	3.00%		
$\beta\%$ from SA	0.30%	0.30%		
$\gamma\%$ from Premium	2.00%	0.00%		
Initial Commission % Premium	35.0%	35.0%		
Renewal Commission % Premium	4.00%	0.00%		
Initial Expenses Fix	2000	0.000		
Initial Expenses % Premium	4.00%	1.50%		
Renewal Expenses Fix	600.0	0.000		
Renewal Expenses % Premium	8.00%	0.50%		
Surrender period	2years	2 years		
Surrender fee	5.00%	5.00%		
Inflation rate of fix expenses	2.00%	2.00%		

Table 4.4:	Other	assumptions	used in	our	example

For estimation of single and regular premium of each policy, we will use the formulas for gross premium [5]. Compared to net premium, the gross premium is already increased by insurance company's expenses. Formula 4.1 shows the calculation of single gross premium.

$$JB_x = A_{x\overline{n}} + \alpha + \beta \cdot \ddot{a}_{x\overline{n}}, \qquad (4.1)$$

where

 $\begin{array}{ll} x & \text{age of a policy holder,} \\ n & \text{policy period,} \\ JB_x & \text{single brutto unit premium,} \\ A_{x\overline{n}|} & \text{endowment value} \\ & A_{x\overline{n}|} = \frac{M_x - M_{x+n} + D_{x+n}}{D_x}, \\ \alpha & \text{inital expenses (in \%),} \\ \beta & \text{regular administrative expenses (in \%),} \\ \ddot{a}_{x\overline{n}|} & \text{temporary annuity} \\ & \ddot{a}_{x\overline{n}|} = \frac{N_x - N_{x+n}}{D_x}. \end{array}$

Commutation function used in notation can be found in [5]

$$D_x = l_x v^x,$$

$$C_x = d_x v^{x+1},$$

$$N_x = D_x + D_{x+1} + \dots + D_{\omega},$$

$$M_x = C_x + C_{x+1} + \dots + D_{\omega},$$

where ω means the maximum age used in life tables, that the probability to reach that age is going to zero. For example, in Czech Republic, $\omega = 105$.

Formula 4.2 shows the calculation of regular gross premium for a policyholder.

$$P_{x\overline{n}|}^{B} = \frac{A_{x\overline{n}|} + \beta \cdot \ddot{a}_{x\overline{n}|}}{(1-\gamma) \cdot \ddot{a}_{x\overline{n}|} - \alpha \cdot n},\tag{4.2}$$

where

 $\begin{array}{ll} P^B_{x\overline{n}|} & \text{regular gross unit premium,} \\ \gamma & \text{collecting expenses (in \%).} \end{array}$

The capital value of each policy will be calculated as cumulated amounts of saving premium components from the inception date, where the capital value is zero, till the valuation date. For evaluation of capital value we will use the technical interest rate.

$$CV_t = (CV_{t-1} + SP_t) \cdot (1 + TIR),$$
(4.3)

where

CV_t	fund (capital) value of one policy at the end of year t ,
$prem_t$	premium amount per one policy (single or regular),
SP_t	saving part of the premium,
TIR	technical interest rate.

As it was seen in the Table 4.1, the generated number of policies with regular premium is 994 and the generated number of policies with single premium is 1006. Table 4.5 shows the summary of variables of single and regular premium, capital value at valuation date and age of policy to valuation date. Variables of premium and capital value are calculated in CZK.

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Single prem.	65 344.4	102 922.0	149 292.0	152 371.0	194 128.0	256 848.0
Regular prem.	2 650.2	8 373.8	14 226.3	18 734.9	23 759.8	74 495.8
Capital Value	0.0	18 151.5	55 093.2	65 624.4	103 297.0	244 296.0
Policy Age	0.0	2.0	4.0	5.1	8.0	15.0

Table 4.5: The summary of dataset variables

The examples of our generated portfolio can be seen in the Table 4.14, that is in the appendix of the thesis.

4.3 Results of Calculation

In this section we will submit the comparison of time needed for calculation by used methods.

For each product, we will use all discussed methods, so we can compare the results of the best estimate of liability amount and time needed for calculation.

