University of Economics, Prague

Faculty of Finance and Accounting

The Department of Banking and Insurance

Field of study: Financial Engineering

Methods for Faster Estimation of Life Insurance Liabilities

Author: Bc. Markéta Švehláková

Supervisor: prof. Ing. Eva Ducháčková, CSc.

Academic Year: 2017/2018

Declaration:

I hereby declare that this master’s thesis, titled “Methods for Faster Estimation of Life Insurance Liabilities”, is my own work and that all the sources I used are properly marked and included in the reference list.

Prague, May 23rd 2018 …………………………….

Acknowledgements:

I would hereby like to thank prof. Ing. Eva Ducháčková, CSc. for her time, valuable advice and supervision regarding the compilation of this master's thesis.

Abstract

The thesis deals with methods for estimation of a life insurance company's liabilities and it aims to find a faster alternative to the traditionally used cash flow analysis. It studies two approximation methods, cluster analysis and stratified random sampling. The thesis is structured into four parts. First, it briefly characterizes life insurance, presents common types of life insurance contracts and shows a typical structure of a life insurance company's liabilities. Then, it describes the process of liabilities estimation using cash flow analysis and characterizes the properties of cluster analysis and stratified random sampling. Next, all the studied methods are applied to a modified portfolio of homogeneous life insurance policies and to a real portfolio of heterogeneous policies. Finally, the thesis compares the methods and defines the contexts in which an insurance company would prefer each of the tested methods.

Keywords: life insurance, life liabilities, cash flow analysis, cluster analysis, stratified random sampling

Abstrakt

Diplomová práce se zabývá metodami odhadu závazků životních pojišťoven a snaží se nalézt rychlejší alternativu k tradičně používané cash flow analýze. Zkoumá dvě aproximační metody: shlukovou analýzu a stratifikovaný náhodný výběr. Práce je členěna do čtyř částí. První část stručně charakterizuje životní pojištění, představuje běžné typy smluv a ukazuje typickou strukturu závazků životní pojišťovny. Druhá část popisuje proces odhadu závazků s využitím cash flow analýzy a ukazuje vlastnosti shlukové analýzy a stratifikovaného náhodného výběru. Třetí část prakticky testuje využití zkoumaných metod pro výpočet závazků z upraveného portfolia homogenních pojistných smluv a portfolia heterogenních smluv. Poslední část srovnává zkoumané metody a určuje, ve kterých situacích by pojišťovna zvolila kterou metodu.

Klíčová slova: životní pojištění, závazky životních pojišťoven, cash flow analýza, shluková analýza, stratifikovaný náhodný výběr

Notation

AIC	Akaike Information Criterion
CF	Cash flow
GLM	General linear model
i_t	investment return related to year t
j	model point number
LAT	liability adequacy test
l_t	lapse rate related to year t
m	number of model points
n	policy period
N_t	expected number of policies in force at the beginning of year t
PV	present value
PVCF	present value of future cash flows
q_x	the probability of death between the ages of x and x+1
t	policy year
x	age of the insured person at policy inception

Table of Contents

Introduction

The value of a company's liabilities is key to both financial modeling and accounting. Unfortunately, the task of calculating liabilities of life insurance companies is rather demanding since it not only requires evaluating the current state but also predicting the future development of all the cash flows from every policy in force. This significantly increases the number of computational steps of each calculation. In addition, it makes it necessary to repeat the computational process multiple times under different assumptions, because it is not possible to determine the future value of all the variables entering into the computation and one needs to test various scenarios. Consequently, most traditional methods intending to obtain the accurate value of life liabilities are extremely time-consuming and even with modern technologies, the results may be derived with high delay.

The problem of the lengthiness of the life liabilities computational process is a widely discussed topic and there is a high demand for a more efficient solution. To prove this, a grant project has been advertised at the University of Economics, Faculty of Informatics and Statistics whose aim was to test cluster analysis and its applicability to the calculation of liabilities from a homogeneous portfolio of life insurance contracts. Engaging in the project, I became interested in the possibilities of faster estimation of life liabilities and decided to study the topic more deeply. The thesis continues the work started by the project and seeks to optimize the settings of the clustering algorithm and to test other methods as well. In addition, it not only applies the methods to the homogeneous portfolio but also to a heterogeneous portfolio of life insurance policies.

The aim of the thesis is to find an alternative method to the traditionally used cash flow analysis which would enable a faster estimation of life liabilities. The thesis studies two approximation methods, namely cluster analysis and stratified random sampling, and it seeks to optimize the settings of each one. In addition, it compares the methods with each other and with cash flow analysis and tries to identify the strengths and weaknesses of each one considering different requirements of the insurance company.

The methods are tested using a real portfolio of life insurance policies. All the methods are first applied to a modified homogeneous portfolio where some properties of the policies are fixed and the settings of the methods can be optimized more easily. Only then, the methods are applied to a real portfolio. The comparison between the methods and between their different settings is based primarily on the accuracy of the estimated results and the computational time of one scenario. All the tests are performed using R software.

The thesis is structured into four chapters. The first chapter briefly characterizes the nature of life insurance and it presents the most common products offered by life insurance companies. Apart from this, it shows a typical structure of a life insurance company's liabilities and discusses the aspects which make the estimation of life liabilities such a challenging task.

The second chapter presents the three tested methods. First, it describes the process of computation of life liabilities using the traditional approach. It mentions various assumptions used in the calculation and the way the assumptions are made. Then, it explains the process of the calculation of the best estimate liability, LAT reserve and the sensitivity of the liabilities to the changes in the assumptions. Second, it deals with cluster analysis. It describes the steps of the analysis, explains the clustering algorithm and mentions how each step should be modified if the method is applied to the calculation of life liabilities. Finally, the chapter characterizes stratified random sampling and describes the process of its application to the life liabilities calculation.

The aim of the third chapter is to test the applicability of the methods to the calculation of a life insurance company's liabilities. It first applies each method to the homogeneous portfolio and compares different settings of each method in order to optimize the methods. Then, the methods are applied to the real portfolio of life insurance policies and the properties of the results, as well as the overall efficiency of the computational process, are measured.

The fourth chapter compares the results obtained using each method and it seeks to find the strengths and weaknesses of each one. The comparison is based primarily on the accuracy of the approximation, variability of the result and computational time. The chapter also aims to define the contexts in which an insurance company might prefer each of the tested methods.

1 Characteristics of Life Insurance and Life Insurance Liabilities

To be able to study various methods for life insurance liabilities estimation, it is necessary to understand the nature of life insurance and to be familiar with the characteristics of life insurance companies as well as the structure of their liabilities. Therefore, the first chapter gives some general information about life insurance companies, characterizes life insurance, presents various types of life insurance products and shows a typical structure of life insurance companies’ liabilities.

1.1 Characteristics of Life Insurance

Like any form of insurance, life insurance represents a means of protection from financial loss incurred due to an uncertain event. The uncertainty in the case of life insurance is connected with human life and the impossibility to predict its length. Thus the insured event can be of two types - death and survival of the insured person. Originally, life insurance only provided a protection against the risk of death and it was intended to give a financial support to the surviving dependents of the insured person. Later, insurance companies started to offer the protection against the second type of risk and enabled their clients to obtain a financial support if they were unable to work due to their advanced age or disability. Since then, many different product types have developed and nowadays, life insurance can not only be used as a protection but it can also serve as an investment instrument.

In legal terms, life insurance could be defined as a contract between a policyholder and an insurance company where the policyholder is obliged to pay either regularly or in a lump sum a policy premium and the insurance company promises to pay a sum of money to a designated beneficiary in the case of death or other prespecified events (Whelehan 178).

1.2 Types of Life Insurance Products

Depending on the events that are insured, life insurance products can be classified into three main groups - life insurance contracts, survival benefit contracts and hybrid contracts. Each group is characterized in the following section.

1.2.1 Life Insurance Contracts

A basic life insurance contract involves the payment of a premium, either monthly or at other intervals, in return for a lump sum payment upon the death of the policyholder. Life insurance contracts are primarily purchased to mitigate the financial consequences of early death for surviving relatives though they may include an investment component as well.

There are various types of life insurance contracts. The primary types include whole life insurance and term insurance. Whole life insurance pays a lump-sum cash amount upon the death of the insured person, whenever that may occur. Term insurance also pays a lump-sum cash amount upon the death of the insured person but only provided that it occurs during the defined policy term (Camilli, Duncan and London 48).

1.2.2 Survival Benefits Contracts

The simplest type of survival benefits contracts is pure endowment which is a contract promising to pay the insured person a stated sum if he or she survives a specified period with nothing payable in case of prior death (Merriam-Webster).

A slightly more complex type of survival benefits contracts is an *annuity contract*. A basic annuity contract involves a single-sum premium payment or a series of premium payments, in return for a regular series of future payments, conditional on the survival of the contract holder. Annuity contracts are typically purchased to mitigate the risk of outliving one’s savings and they may be an alternative or a supplement to investing in various pension funds (Camilli, Duncan and London 49).

There are different types of annuity contracts. The main distinction is between a whole life annuity, whose payments are contingent on the annuitant’s survival, and a temporary annuity, whose payments continue until the earlier of the annuitant’s death or a certain pre-specified length of time. Another distinction is between an immediate and deferred annuity. In the case of an immediate annuity, the payments begin immediately after the contract comes into force while in the case of a deferred annuity, the payments begin at a specific date in the future. There are also annuities linked to more than one life such as a last survivor annuity which offers payments until the death of the last person in the group, and a joint life annuity in which case the payments already cease upon the first death in the group (Camilli, Duncan and London 50).

1.2.3 Hybrid Insurance Contracts

Hybrid insurance contracts combine the previous two groups of contracts and they guarantee a benefit payment either upon the death of the insured person or upon their survival of a pre-specified time period. The most common types of hybrid insurance contracts include endowment insurance and unit-linked insurance plan.

Endowment insurance pays a lump-sum cash benefit upon the earlier of the death of the policyholder, or the end of a specific time period (Camilli, Duncan and London 48).

Unit-linked insurance plan differs significantly from all the life insurance products mentioned so far because it combines life insurance with some form of an investment. Thus it is sometimes referred to as life assurance while all the above-mentioned products, which are only intended to cover a risk of an uncertain event, are referred to as life insurance. However, both the terms are often used interchangeably.

Unit-linked insurance plan allows the policyholder to allocate a portion of the premium to a separate account comprised of various instruments and investment funds within the insurance company's portfolio. The amount invested in the investment fund is market-linked and it appreciates with increasing value of the fund. Thus the investment risk is carried by the purchaser of the insurance, not the insurance company (Česká Asociace Pojišťoven).

1.2.4 Current Situation

The dominating types of life insurance products in the Czech insurance market are unit-linked insurance and simple life insurance. On the other hand, annuities and endowment insurance play a negligible role and many insurance companies do not offer these products anymore.

Most marketed contracts are highly flexible and they give the policyholders freedom in terms of premium paid since only the risk premium, that is to say, the premium covering the risk of death of the insured person, is set and the clients can choose the value they want to invest in the separate account themselves.

Another current market trend is to offer supplementary insurance next to life insurance contracts. Such insurance may cover, in addition to the risk of death, the financial risk of being unable to generate income due to disability, lengthy hospital stay, long-term illness etc. (Česká Asociace Pojišťoven).

Since the dominating types of life insurance products in the Czech insurance market are policies offering a single claim benefit payment in case of death, this text concentrates primarily on these products.

1.3 Life Insurance Companies' Liabilities Structure

The liabilities structure of a life insurance company is closely connected with the types of products the insurance company offers. Figure 1.1 shows a typical structure of life insurance companies' liabilities in the Czech Republic. One can notice that the greatest proportion of liabilities is represented by technical reserves.

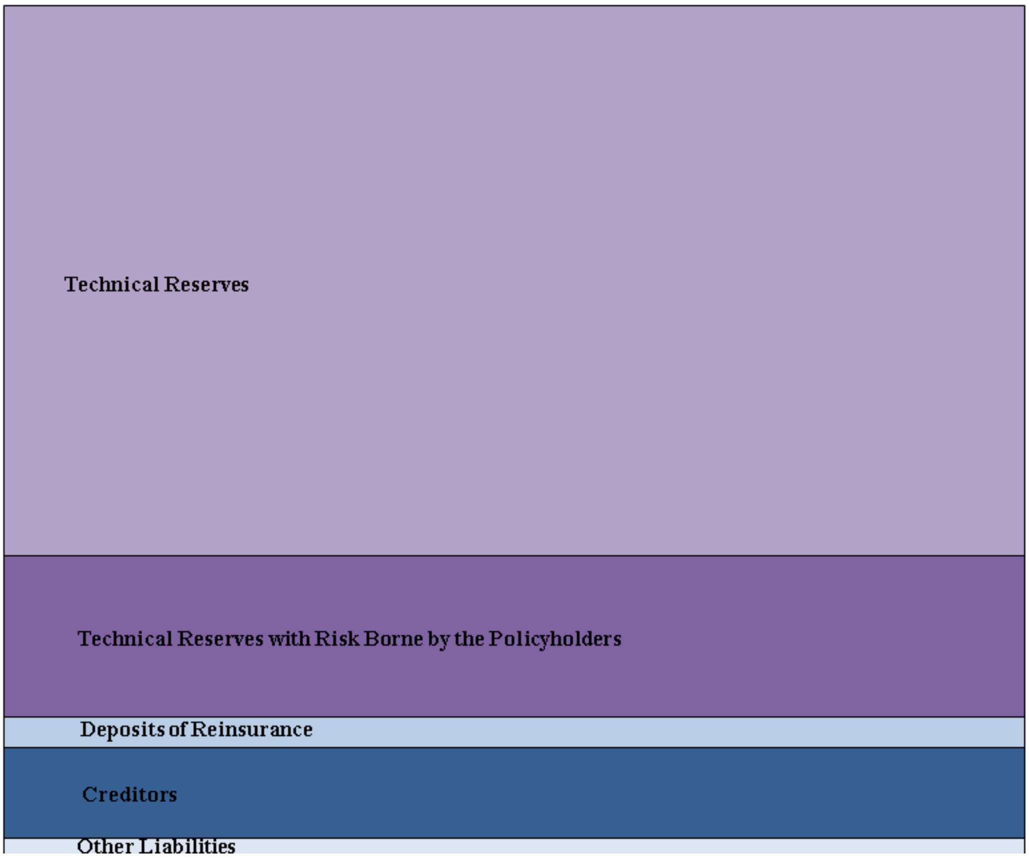

Figure 1.1: Liabilities Structure
Source: Personal collection

Technical reserves are liabilities for amounts an insurance company is obliged to pay in accordance with the policy contract. They include the benefits the company has promised to pay in case of a claim, as well as all the expenses related to the policy contracts such as agent commissions, marketing costs, transmission costs, staff salaries and other administrative costs. (Janeček 5).

The amounts an insurance company is obliged to pay are usually uncertain as to the exact amount and the time of payment. About 95 % of reserves are represented by two categories: claim reserves and policy reserves. Claim reserves are held because the event insured against has already happened, but the amount of claim is not yet known by the insurance company since the claim has not yet been reported or insufficient information has been furnished. A much greater proportion of reserves is represented by policy reserves, which are held because the event insured against has not yet happened but the company is obliged to pay if the event does happen (Lombardi 1).

Technical reserves are not on the same level in every policy year. Even if the insurance company does not make any new contract, the reserves can change over the policy period, for example, owing to the changes in interest rates, expenses, mortalities, lapses etc. This makes the reserving process even more difficult as the insurance companies cannot simply evaluate the new policies but they need to constantly evaluate all the policies in their portfolio.

As the payment amount and time are not certain, the reserving process requires the use of various assumptions, estimates, and judgments about the future. The primary assumptions are expenses, investment returns, mortality, morbidity, voluntary terminations (i. e. expiries, lapses, surrenders, and withdrawals) and taxes. These assumptions are usually based on the company's past experience, industry studies, regulatory requirements, and judgments about the future (Lombardi 3).

It is the calculation of technical reserves that makes the calculation of liabilities of life insurance companies such a demanding and time-consuming task. When calculating other liabilities, one can use the accounting methods applicable to calculation of liabilities of most companies. However, the process of technical reserves estimation differs significantly since it requires the use of stochastic modeling.

For this reason, the thesis focuses on the calculation of technical reserves and on the methods which may make the calculation more efficient, because once the technical reserves have been determined, it is relatively simple to calculate total liabilities of the life insurance company.

2 Valuation Methodologies

The second chapter of the thesis introduces three methods for estimation of life insurance liabilities: cash flow analysis, cluster analysis, and stratified random sampling. The first method is the one most commonly used by insurance companies. The two other methods are based on the first one though they seek to simplify and speed up the calculation process by using a certain level of approximation.

2.1 Cash Flow Analysis

The first method discussed in the thesis is cash flow analysis which is a method commonly used by Czech insurance companies. The main objective of the method is to forecast future cash flows from each contract in the company's portfolio and calculate their sensitivity to the changes in the input variables.

In order to achieve maximum accuracy, cash flows are modeled individually for each policy based on its terms and conditions and the client's characteristics. Alternatively, the contracts can be grouped into various model points which represent a cohort of policies with the same important characteristics such as product type, the client's sex and age at entry, annual premium, the frequency of premium payments, policy term, sum insured etc.