4.3.1 Death benefit as a summation of sum assured and fund value

Analytic function

Method of analytical function based on separation of the formula of cash flow into the parts that depend on the values of investment return and the part that does not depend. As a result we can evaluate the cash flow not policy-by-policy, but on the aggregated portfolio.

Table 4.6 shows the calculation time in seconds using policy-by-policy approach and analytic function approach. We can see that the time needed to calculate the present value of cash flow for our generating policies is more than 10 minutes, whereas the time needed for analytic function is about 20 seconds. As we discussed in section 2.1.1 the cash flow for such a product can be easily separated into required parts, so the average difference between the values of present value of cash flows is zero.

Time (in sec)	Time (in sec)	Mean of
Policy-by-policy	Analytic function	rel. errors
609.45	19.97	< 0.01%

Table 4.6: Time comparison of standard cash flow calculation and analytic function

In reality, the insurance companies have much more contracts in their portfolio and with analytic function they can reach the results within a reasonable time. The time difference in calculation for bigger companies can be from hours to many days depending on the size of the portfolio.

Cluster analysis: K-means

The idea of cluster analysis used within acceleration techniques of liability calculation is to speed up the calculation by decreasing the number of policies. This procedure will group the policies into clusers. For each cluster we will have a representer and a scale - a count of policies in each cluster.

We will create the clusters according to the value of present value of cash flows for each of the policy. We have run the policy-by-policy approach for one of the scenarios, and then we have to use the cluster techniques on scaled values of cash flows.

$$PVCF_{j}^{scale} = \frac{PVCF_{j} - \overline{PVCF}}{\sigma_{PVCF}}, \text{ where}$$
$$\overline{PVCF} = \frac{1}{J} \sum_{j=1}^{J} PVCF_{j} \text{ and}$$
$$\sigma_{PVCF} = \sqrt{\frac{1}{J-1} \sum_{j=1}^{J} \left(PVCF_{j} - \overline{PVCF}\right)^{2}}$$

where \overline{PVCF} is a sample mean, σ_{PVCF} is a sample standard deviation of the present value of cash flows, and J = 2000 is a number of policies in our portfolio.

For K-means method we created new model points as representers of the clusters. The new representers are results of average values of policies within each group.

The representer's continuous variables such as entry age, policy period, sum assured, premium and fund value at the valuation date are calculated as an average of each values from the cluster. For example:

$$SA_{representer} = \frac{1}{n_k} \sum_{l=1}^{n_k} SA_l,$$

where n_k is the number of policies in some cluster k, k = 1, ..., K.

The indicator variables such as policy type or sex for the representer model point are calculated as an integer value of each average.

$$Pol.type_{representer} = \left[\frac{1}{n_k}\sum_{l=1}^{n_k}Pol.type_l\right],$$

Our strategy for the selection of the number of clusters is based on minimalization of relative error of BEL values. We ran the cluster K-means 50 times to see how the relative error according to selected number of cluster can change. The results can be seen in Figure 4.5.

K-means, Numb. of clust. vs. Rel. error of calculation Relative error



Figure 4.5: Dependence of relative error of BEL on number of clusters K

To describe the dependence of relative error of callulation on the selected number of clusters (*Rel.error* ~ Kt) we have used the simple regression method, which is briefly described in the Appendix of this thesis, or can be found in [23]. Equation of red fitted line (linear trend) shown in the Figure 4.5 is:

$$\mathbb{E}[Rel.error] = 0.1993 - 0.0002K$$

The estimated coefficients can be interpreted as follows. The intercept, which is equal to 0.1993, is the expected mean value of relative error for zero number of clusters. The interpretation of intercept has no intrinsic meaning in this situation, because the number of clusters can't be zero. With every additional cluster in calculation we can expect the decrease of the relative error by an average of 0.02%. The time needed to process the liabilities increases with increasing of number of clusters.

From the Figure 4.5 we can see, that the relative error ranges from the values of 10% up to 20%. To compare the results of calculation, we select the observation with minimum error, which is 9.25% and related number of clusters is 551.

Table 4.7 contains the time comparison of results using the standard method and the cluster analysis. We can see that for this product in our example, using cluster analysis, it took about 3 minutes to calculate the present value of future cash flows.