The predicted development of future cash flows can be further used for estimating the company's technical reserves and thus its liabilities. In addition, it can be used for calculating various profit metrics, the LAT reserve, and many other important metrics.

The method seeks to show a realistic state of the company. For this reason, it avoids using predefined and unchanging assumptions, as is the case with traditional or statutory valuation approach often used in the past for valuation of life insurance liabilities. Instead, best estimate of future evolution is assumed. However, as the model is predicting future development from each contract or model point individually, it may be rather complex and time-consuming and some assumptions are unavoidable.

2.1.1 Assumptions

The evolution of future cash flows is determined by the evolution of mortality, lapse rate and various economic variables such as investment return and expenses. The insurance company does not know precisely when the benefits and service costs will occur, nor do they know exactly how much they will be. Therefore, the process of liabilities estimation requires the use of various assumptions and estimates.

All these assumptions and estimates are set on the level of their best estimates, that is to say, on the level of their expected value. The best estimates should be based on the company's past experience, industry studies, regulatory requirements and judgments about the future (Lombardi 3).

Mortality

In order to forecast future claims, the company needs to forecast future mortalities. Their prediction is usually based on mortality tables of a general population. However, the tables need to be adjusted since the clients' mortality tends to be lower than that of a general population. This tendency is caused by various factors. For policies insuring the risk of death, the clients' mortality is generally lower because insurance companies can compel the insured person to attend an insurance medical examination prior the conclusion of a contract and they may refuse to insure a person that is not in good health. On the other hand, the expected mortality of the policies insuring the risk of survival such as annuities and endowments are lower due to the fact that these insurance products are purchased almost exclusively by healthy people who expect to profit from them.

Therefore, mortality rates are calculated as a product of general population mortality rates and a selection ratio which is a ratio reflecting the relationship between the general mortality rate and the clients' mortality rate. The ratio should be estimated using the company's past experience and it should approach one with increasing time because the clients' mortalities go to general mortalities in time.

Next to policy type and policy year, the estimated mortality rates are based on sex[1] and age of the insured person. In addition, smoker status, profession and other client's characteristics potentially influencing their health status may be taken into consideration.

Table 2.1 shows an example of the expected mortality rates calculation. The mortality rates are based on the Czech mortality tables for a forty-year-old man. One can notice that the selection ratio for life insurance is lower than the one for annuities. This is often the case since insurance companies have the power to influence the underwriting process more

[1] Council Directive 2004/113/EC on equal treatment between men and women in access to and supply of goods and services bans European insurance companies from using gender to set the charges of life insurance products on the grounds of sex. However, even though insurance companies have to charge the same prices to women and men, gender is still used as one of the most important criteria for mortality modeling for the purpose of forecasting future cash flows, profits etc.

strongly than the client. However, the ratios depend on the company's personal experience and they differ between companies. Some insurance companies (e.g. Česká podnikatelská pojišťovna) use Czech Statistical Office's mortality rates without applying any adjustments to the rates at all.

Policy year	**Selection ratio (life insurance)**	**Selection ratio (annuities)**	**Mortality rate (life tables)**	**Expected mortality rate (life insurance)**	**Expected mortality rate (annuities)**
1	22%	70%	0.001473	0.00032406	0.0010311
2	33%	75%	0.001682	0.00055506	0.0012615
3	44%	80%	0.001894	0.00083336	0.0015152
4	55%	85%	0.002049	0.00112695	0.0017417
5+	66%	90%	0.002318	0.00152988	0.0020862

Table 2.1: Expected Mortality
Source: Personal collection based on CSO and Allianz

Lapses

Insurance companies need to find the best estimate of future lapse rates as well as some forecast for how the actual lapses may depart from the best estimate if the economic environment changes. Unfortunately, there are no lapse rates estimated by the Czech Statistical Office which actuaries could use as reference values as it is the case for mortality rates. Therefore, they have to estimate the values themselves using the company's experience. However, many companies have insufficient data for the estimation and even with sufficient data, it is extremely difficult to forecast the lapse rates because the lapses are strongly influenced by the changes in the economic environment.

This could be illustrated with the example of the Czech insurance market development at the turn of the century. The Czech insurance market only started developing in the 1990s after the state insurance monopoly had been abolished. Life insurance products soon became popular, but due to lack of experience of the newly established insurance companies, the products were often of low quality. This resulted in a rapid increase in the number of withdrawals in 2000 and the lapse rate increased by about 6 percentage points (Janeček 21).

When developing a lapse function, one should take into consideration the following factors:

1. Policy year - usually the lapse rates are highest in the first policy year and they tend to decrease every year.

2. Product type - products that are primarily investment products such as single premium deferred annuities might be more subject to excess lapsation than more protection-oriented products.
3. The presence and level of any surrender charges - the higher the surrender charges are, the less likely the policyholders are to cancel their life insurance policy.
4. Investment return on the policy compared to the market interest rate (Lombardi 231).

As actuaries often do not have sufficient data source, policy year is usually the only variable used for constructing the lapse rate function. Table 2.2 shows lapse rates in the Czech insurance market in 2015. One may notice that more than 50 % of the policies are canceled within 5 years from the issue.

Policy year	1	2	3	4	5
Lapse rates	13.7%	12.2%	10.6%	9.4%	9.0%
Cumulative	13.7%	25.9%	36.5%	45.9%	54.9%

Table 2.2: Lapse Rates
Source: Matoušek (5)

Commissions

Commissions represent the primary way that producers get paid for selling life insurance. They are typically determined by the type of insurance, being higher for products with higher margin, and they are tied to premiums generated. One can distinguish between two main types of commissions – initial and renewal.

Initial commissions are paid in the first policy year or sometimes in the first few years. After this period, the renewal commissions are paid for a specific number of years. Most companies pay commissions annually as a percentage of the premium generated in each policy year. The percentage is typically much higher in the case of initial commissions and it equals something between 40% and 90% (depending on the company and product) of the premium paid during the first year. If the initial commissions are paid for more than one year the percentage usually decreases every policy year. On the other hand, renewal expenses are generally much lower and they range from about 2% to 5% of premiums paid into the policy during those specified years, but it can be higher depending upon the commission structure of the company (Rockford).

Forecasting the value of commissions is connected with forecasting future premium payments. Since the percentage of premium paid as commissions is fixed, the estimation does not require building an extra model.

Expenses

Insurance companies need to be able to estimate their future expenses. The expenses could be classified into two groups – initial and maintenance. Initial expenses are one-off expenses related to the policy inception. They include medical underwriting, product marketing and remuneration of sale forces excluding commissions (Janeček 23).

Maintenance expenses include the expenses incurred every policy year to keep the policies in force. They include the expenses of contribution billing and collection, contribution tax payment, policy record maintenance, accounting, valuation, pricing and other policyholder services (Camilli, Duncan and London 440).

Expense charges allocated to a contract could be divided into four groups according to the way they are calculated. The first group is charges calculated as a percentage of the gross premium. These expenses arise as the premium is paid and they include for example state premium taxes. Some initial expenses are included in this group, which is why they are usually higher in the first policy year. The second group includes charges calculated as a fixed amount per unit of face value. The amount per unit of insurance might be expected to cover such things as underwriting expenses (i.e. risk classification), policy issue expenses and subsequent policy maintenance expenses. Like the percent of premium expenses, these expenses can be expected to be higher in the first year than in the subsequent years of the policy. The third group includes charges calculated as a fixed amount or percentage of benefit amount incurred when the benefit payment is made. These expenses are known as settlement expenses and they are one-time charges incurred at the same time as the benefit is paid. Therefore they can be added to the liabilities model by simply adding them to the benefit payment amount. The final group is the expenses which are calculated as a fixed amount for the contract itself, regardless of benefit amount. Again, some initial expenses are included in this group and they are, therefore, higher in the first year (Camilli, Duncan and London 225).

Unlike commissions, the expenses can change during the policy period. They naturally increase due to the inflation but some expenses may decrease for example due to the technological progress. All these changes should be taken into consideration. However, it is very difficult to forecast the development of expenses, which is why inflation is usually the only aspect taken into consideration. As salaries represent large part of the expenses, not only consumer price index but also wage index should be taken into account.

Table 2.3 shows an example of expense inflation calculation based on consumer price index and wage index in the Czech Republic. It is noticeable that the inflation is only modeled up to 2019 and constant growth rate is assumed after that. This is because it is quite difficult to forecast future inflation and Czech National Bank does not offer any further prediction. Therefore, the model should be updated frequently and newly available information on inflation development should be added.

Year	CPI (annual growth rate)	Salary inflation	Salary part of total expenses	Expense inflation
2014	0.4%	4.7%	60%	3.0%
2015	0.3%	1.3%	60%	0.9%
2016	0.7%	2.4%	60%	1.7%
2017*	2.4%	2.9%	60%	2.7%
2018*	2.4%	2.7%	60%	2.6%
2019+*	2.0%	2.1%	60%	2.1%

Table 2.3 Expense Inflation

Source: Personal collection using data from CSO, CNB, Janeček

Interest Rates

Interest rates play an important part in the model. The insurance company invests the client's funds in financial markets and it can retain a percentage of the profit. On the other hand, some contracts guarantee the client a minimum investment return and should it happen that the investment return is lower than the guaranteed rate, the insurance company is obliged to cover the difference between the rates. Next to the investment return, the actuaries also need to determine discount rate which is used for discounting the projected future cash flows.

For both the investment return and the discount rate, the best estimate should be used. The best estimate of the investment return should reflect the expected future returns on the assets held at the valuation date. Reinvestment of future cash flows should be taken into

* Forecast

consideration as well and its valuation should be in accordance with the future investment strategy agreed within the company (Janeček 24).

Other Assumptions

There are many other assumptions that need to be taken into consideration. For example, with some contracts, the policyholder has the right to increase or decrease the sum assured or the premium, some contracts allow partial withdrawals before the end of the policy period or the policyholder may be given the right to stop paying the premium for a certain time period.

2.1.2 Cash Flow Projection

Cash flow projection for each contract is based on the following characteristics of the contract:

- Product type (the event insured against, specification of benefit payment)
- Client’s characteristics (sex, age at entry)
- Premium (annual premium amount, payment frequency)
- Inception year
- Policy period
- Sum assured
- Current value of fund

It is also possible to add other characteristics such as guaranteed interest rate or profit share. Groups of policies with the same characteristics represent individual model points for which the projected cash flows are calculated.

Cash flow from model point j in valuation year t is given by

$$\begin{aligned} CF_{j,t} = {} & Premium\ written_{j,t-1} - Surrender\ paid_{j,t} - Claims\ paid_{j,t} \\ & - Maturity\ paid_{j,t} - Commissions\ paid_{j,t-1} \\ & - Expenses_{j,t-1} \end{aligned} \tag{1}$$

where

$$Premium\ written_{j,t-1} = Annual\ premium_{j,t-1} * N_{j,t-1}$$

$$Surrender\ paid_{j,t} = Accounting\ reserve_{j,t} * (1 - Surrender\ fee_{t})$$

$$Claims\ paid_{j,t} = Expected\ number\ of\ claims_{j,t} * Claim\ Benefit_{j,t}$$

$$Maturity\ paid_{j,t} = Expected\ number\ of\ maturities_{j,t} * Maturity\ payment_{j,t}$$

$$Commissions\ paid\ _{j,t-1} = Commissions_t * N_{j,t-1}$$

$$Expenses\ _{j,t-1} = Expected\ expenses\ per\ policy_t * N_{j,t-1}$$

N_{t-1} denotes the expected number of policies in force at the beginning of year t and it can be obtained using the recursive formula

$N_t = N_{t-1} * (1 - expected\ q_x) * (1 - l_t)$ for $t < policy\ policy\ period$

$N_t = 0$ for $t \geq policy\ period$

Using equation (1), one can forecast future cash flows from each model point for each year. These values can be further used to obtain the expected present value of future cash flows from all policies of the insurance company using

$$PV(CF) = \sum_{j=1}^{m} \sum_{t=1}^{n} CF_{j,t} \cdot \prod_{k=1}^{t} \frac{1}{1+i_k} = \text{-BEL} \qquad (2)$$

where m denotes the number of model points, n is the number of policy years to maturity and i_k is the expected investment return at time k.

The present value of future cash flows can be used to obtain the value of technical reserves or best estimate liability and thus determine total liabilities of the life insurance company. In addition, the cash flow model can be used for determining other values for each model point, namely present value of future gross and net profit, present value of distributable earnings, and present value of premium. The values are calculated as the sum of discounted expected values of future gross and net profit, distributable earnings and premium respectively. In addition, the model can be used for calculating various profit criteria which can be used for evaluating the profitability of each model point.

An example of three profit criteria can be seen below.

$$\text{Profit criterion}_1 = \frac{\text{PV(Gross profit)}}{\text{PV(Premium)}}$$

$$\text{Profit criterion}_2 = \frac{\text{PV(Net profit)}}{\text{PV(Premium)}}$$

$$\text{Profit criterion}_3 = \frac{\text{PV(Distributable earnings)}}{\text{PV(Premium)}}$$

2.1.3 Sensitivity

It is not enough to calculate the best estimate liability because the expected future cash flows might be highly sensitive to modest changes in some of the many assumptions and the reserves based on best estimate liability may not be sufficient. Therefore, actuaries need to perform sensitivity testing of liabilities and economic profit.

To evaluate the sensitivity of the portfolio to a change in an assumption, actuaries calculate the expected value of future cash flows or other metric using the changed assumption while holding all other assumptions unchanged. The sensitivity is then calculated as the percentage change in the value of the calculated metric in response to a one percent change in the tested variable.

Sensitivity calculation enables actuaries to estimate negative impact of assumption changes and thus calculate the liability adequacy test required by the International Reporting Financial Standard 4 for insurance contracts. It takes into account possible unfavorable changes in all the input parameters and calculates the reserves based on the most unfavorable scenario. Holding the LAT reserve should ensure that the insurance company holds sufficient reserves to meet their liabilities at all times (Perna and Sibillo 75).

Sensitivity calculation significantly increases the computational time as it requires applying the formulas multiple times under various assumptions. Finding a faster estimation enables actuaries to test more scenarios and obtain a more detailed information concerning the possible future development of the value of liabilities.

2.2 Cluster Analysis

Cluster analysis is an explanatory analysis that tries to identify similarities and dissimilarities within a set of objects and based on these similarities, it groups the objects into disjoint subsets known as clusters so that the entities in the same cluster are alike and dissimilar to the entities in other clusters. (Hennig et al. 2).

All objects belonging to the same cluster can be characterized using a single object assigned some kind of weight. This enables one to reduce the size of the dataset to the number of clusters. The reduced dataset of representative objects allows a better insight into the tested dataset and it reduces the computational time of all calculations performed on the dataset.

Cluster analysis of life insurance policies portfolio comprises six main steps (Romesburg 9).

1. Obtain and standardize the data matrix
2. Select a distance measure
3. Execute the clustering method
4. Create the representative portfolio
5. Test the precision

All these steps are further described in the following section.

2.2.1 Data Matrix

The first step of the cluster analysis is to construct a data matrix. In this matrix, columns stand for observations whose similarities one wants to estimate and its rows stand for clustering variables the values of which are used to determine the similarities (Romesburg 10). In the case of insurance portfolio dataset, the columns stand for individual model points and the rows stand for the chosen clustering variables.

A clustering variable can be any variable, the value of which one wants to closely reproduce with the compressed model. One might use attributes which are already available for each model point such as the characteristics of the insured person or the properties of the policy. These variables include age, annual premium, policy term, sum assured etc. Such variables can be easily obtained and they can describe well the similarities between the characteristics of each contract but they may not lead to very accurate results when estimating liabilities due to their ambiguous impact on the cash flow development. Therefore, it seems more appropriate to use derived variables such as individual cash flows values, expected present value of future cash flows, future gross and net profit, distributable earnings or premium (Freedman 6). Although these variables need to be calculated for each model point first, which increases the computational time, they better describe the dynamics and development of each contract and they should lead to more accurate results when calculating liabilities development.

Measuring the distances between model points is closely related to the scale on which measurements are made. For this reason, some data matrices need to be standardized before applying cluster analysis. Standardizing the data matrix converts the original variables to new unitless variables in order to make them comparable (Romesburg 78).

This step is particularly needed when using characteristics of the model points such as age as attributes. For instance, it is obvious that one year difference in the age of the insured person has a different impact on the liabilities value than one crown difference in the sum insured.

The most widely used standardizing function is

$$Z_{ij} = \frac{X_{ij} - \bar{X}_j}{S_j} \tag{3}$$

where X_{ij} represents the value of the *j*-th clustering variable of the *i*-th model point, $\bar{X}_i$ is the mean of the *j*-th clustering variable and S_j is the standard deviation of the *j*-th clustering variable (Romesburg 78).

2.2.2 Clustering Distance Measures

The classification of observations into groups requires a method to compute the distance between each pair of observations. The result of this computation is known as a distance or dissimilarity matrix whose elements a_{ij} represent the dissimilarity coefficients which measure the distance between observations *i* and *j* (Kassambara 25).