Time[in sec]	Time [in sec]	Mean of
Policy-by-policy	Cluster Analysis	rel. errors
609.45	169.59	9.25%

Table 4.7: Time comparison of standard cash flow calculation for full portfolio and clustered

Cluster analysis: K-medoids

In K-medoids, we selected the representer as a policy with smallest average distance to all other observation in the cluster.

$$x_{medoid} = \operatorname*{argmin}_{y \in \{x_1, \dots, x_{n_k}\}} \sum_{i=1}^{n_k} d(y, x_i),$$

where n_k is the number of observations in the cluster k, k = 1, ..., K.

Figure 4.6 shows the dependence of absolute value of relative error on the number of clusters used in calculation.

K-medoids, Numb. of clust. vs. Rel. error of calculation



Figure 4.6: Dependence of relative error of BEL on number of clusters K

The equation of dependence between the absolute value of relative error and the number of cluster is:

$$\mathbb{E}[|Rel.Error|] = 0.0006 - 1.1512 \cdot 10^{-6} K$$

From the Figure 4.6 We can see, that the absolute value of relative error ranges from almost 0% to 0.1%. Even for small numbers of clusters, the K-medoids algorithm showed the accuracy of calculation about 99,9%.

For demonstration of the results, we have selected the observation with 264 clusters. The time needed to process such an amount of clusters is about one minute (Figure 4.8).

Time[in sec]	Time [in sec]	Mean of
Policy-by-policy	Cluster Analysis	rel. errors
609.45	78.06	< 0.01%

Table 4.8: Time comparison of standard cash flow calculation for full portfolio and clustered

Comparison

Table 4.9 introduces the results of above discussed method for the insurance product, where the death benefit is paid as a summation of sum assured and policyholder's fund value. Values of best estimate liability, calculated for our generated portfolio, are shown in milion of Czech Crowns. We can see, that the methods of analytic function and cluster K-medoids show the more precise result of BEL valuation. The clustered K-means method is more sensitive to noise and outliers in the cluster. The range of the relative error using the K-means algorithm were up to 20%. Compare the both: the time and the errors in calculation, the analytic function shows the best result. K-medoid shows also accurate results of calculation and as K-means, it can be used for any type of product without any additional settings.

	Policy-by-	Analytic	Cluster	analysis
	policy	function	K-means	K-medoids
BEL [in MCZK]	-154.25	-154.25	-168.53	-154.23
Time [in sec]	609.45	19.97	169.59	78.06
Abs. error		0.00	-14.28	0.02
Rel.error		< 0.01%	9.25%	< 0.01%
Numb.of clusters			551	264

Table 4.9: Results comparison

4.3.2 Death benefit as a maximum of sum assured and fund value

Analytic function

The idea of partitioning the formula of cash flows into parts that depend and doesn't depend on the return of investment is based on bisection method. We

are able to evaluate the "time of change" t^{*j} for each policy in the portfolio j and then calculate the cash flow according to the projection year t. For this type of product, we are also able to separate the cash flows formula on the parts that depend and doesn't depend on investment return.

The total time of calculation by analytic function also obtains the time needed to calculate the values of t^{*j} , which is about 7 seconds for our portfolio. We calculated "times of change" of death benefit for one selected interest rate scenario. Further, for the purpose of accelaration, we assumed the same values of t^{*j} for all other scenarios used in calculation.

Table 4.10 shows the time comparison of used methods. The time needed to calculate cash flows projection for our generated portfolio using standard policyby-policy method is about 10 minutes, whereas the time needed for analytic function is about 23 seconds. Even with added time of processing the values of t^{*j} the method shows very good difference in time.

Time (in sec)	Time (in sec)	Mean of
Policy-by-policy	Analytic function	rel. errors
617.262	23.743	< 0.01%

Table 4.10: Time comparison of standard cash flow calculation and analytic function

Cluster analysis: K-means

Figure 4.7 shows the output of 50 runs of cluster K-means method with different number of clusters.