There are many methods to calculate the dissimilarity coefficients. They are typically calculated using a mathematical formula whose choice may be a critical step in clustering, because it may determine the shape of the clusters (Kassambara 26).

The most common distance measure is Euclidean distance. Euclidean distance between the *i*-th model point MP_i and the *j*-th model point MP_j is defined as

$$d_{euc}(MP_i, MP_j) = \sqrt{\sum_{k=1}^{K} (X_{ki} - X_{kj})^2} \tag{4}$$

where K is the number of variables.

Another widely used distance measure is Manhattan distance defined as

$$d_{man}(MP_i, MP_j) = \sum_{k=1}^{K} |X_{ki} - X_{kj}| \tag{5}$$

It is also possible to measure distances using correlation-based measures the most commonly known being Pearson correlation distance defined as

$$d_{cor}\left(MP_i, MP_j\right) = 1 - \frac{\sum_{k=1}^{K}(X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j)}{\sqrt{\sum_{k=1}^{K}(X_{ki} - \bar{X}_i)^2 \sum_{k=1}^{K}(X_{kj} - \bar{X}_j)^2}} \tag{6}$$

where $\bar{X}_i$ is the mean of the *i*-th model point (Kassambara 26).

2.2.3 Clustering Method

There are many clustering methods, each of which is applicable to different dataset types. Most methods are only suitable for datasets of a limited size and applying them to a large portfolio would be extremely time-demanding and inefficient. This is not the case with CLARA algorithm which is a special clustering algorithm derived for large datasets in order to reduce the computational time and RAM storage requirements (Kassambara 48). As the aim of this thesis is to apply a clustering method to a large portfolio of policies, CLARA algorithm seems a suitable clustering method.

CLARA algorithm is a special form of k-medoids algorithm adapted to large datasets. Generally, k-medoids algorithms aim to partition a dataset into *k* clusters and represent each cluster with one of the model points from the cluster known as a medoid (Kassambara 48). This enables one to reduce a portfolio of any size to a portfolio of *k* representative model points.

Objects selected as medoids are model points with a minimum average distance to all other members of the relevant cluster. They correspond to the most centrally located model points in each cluster and they can, therefore, be considered representative examples of the members of the cluster they belong to.

There are other methods for setting the representative model point in each cluster such as using the mean of all model points within the cluster. However, such methods might be more sensitive to noise and outliers and they may lead to biased results (Kassambara 48). For this reason, only the use of medoids is tested in this thesis.

The process of finding the medoids is based on PAM algorithm. The PAM algorithm is an iteration process which tries to find *k* clusters among the model points by analyzing all possible model points pairs such that one of them is a medoid and the other is not and repeatedly trying to make a better choice of medoids (Ray and Acharya 44).

The algorithm comprises several steps. First, the initial choice of *k* medoids is made. The medoids can either be provided or they are randomly selected. Second, the algorithm measures the distances between the medoids and all other model points using the selected distance measure. Third, each model point is assigned to its closest medoid to create *k* clusters. Finally, each cluster is tested for an object which could decrease the average dissimilarity coefficient. If such object is found in any of the clusters, it is assigned a new medoid and the whole process of clusters creation is repeated with a new set of medoids (Kassambara 49).

CLARA algorithm is an extension to PAM algorithm which utilizes the technique of sampling in order to reduce the computational time. It starts with selecting a random sample of a fixed size from the dataset. In this sample, *k* medoids are found using PAM algorithm. The medoids selected from the random sample are used for clustering the whole dataset and each model point from the dataset is assigned to its nearest medoid. The quality of the obtained medoids is measured by their average distance from all the model points within their cluster. CLARA repeats the sampling and clustering processes a pre-specified number of times and the final clustering result corresponds to the set of medoids with a minimum average dissimilarity (Kassambara 57).

2.2.4 Representative Portfolio

Once the clustering method has been executed the objects are grouped into a selected number of clusters represented by medoids. The medoids can be used to create a reduced dataset with similar behavior to the original one. However, since each cluster consists of a different number of model points, the medoids should be assigned different weights.

There are more possible methods for defining the weights. The simplest method is based on the number of model points within clusters. This method works quite well in the case of ordinarily distributed observations. However, it might underrate or overrate datasets containing extreme values. Therefore, it seems better to derive the weights from the tested variables and use for instance the sum of the values of one of the tested variables over all the model points within the cluster as weights (Fojtík et al. 6).

All calculations based on the original dataset can be performed on the reduced portfolio of weighted medoids and if the clustering method has been applied correctly, they should give similar results as the exact calculations performed on the whole dataset.

2.2.5 Precision Measure

Clustering approach is an approximation method and therefore, it does not give exact results. This is why it is necessary to test the precision of the method and decide whether the error is acceptable and the method can be used. In addition, the precision test can serve as a good criterion to choose between various clustering approaches.

In order to quantify the precision of clustering, one needs to define a precision measure. Such measure compares the results of the calculations performed on the whole dataset with the results calculated using the clustering approach. Since it is equivalent to test the precision and the error of the model, most precision measures are based on quantifying the model's error.

One of the most commonly used precision measures is the relative difference between the exact and the approximate results. The error of the k-th variable is defined as

$$err_k = \frac{approx_k}{real_k} \tag{7}$$

where $approx_k$ is the result obtained using the clustering approach and $real_k$ is the result of the exact calculation. To compare multiple clustering solutions, the error can be compared using the maximum, average and median value of an absolute value of error (Fojtík et al. 4).

2.3 Stratified Random Sampling

The simplest way of reducing the computational time of life insurance liabilities is reducing the size of the portfolio. In the case of cluster analysis, this is performed in a rather complex way using many iterations in order to find a group of model points which are the best representatives of the whole portfolio. A question arises whether the selection of representative model points could be performed in a more time-effective way. A very simple method is to select a random sample from the dataset and perform all the required calculations on this sample. However, since the portfolio offers some useful information about the model points it seems appropriate to use the information in order to select a better random sample.

Stratified random sampling is a sampling method that consists in a classification of model points into mutually exclusive groups called strata and choosing a simple random sample from each stratum. The main advantage of stratified random sampling over simple random sampling lies in the fact that stratified random sampling uses available information about the data and thus it leads to a smaller estimation error (Kim et al).

As the selections in different strata are made independently, the variance of the estimation for the whole dataset can be obtained adding the variances of estimations for individual strata. This means that only the within-stratum variances enter into the variance of the estimation for the whole dataset. For this reason, the stratification should be performed in such a way that the units within a stratum are as similar as possible. Then even if each stratum differs markedly from other strata, the stratified sample should behave in a similar way to the whole dataset (Thompson 141).

A portfolio of life insurance policies may be stratified based on known variables such as age or sex of the insured person or other properties of the policy so that policies with the same characteristics are grouped into the same strata. The process of life liabilities estimation using stratified random sampling could be divided into the flowing steps:

1. Obtain the data matrix
2. Divide the model points into strata
3. Select a random sample from each stratum
4. Create a representative portfolio
5. Test the precision

Some of the steps are further described in the following section.

2.3.1 Data Matrix and its Stratification

The first step of stratified random sampling is to construct a data matrix. In such matrix, rows stand for different observations and columns represent various variables. First, the data needs to contain a reference variable whose value is intended to be reproduced using the representative portfolio. In the case of a portfolio of life insurance policies, the reference variable can be the present value of future cash flows from each model point, the present value of future profit, some of the profit criteria etc.

Apart from the reference variable, the data matrix should contain a set of stratification variables which are used for grouping the observations. Typically, these variables describe some qualitative properties of the model points such as age, sex or type of insurance. In contrast to clustering, the variables are required to be categorical so that more observations can have the same value and thus can be classified in the same group.

This does not necessarily mean that continuous variables such as fund value cannot be used at all but they need to be discretized first and intervals or levels of continuous variables can be used as categorical variables. In addition, some discrete variables such as age should be joined into intervals as well in order to reduce the number of strata created.

Since the portfolio of life insurance policies usually contains detailed information about the client and the contract settings, not all this information can be used for grouping the model points because some groups would be too small and the effect of random sampling would be diminished. It is, therefore, important to find a way of measuring the discriminatory power of the variables and choose those with the strongest impact on the liabilities development.

There are various methods for measuring the discriminatory power of the variables. One can perform for example the analysis of variance (ANOVA) and compare the between-strata and within-strata variances of the reference variable. The comparison between variables can be then performed for instance on the basis of ANOVA p-value. The values of the variables with the best results can be used for categorizing the model points.