Figure 4.7: Dependence of relative error of BEL on number of clusters K

From the Figure 4.7 we can see that the relative error is ranging from 10% up to 20% according the selected number of clusters. The equation of linear trend is:

 $\mathbb{E}[Rel.Error] = 0.2209 - 0.0002K$

Our aim of selection the number of clusters is minimize the error of calculation. For comparison of the results we have selected the observation with 604 clusters and the relative error about 9%. Table 4.11 shows the comparison of time of valuation of portfolio's BEL. The time needed to process the liabilities for this type of product using the standard "policy-by-policy" method is about 10 minutes, and using clustered K-means is about three minutes.

Time (in sec)	Time (in sec)	Mean of
Policy-by-policy	Cluster Analysis	rel. errors
617.262	184.82	8.91%

Table 4.11: Time comparison of standard cash flow calculation for full portfolio and clustered

Cluster analysis: K-medoids

Figure 4.8 shows the results of relative error depending on the number of clusters of 50 runs of K-medoids method.

K-medoids, Numb. of clust. vs. Rel. error of calculation Relative error



Figure 4.8: Dependence of relative error of BEL on number of clusters K

From the Figure 4.8 we can see the absolute value of relative difference of BEL ranges from 0.1% to 0.7%. Even with low number of clusters, the K-medoids presents the error in BEL lower that 1%.

The equation of linear trend is:

$$\mathbb{E}[|Rel.Error|] = 0.0035 - 5.9718 \cdot 10^{-6} K$$

Table 4.12 shows the time comparison needed to calculate the liabilities. For comparison we have selected the observation with 299 clusters and relative error equals to -0.01%. The time needed to process the liabilities using K-medoids is about minute and a half.

Time (in sec)	Time (in sec)	Mean of
Policy-by-policy	Cluster Analysis	rel. errors
617.262	94.62	-0.01%

Table 4.12: Time comparison of standard cash flow calculation for full portfolio and clustered

Comparison

Table 4.13 introduces the results of calculation using all discussed methods. Values of best estimate liability are calculated on the generated portfolio with a death benefit paid as a maximum amount of sum assured and policyholder's fund value. As for the first type of product, the method of analytic function shows the better result compared the both: time of calculation and deviation of the output value of BEL . However, the big disadvantage of analytic function is highly demanding initial preparations and settings.

	Policy-by-	Analytic	Cluster	analysis
	policy	function	K-means	K-medoids
BEL [in MCZK]	-144.28	-144.26	-157.12	-144.26
Time [in sec]	617.26	23.74	184.92	94.62
Abs. error		0.02	-12.85	0.02
Rel.error		-0.01%	8,91%	-0.01%
Numb.of clusters			604	299

Table 4.13: Results comparison

Conclusion

Proper and consistent valuation of liabilities is highly demanding task for insurance companies. The standard techniques of policy-by-policy cash flow projection estimates the insurance company's liabilities with high accuracy but usually it takes an extreme time. The purpose of this thesis was to present the two possible ways of acceleration valuation of life insurance liabilities.

Within this thesis, we focused on the unit-link insurance and discussed two products, which are mainly used in the companies. Products differs in payments of death benefit, for the first type of products it is defined as a summation of sum assured and policyholder's fund value and for the second it is the maximum amount of these two values. We introduced the main formulas and principles of calculation of liabilities in insurance companies. This thesis presents the accelaration techniques of analytic function and cluster analysis. We also presented the theory for interest rate modelling and we choose a Hull-White model to simulate the investment rates of return.

We ran all our calculations in Wolfram Mathematica software. We started with the simulation of 50 interest rate scenarios, then we set the assumption and generated the sample portfolio of 2000 policies in-force for life insurance company. We were able to use the standard "policy-by-policy" method, analytic function and cluster anlysis to calculate the present value of portolio's cash flows. The time needed to calculate the cash-flows projection for 2000 policies under 50 scenarios of interest rate using policy-by-policy approach was about 10 minutes for both types of products.

For each of two products the analytic function showed the best result according to the deviation of values and calculation time. To calculate the best estimate liability with analytic function took about 20 seconds for each type of the products and the relative errors in both cases is less than 0,01%. The method of analytic function described in this thesis can be used with not only strictly mathematical software but also in ordinary available softwares such as MS Excel. It also can be adjusted on other insurance products with death benefit paid as fund value of sum assured only, or with different types of maturity payments. The big disadvantage of the method is that initial preparation is highly demanding.

The method of clustering can decrease the number of policies and so decreases the final time of calculation. We also compared the effectivness of two clustering method: K-means and K-medoids. To select the number of policies needed for precise calculation and see the dependence between these values, we made 50 runs of each method and each type of product. The representer of cluster was selected as an average value within each group. The maximum precision that was obtained using the K-means in our eample for both types of products was about 90%. To obtained such a precision we needed to create the maximum possible number of clusters, as a result, more time was needed to calculation. For both products in our example it took about 3 minutes to calculate the best estimate liability.

More precise results were obtained using K-medoids clustering method. Compared to K-means, the K-medoids algorithm is less sensitive to outliers in the groups. The representer was choosen as the most centered policy within each group. K-medoids clustering method showed the high precision results of BEL even with low number of clusters. The presision of obtained results in our example for both products were about 99% and time needed to process the liabilities was about minute and a half.

The advantage of clustering methods is that it can be used for any type of data without any initial setting. The disadvantage can be producing different cluster results after each usage.

There are other possibilities to accelerate the calculations that are not discussed in this thesis, for example so-called interpolation approach that uses the grid scenarios [6], methods of antithetic variates or control variates. Further, calculation might be extended on calculation under the stress scenarios. This, however, would be a seperate topic to be researched.

Bibliography

- [1] Official Journal of the European Union. Directive 2009/138/EC of the European Parliament and of the Council of 25 November 2009 on the taking-up and pursuit of the business of Insurance and Reinsurance (Solvency II).
- [2] Official Journal of the European Union. II Non-legislative acts. Commission Delegated Regulation (EU) 2015/35 of October 2014.
- [3] S. Singor T.Possen. Dealing with options and guarantees. *Ortec Finance*.
- [4] M. Janeček. Valuation Techniques of Life Insurance Liabilities. PhD thesis, Charles university, Faculty of Mathematics and Physics, Prague, 2006.
- [5] T. Cipra. *Pojistná matematika teorie a praxe*. Druhé aktualizované vydání. EKOPRESS, 2006.
- [6] M. Janeček. Acceleration Techniques for Life Cash Flow Projection Based on Many Interest Rates Scenarios - Cash flow Proxy Functions. 2017.
- [7] M. Janeček. Techniques for substantial acceleration of life insurance calculations. 2017. Economic University in Prague.
- [8] Marija J. Norušis. IBM SPSS Statistics 19 Statistical Procedures Companion.
- [9] F.S.A. Emiliano A. Valdez, Ph.D. Data Clustering. 2018. University of Connecticut.
- [10] Kassambara A. Practical Guide to Cluster Analysis in R. STHDA, 2017.
- [11] Wu J. Gan G., Ma C. Data Clustering: Theory, Algorithms, and Applications. ASA-SIAM, 2007.
- [12] Wolfram Mathematica. Documentation Center.
- [13] Rousseeuw P.J. Kaufman L. Finding Groups in Data: An introduction to Cluster Analysis. John Wiley, 1990.
- [14] P.Tryfos. Methods for Business Analysis and Forecasting: Text and Cases. Wiley, 1998.
- [15] Wolfram Mathematica. Wolfram MathWorld. http://mathworld.wolfram. com.
- [16] Michal Hakala. Modely úrokových měr praktické aspekty. Master's thesis, VŠE, Fakulta informatiky a statistiky, Prague, 2017.
- [17] F.Mercurio D. Brigo. Interest Rate Models Theory and Practice. Springer, 2006.

- [18] European Insurance and Occupational Pensions Authority. Risk-Free Interest Rate Term Structures. https://eiopa.europa.eu/regulation-supervision/ insurance/solvency-ii-technical-information/ risk-free-interest-rate-term-structures.
- [19] Český Statistický Úřad. Úmrtnostní Tabulky. https://www.czso.cz/csu/ czso/umrtnostni_tabulky.
- [20] Tom M. Apostol. One-Variable Calculus, with an Introduction to Linear Algebra. Second edition. John Wiley, 1967.
- [21] O.John O. F. K. Kalenda M.Zelený V, Hájková. Matematika. Matfyz Press, 2006.
- [22] A. Komárek. Linear Regression. Course Notes. 2017.
- [23] André I. Khuri. Linear Model Methodology. CRC Press, 2010.