An alternative approach is to skip the first step and instead of starting with selecting the stratification variables, one can start directly with defining the stratification categories. As was mentioned in the previous section, the categories should be made in such a way that the model points in each stratum are as similar as possible while they can be quite dissimilar from the model points in other strata. This is exactly what cluster analysis can do so the categories may be created using some form of cluster analysis with the stratification reference variable used as the clustering variable. In contrast to the clustering approach described in the previous chapter, the clusters are only used for grouping the objects but rather than finding a medoid in each cluster, a random sample is drawn from each one.

At the end of this step, several categories are defined based on different values of the stratification variables and each model point can be classified into one of the strata. Assuming the portfolio of N model points has been divided into m strata of the size N_i, it must hold that $N_1 + N_2 + \cdots + N_m = N$.

2.3.2 Random Sample Selection and Creation of the Representative Portfolio

Once all the model points have been grouped, a simple random sample without replacement of the size n_i is drawn from each stratum so that $n_1 + n_2 + \cdots + n_m = n$.

The most important aspect of this part is to determine the size of each sample. The simplest option is to use *uniform allocation* and draw the same number of model points from each stratum. This approach is similar to cluster analysis, in which case each cluster is represented by a single model point. Since each sample represents a different number of model points, the sample needs to be assigned a certain weight which should reflect the proportion of model points the sample represents. The weight of the sample drawn from the i-th strata is defined as

$$w_i = \frac{n \cdot N_i}{n^* \cdot N} \tag{8}$$

where N_i is the size of the *i*-th stratum, N is the size of the whole portfolio and n^* is the size of the sample drawn from each stratum.

It seems a better option to use *proportional allocation.* In such approach, the proportion of the sample taken from each stratum is equal to the proportion of each stratum in the entire portfolio so that the size of the sample reflects the size of the strata and there is no need to weight the randomly selected model points (Aczel and Sounderpandian 743).

The size of the sample drawn from the *i*-th stratum is defined as:

$$n_i = n\frac{N_i}{N}$$

Another option is to use *Neyman allocation* whose purpose is to maximize the precision of the model given a fixed sample size. With Neyman allocation, the optimum size of the sample is

$$n_i = n\frac{N_i\sigma_i}{\sum_{j=1}^{m} N_j\sigma_j}$$

where σ_i is the standard deviation of stratum *i* (Aczel and Sounderpandian 743).

The randomly selected model points are used to create the representative portfolio. All calculations can be then performed on the representative portfolio and they should lead to similar results as the calculations performed on the whole portfolio. However, since the number of model points in the representative portfolio is lower than in the original one, the model points need to be weighted in some way. In the case of uniform allocation, the weights are assigned using equation (8). In the case of proportional and Neyman allocation, the weights are already reflected by the number of model points selected from each sample. However, when computing values of variables of summation such as BEL, it is necessary to multiply the result by the inverse value of the proportion of model points which have been drawn from the portfolio N/n.

2.3.3 Precision Test

One needs to bear in mind that the stratified random sampling method is an approximation method and it leads to a certain error. If all the stratum sizes are sufficiently large and contain at least 30 model points, an approximate 100(1 - α)% confidence interval for the mean of the reference variable $\bar{X}$ is

$$\bar{X}_{st} \pm Z_{\frac{\alpha}{2}}\sqrt{\widehat{var}(\bar{X}_{st})}$$

where $\bar{X}_{st}$ is the mean of the stratified sample, $Z_{\frac{\alpha}{2}}$ is the $\frac{\alpha}{2}$ quantile from the standard normal distribution and $\widehat{var}(\bar{X}_{st})$ is the estimator of the variance of the stratified sample mean defined as

$$\widehat{var}(\bar{X}_{st}) = \sum_{i=1}^{m}\left(\frac{N_i}{N}\right)^2\left(\frac{N_i - n_i}{N_i}\right)\frac{\sigma_i^2}{n_i}$$

Apart from measuring the confidence interval, one can measure the approximation error. This can be done by comparing the results obtained from the calculations performed on the representative portfolio of randomly selected model points with the results based on the whole portfolio. The error can be measured using equation (7).

3 Application

The third chapter of the thesis shows the application of the discussed methods to a real dataset of life insurance policies. It aims to test the overall applicability of the methods with respect to their accuracy and computational time. In addition, it examines various approaches and settings of each method and tries to identify the most efficient use of each method.

The whole case study is performed using R software. It should be mentioned that R may not be the most time efficient software and the computations could be performed faster using different language. However, the concern of the thesis is to compare computational times of different methods rather than show the exact computational time of each one. It seems, therefore, rather irrelevant what software is used since the relationships between the computational times are identical whichever software is used.

3.1 Input Data and Assumptions

The following chapter gives information about the input model points and assumptions used in all the tested methods. In addition, it introduces the modification of model points to a homogeneous portfolio.

3.1.1 Model points

The dataset used in the thesis comes from a real portfolio of life insurance policies. It comprises 106,524 model points each of which contains personal information about the insured person and the contract settings. However, in accordance with the privacy policy of the insurance company, some confidential information had to be omitted or modified.

All policies in the portfolio combine some kind of life insurance with investing. Unfortunately, there are no policies on other types of contracts such as survival benefit contracts which makes it impossible to test the methods on other forms of insurance. Although this slightly simplifies the process of life insurance liabilities estimation, it does not seem a big problem because the insurance companies nowadays offer primarily the products present in the tested portfolio. Also, all the tested methods can be easily applied to other forms of insurance as well and it is likely that they would lead to similar results since the calculations of liabilities are similar for all the types of insurance contracts. The portfolio of 106,524 model points could be described as a medium-sized portfolio for a smaller insurance company and it seems more than sufficient for the purpose of this thesis.

An example of four model points can be seen in Table 3.1. Apart from the policy number, each model point is distinguished by twelve variables. The first variable is the *Product Type.* Altogether, there are four product types which differ in the settings of charges, surrender fees and surrender period. The second and the third variables give personal information about the client which is used for prediction of deaths. *Premium frequency* shows how many times a year the premium is paid. *Duration* gives information on how many months have passed since the insurance policy came into effect, while *Policy term* is the period from policy inception to maturity. *Sum Assured* is the amount paid in the case of death and *Fund Value* is the current value invested in an investment fund. *Benefit* is the payment in the case of death. It can be of four types. The type SA only pays sum assured in the case of death, type SA_CV pays cash value at death in addition to sum assured. M_SA_CV pays either sum assured or cash value, whichever one is bigger. Finally, the type decreasing pays a benefit whose value equals a temporary annuity until the contract's maturity. This means that the benefit is decreasing and at maturity, it equals zero. *TIR* is the value of technical interest rate which is used for calculating the guaranteed income. The last variable is the current c*ash value of profit share*. It equals the accrued sum of the income guaranteed on technical interest rate and the part of the profit on the contract which the insurance company shares with the client.

Policy Number	**Product Type**	**Sex**	**Age at Entry**	**Annual Premium**	**Premium Frequency**	**Duration (Months)**	**Policy Term (Months)**
1	A	0	18	2488	12	232	264
2	B	1	33	4154	4	200	408
3	C	0	27	639	12	224	552
4	D	1	53	1170	1	177	216

Sum Assured	**Fund Value**	**Benefit**	**TIR**	**Cash Value Profit Share**
100,000	54,097.64	SA	0.015	0
200,000	56,647.04	SA_CV	0.02	0
30,000	0	M_SA_CV	0.025	0
30,000	12,645.78	Decreasing	0.025	0

Table 3.1: Sample Model Points
Source: Personal collection

3.1.2 Assumptions

All the methods are tested using the same assumptions about mortality, lapses, investment return and expenses. The assumptions are summarized in table Table 3.2. Mortalities assumption is based on the Czech life tables published by the Czech Statistical Office and it is adjusted using adjustment ratios shown in the table. Since only life insurance products are considered, it is not necessary to distinguish between life insurance and annuity adjustment ratios. The expenses development is estimated using the growth rate of 2.5%.

Year (n)	**Mortality Adjustment**	**Lapse Rates**	**Investment Return**	**Commissions**	**Initial Expenses**	**Renewal Expenses**
1	0.2	0.3	0.025	0.6	2,000	800
2	0.3	0.25	0.0255	0.2	2,050	820
3	0.4	0.2	0.026	0.04	2,101.3	840.5
4	0.5	0.15	0.0265	0.04	2,153.8	861.5
5+	0.6	0.1	0.027	0.04	2,000*1.25^n	800*1.25^n

Table 3.2: Assumptions
Source: Personal collection

It should be mentioned that the assumptions are rather simplified and they may not exactly reflect the current market situation. However, the aim of the thesis is not to calculate the exact value of liabilities of this particular portfolio but rather to find a method which could fast and accurately calculate liabilities of any life insurance portfolio given realistic assumptions. The task of finding the assumptions is left to life insurance companies.

3.1.3 Homogeneous portfolio

In order to make the methods testing more straightforward and to avoid incorrect conclusions caused by an unusual behavior of the portfolio, all the methods are first tested using a simplified dataset assuming a homogeneous portfolio of policies and only then, the methods are applied to the real portfolio. The homogeneous portfolio is created reducing the original thirteen variables to the policy number, age at entry, sex, annual premium, policy term, duration, sum assured and fund value. All the other variables are assumed to be fixed to the same value for all the model points. Table 3.3 shows a sample of the homogeneous portfolio.

Policy Number	**Age at Entry**	**Sex**	**Annual Premium**	**Policy Term (Months)**	**Duration (Months)**	**Sum Assured**	**Fund Value**
1	18	0	2,488	264	232	100,000	54,097.64
2	33	1	4,154	408	200	200,000	56,647.04
3	27	0	639	552	224	30,000	0
4	53	1	1,170	216	177	30,000	12,645.78

Table 3.3: Homogeneous Portfolio

Source: Personal collection

The rest of the variables are fixed as follows: Firstly, all policies are assumed to be of the same product type (type A) with the same settings of charges, surrender fees and surrender period. Secondly, the premium is always paid annually. Thirdly, only sum assured is paid as the benefit in the case of death. Finally, TIR is set to 2.5% and the cash value is assumed to be zero.

3.2 Cash Flow Analysis

Before applying any approximation methods, liabilities are computed using the traditional approach giving exact results though at a very high computational time.

Applying cash flow analysis is quite straightforward. It consists in applying equation (1) to every model point and calculate the expected cash flow for every year within the policy term. As there are 106,524 model points with the policy term reaching up to 25 years the formula needs to be applied 2,002,228 times resulting in 2,002,228 values of cash flows. These values can be further used for obtaining the expected present value of future cash flows or best estimate liability from all policies of the insurance company using equation (2).

The process is then repeated several times under different assumptions in order to calculate the sensitivities of best estimate liability to the changes in the assumptions. In this study, five scenarios are tested increasing one by one expected mortalities, lapses, investment returns, commissions and expenses by 1%. This increases the number of times equation (1) is applied to 12,013,368. Comparing each scenario result with the best estimate, one can learn whether the rise in each scenario influences the liabilities positively or negatively and how strong the influence is.

Finally, LAT reserve is calculated assuming that all the assumptions changed in an unfavorable direction. Insurance companies need to study the likely negative development more closely and change the assumptions by a realistic value. However, this thesis simplifies the matter and assumes a change of all assumptions by 10% in the unfavorable direction.

3.2.1 Homogeneous portfolio

Cash flow analysis is first performed on a homogeneous portfolio. This does not particularly simplify the computation because the calculations are performed on each model point separately and so it does not really matter how many different formulas are used. However, the homogeneous portfolio will be very helpful later when approximation methods are tested and compared with cash flow analysis.

The value of BEL in the case of a homogeneous portfolio is approximately 253 million CZK and its calculation takes approximately 14 minutes.

Table 3.4 summarizes the changes in BEL value which follow a 1% change in each assumption respectively. One can see that BEL function is increasing in all the assumptions except lapses. The increase in lapses causes a decrease in BEL which shows that the surrender fee is so high that the insurance company profits from the lapses. BEL is most sensitive to the changes in expenses and it increases 3.5 times faster than the expenses. Surprisingly, mortality, which many people believe to be the most important factor of life insurance, plays the least important part in BEL development and it increases five times faster than BEL.

Scenario	Best Estimate Liability
Mortality increased by 1%	Increased by 0.2%
Lapses increased by 1%	Decreased by 1.3%
Investment return increased by 1%	Increased by 1.9%
Commissions increased by 1%	Increased by 1.3%
Expenses increased by 1%	Increased by 3.5%

Table 3.4 BEL Sensitivities to Assumptions Changes (homogeneous portfolio)
Source: Personal collection

LAT reserve is calculated increasing all the assumptions except for lapses by 10% and decreasing lapses by 10%. Under such assumptions, the present value of future cash flows is approximately -489 million CZK which means that the LAT reserve is 489 million CZK. The whole computational process took 2 hours and 5 minutes.

3.2.2 Heterogeneous portfolio

The same calculations are applied to the real portfolio of heterogeneous model points. This time, however, the computations for each model point differ more because they need to reflect the different contract settings.

The value of BEL of a heterogeneous portfolio is approximately 563 million CZK and its calculation takes 16 minutes.

Table 3.5 summarizes the changes in the BEL value which follow a 1% change in each assumption respectively. One can notice that the heterogeneous portfolio is much less sensitive to the changes in all the assumption. This demonstrates how portfolio diversification can reduce the risk of a change in the liability value.

Similarly to the case with homogeneous portfolio, BEL is increasing in all the assumptions except lapses. It reacts most strongly to the changes in lapses and expenses which are the only two assumptions that change at a slower rate than BEL. All other assumptions changes cause a comparatively lower percentage change in BEL.

Scenario	**Best Estimate Liability**
Mortality increased by 1%	Increased by 0.1%
Lapses increased by 1%	Decreased by 1.6%
Investment return increased by 1%	Increased by 0.9%
Commissions increased by 1%	Increased by 0.6%
Expenses increased by 1%	Increased by 1.5%

Table 3.5 BEL Sensitivities to Assumptions Changes (heterogeneous portfolio)
Source: Personal collection

The LAT reserve is again calculated decreasing lapses by 10% and increasing all other assumptions by 10%. This gives the LAT reserve of approximately 865 million CZK. The whole computational process is slightly slower than the one with the homogeneous portfolio and it takes about 2 hours and 11 minutes.

3.3 Cluster Analysis

The application of cluster analysis follows the five steps described in chapter 3.2

1. Obtain and standardize the data matrix
2. Select a distance measure
3. Execute the clustering method
4. Create the representative portfolio

5. Test the precision

Those steps are first applied to a homogeneous portfolio in order to optimize the clustering algorithm and only then, the method is applied to a heterogeneous portfolio.

Since CLARA algorithm utilizes random sampling, the results are random as well and unless an extremely high number of iterations are used, every time a different selection of medoids is made. For this reason, the results may differ every time. In order to get an unbiased view of the method's properties, each trial is performed ten times and the results presented are calculated as the average value.

3.3.1 Homogeneous Portfolio Clustering

Optimizing the clustering algorithm using homogeneous portfolio has many advantages. Chapter 3.2. showed that the computation is slightly faster. In addition, the model points are more similar and therefore, fewer samples need to be selected from the portfolio to achieve sufficient accuracy. This simplifies and speeds up the computational process and enables one to test more clustering settings in order to optimize the clustering algorithm.

Data Matrix

The first step consists in selecting clustering variables and obtaining and standardizing the data matrix. It seems an obvious choice to use the present value of future cash flows from each model point as the clustering variable. The present value of future cash flows from model point *i* is defined as

$$PVCF_j = \sum_{t=1}^{n} CF_{j,t} \cdot \prod_{k=1}^{t} \frac{1}{1+i_k} \tag{9}$$

where $CF_{j,t}$ denotes cash flow from the j-th model point in year t, *n* is the number of policy years to maturity and i_k is the expected investment return at time k.

The main advantage of using PVCF as the clustering variable is that it reflects the dynamics and development of each contract and that all BEL, LAT and sensitivities are directly derived from PVCF. The main weakness of this variable lies in the fact that the values cannot be obtained directly from the dataset but they need to be calculated first, which increases the computational time. Other clustering variables are tested at the end of this chapter.

Model Point Nr.	PVCF
1	-8,633
2	-1,963
3	-4,318
4	-8,667
5	-7,601
…	…
106,524	-4,442

Table 3.6: Data Matrix
Source: Personal collection

Table 3.6 shows a sample of the data matrix derived from the homogeneous portfolio of model points. The data matrix is standardized using equation (3). A sample of the standardized matrix can be seen in Table 3.7.

Model Point Nr.	PVCF
1	-1.15
2	-0.08
3	-0.36
4	-1.16
5	-0.96
…	…
106,524	-0,38

Table 3.7: Standardized Data Matrix
Source: Personal collection

Distance Measure

Two distance measures are considered in this thesis - Manhattan and Euclidean. To decide which one works better, two clustering trials are performed using Manhattan and Euclidean distance measures respectively. Both trials are performed using a modified version of R package *Cluster* with 50 samples.

The most important criteria of the efficiency of the distance measures are the computational time and approximation accuracy of the cluster analysis performed on their basis. Since both computational time and accuracy might change with the increasing number of clusters, several trials are performed with a different number of clusters. The accuracy is measured based on BEL and LAT using the error measure defined as

$$err_k = \frac{approx_k}{real_k}$$

where $approx_k$ is the value of BEL or LAT obtained using the clustering method and $real_k$ is the value of BEL or LAT obtained using the accurate calculation described in chapter 3.2.