List of Tables

1.1	Cash flow of Endowment policy	13
4.1	The summary of simulated dataset variables	36
4.2	Mortality experience	37
4.3	Lapses assumptions	37
4.4	Other assumptions used in our example	38
4.5	The summary of dataset variables	39
4.6	Time comparison of standard cash flow calculation and analytic	
	function	40
4.7	Time comparison of standard cash flow calculation for full portfolio	
	and clustered	42
4.8	Time comparison of standard cash flow calculation for full portfolio	
	and clustered	43
4.9	Results comparison	43
4.10	Time comparison of standard cash flow calculation and analytic	
	function	44
4.11	Time comparison of standard cash flow calculation for full portfolio	
	and clustered	45
4.12	Time comparison of standard cash flow calculation for full portfolio	
	and clustered	46
4.13	Results comparison	46
4.14	Simulated policies in-force used in our example	52

$M \dots h = h$	Policy	Inception	Entry	0.000	Policy	Sum	D	CV at
INUITIDEL	$type^{a}$	date	age	Xəc	period	assured	Fremum	valuation date
	2	1.1.2012	23	0	15	250000	$240\ 965.0$	$169 \ 061.0$
2	2	1.1.2017	17	1	10	250000	$248 \ 512.0$	0.0
3	Η	1.1.2015	21	0	20	$120\ 000$	$8 \ 653.1$	$10\ 423.3$
4	Η	1.1.2014	19	1	ų	80000	$22 \ 720.4$	$48 \ 379.6$
ហ	2	1.1.2015	18	0	15	250000	$240 \ 950.0$	228 714.0
9	Η	1.1.2011	21	0	15	$200\ 000$	19 117.8	$56 \ 393.1$
2	1	1.1.2016	21	0	30	$200\ 000$	$9\ 785.1$	414.0
8	Ц	1.1.2017	17	0	10	$180\ 000$	25669.0	0.0
6	2	1.1.2015	17	1	40	80000	66 651.3	$62 \ 956.5$
10	1	1.1.2013	20	μ	10	$120\ 000$	$17 \ 072.4$	$40 \ 983.8$
			•••					
2000	-	1.1.2009	56	0	10	$180\ 000$	27501.9	82 253.2

Table 4.14: Simulated policies in-force used in our example

a 1 - regular, 2 - single b 0 - males, 1 - females

52

Appendix

The intermediate value theorem for continuous functions

Theorem: Let f be continuous at each point of a closed interval [a, b]. Choose two arbitraty points $x_1 < x_2$ in [a, b] such that $f(x_1) \neq f(x_2)$. Then f takes on every value between $f(x_1)$ and $f(x_2)$ somewhere in the interval (x_1, x_2) [20].

The proof can be seen in [20], or [21]

Simple Linear Regression

Basisc of linear regression can be found in [22], or [23].

We define Y as a response vector, and x_1, x_2, \ldots, x_k as a set of input variables [23].

$$Y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \epsilon,$$

where ϵ is an experimental error term associated with the measured, or observed, response at a point $\mathbf{x} = (x_1, x_2, \dots, x_k)^T$ in a region of interest, \Re , and $\beta_0, \beta, \dots, \beta_k$ are fixed unknown parameters. When, k = 1, the model is called a simple linear regression model.

Consider the simple linear regression model,

$$Y_u = \beta_0 + \beta_1 x_u + \epsilon_u, \quad u = 1, \dots, n,$$

where the ϵ_u 's are mutually independent with zero mean and variance σ^2 , and $u = 1, \ldots, n$ is an number of experimental run. The best linear unbiased estimators (BLUE) of β_0 and β_1 are

$$\hat{\beta}_1 = \frac{S_{xY}}{S_{xx}}$$
$$\hat{\beta}_0 = \bar{Y} - \beta_1 \bar{x},$$

where $\bar{Y} = \frac{1}{n} \sum_{u=1}^{n} Y_u$, $\bar{x} = \frac{1}{n} \sum_{u=1}^{n} x_u$, $S_{xY} = \sum_{u=1}^{n} (x_u - \bar{x})(Y_u - \bar{Y})$, and $S_{xx} = \sum_{u=1}^{n} (x_u - \bar{x})^2$.