Altogether, 56 trials have been made with the number of clusters ranging from 1 to 500. A sample of a relationship between the number of clusters, computational time and error for either distance measure can be seen in Table 3.8. The computational time includes the time of the selection of medoids but it does not include the time of calculating the present values of future cash flows. Note that the results are only illustrative since some clustering properties such as the number of samples have not been optimized yet.

Number of Clusters	BEL Error		LAT Error		Computational Time	
	Euclidean	Manhattan	Euclidean	Manhattan	Euclidean	Manhattan
1	39.84%	42.68%	24.00%	18.16%	0.48	0.46
2	17.18%	19.63%	12.80%	6.28%	0.49	0.51
5	8.05%	6.54%	3.71%	4.06%	0.55	0.59
10	1.91%	2.84%	1.59%	2.14%	0.86	0.77
50	0.94%	0.85%	0.86%	0.71%	2.22	2.24
100	0.42%	0.66%	0.53%	0.24%	3.72	3.92
150	0.31%	0.45%	0.45%	0.23%	5.39	5.6
300	0.11%	0.02%	0.24%	0.11%	10.91	10.45
500	0.10%	0.02%	0.21%	0.05%	17.84	17.3

Table 3.8: Distance Measures Efficiency
Source: Personal collection

Generally, computational time increases and error decreases with the increasing number of clusters. The relationship between computational time and error when calculating the LAT reserve can be observed in Figure 3.1 and Figure 3.2.

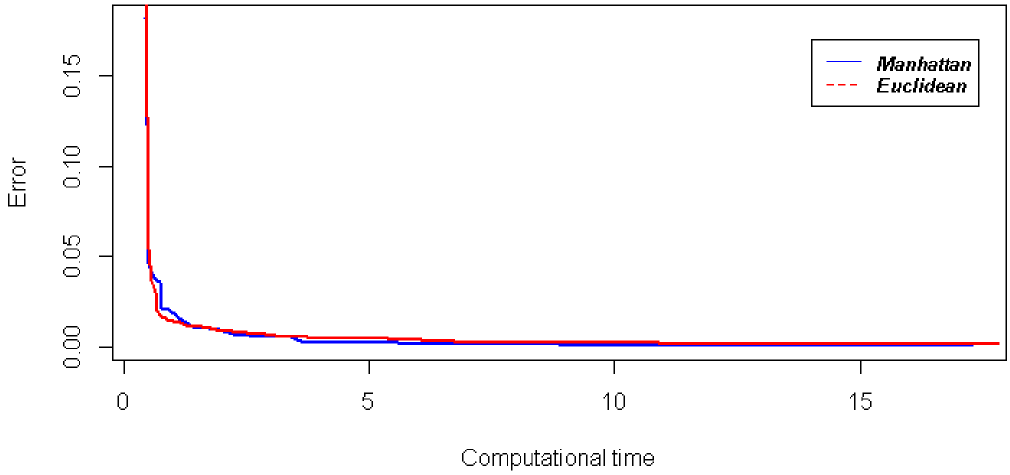

Figure 3.1: Distance Measures Efficiency
Source: Personal collection

Figure 3.2 shows a detail of the relationship based on the number of clusters ranging from 20 to 150. This interval of the number of clusters seems to offer the best ratio of accuracy to computational time and it is therefore likely that the final choice of the number of clusters will be in this interval. For this reason, the choice of distance measure is primarily made with respect to the efficiency of the measures when 20 to 150 clusters are used.

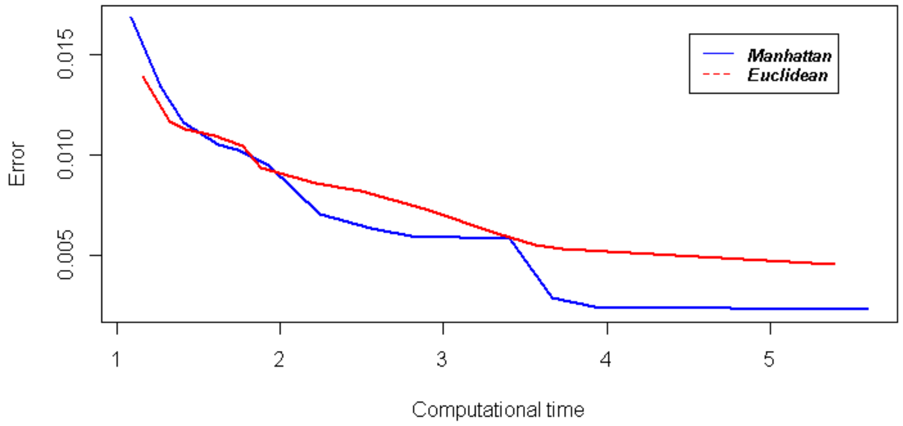

Figure 3.2: Distance Measure Efficiency Detail
Source: Personal collection

The figures and the table show that both measures give comparable results. However, if the number of clusters ranges from 20 to 150, Manhattan measure gives results with a lower error at the same computational time. For this reason, Manhattan measure is chosen as the distance measure in this case study.

Clustering Method

The method used in this case study is CLARA algorithm. The main arguments of the algorithm are the data matrix, the distance measure, the number of clusters, the number of samples and the size of the samples. The setting of these arguments may influence both the computational time and the accuracy of the algorithm. Therefore, possible combinations of these arguments are tested in order to find an optimal setting.

First, optimal combinations of the number of samples and the size of the samples are searched holding the number of clusters fixed to 50. In order to find an optimal combination, many trials with a different combination of the number of samples and the size of the samples are made measuring their accuracy and computational time. The computational times of some of the combinations in minutes can be seen in Table 3.9 while the computational errors can be seen in Table 3.10. It is not surprising that computational time generally increases with the increasing number of samples as well as with the increasing size of samples whereas the approximation error decreases when either of the arguments increases. This means that one cannot generally increase the accuracy of the calculation without increasing the computational time as well.

Time		Number of Samples								
		50	60	80	100	300	400	500	750	1000
Size of Samples	50	1.85	1.89	1.90	1.93	1.96	1.97	1.97	1.98	2.00
	60	1.91	1.94	1.96	1.97	1.98	2.00	2.00	2.01	2.01
	80	1.98	1.99	2.00	2.03	2.03	2.03	2.04	2.04	2.06
	100	2.01	2.03	2.03	2.05	2.07	2.08	2.08	2.08	2.09
	300	2.05	2.06	2.06	2.10	2.11	2.11	2.12	2.18	2.27
	400	2.05	2.07	2.11	2.13	2.16	2.16	2.17	2.20	2.32
	500	2.07	2.08	2.11	2.13	2.20	2.22	2.26	2.29	2.38
	750	2.12	2.21	2.25	2.25	2.27	2.32	2.34	2.42	2.43
	1000	2.28	2.33	2.38	2.39	2.53	2.61	2.61	2.64	2.65

Table 3.9: Computational Time Based on the Number of Samples and the Samples Size
Source: Personal collection

Error		Number of Samples								
		50	60	80	100	300	400	500	750	1000
Size of samples	50	3.26%	3.18%	2.92%	2.37%	2.21%	2.11%	1.74%	1.47%	1.41%
	60	2.88%	2.64%	2.35%	1.67%	1.64%	1.63%	1.34%	1.32%	1.08%
	80	1.97%	1.92%	1.33%	1.28%	1.23%	1.22%	1.07%	1.05%	0.61%
	100	1.88%	1.29%	1.15%	0.97%	0.83%	0.77%	0.65%	0.61%	0.51%
	300	1.14%	0.98%	0.79%	0.72%	0.61%	0.49%	0.45%	0.44%	0.39%
	400	0.93%	0.92%	0.76%	0.61%	0.57%	0.43%	0.39%	0.38%	0.29%
	500	0.80%	0.77%	0.69%	0.51%	0.42%	0.39%	0.23%	0.23%	0.15%
	750	0.67%	0.63%	0.39%	0.30%	0.26%	0.17%	0.13%	0.11%	0.05%
	1000	0.60%	0.55%	0.19%	0.14%	0.10%	0.09%	0.08%	0.04%	0.00%

Table 3.10: Error Based on the Number of Samples and the Samples Size
Source: Personal collection

From the results of the tested combinations, one can derive the achievable accuracy given a computational time or the efficiency of each combination. The relationship between computational times and errors of the tested combinations of are shown in Figure 3.3. The function is convex and decreasing. The best combinations are such that give the lowest error given computational time or the lowest computational time given error so that one cannot increase the accuracy of the calculation without increasing the computational time as well. These optimal combinations lie on the lower left border and they are marked blue.

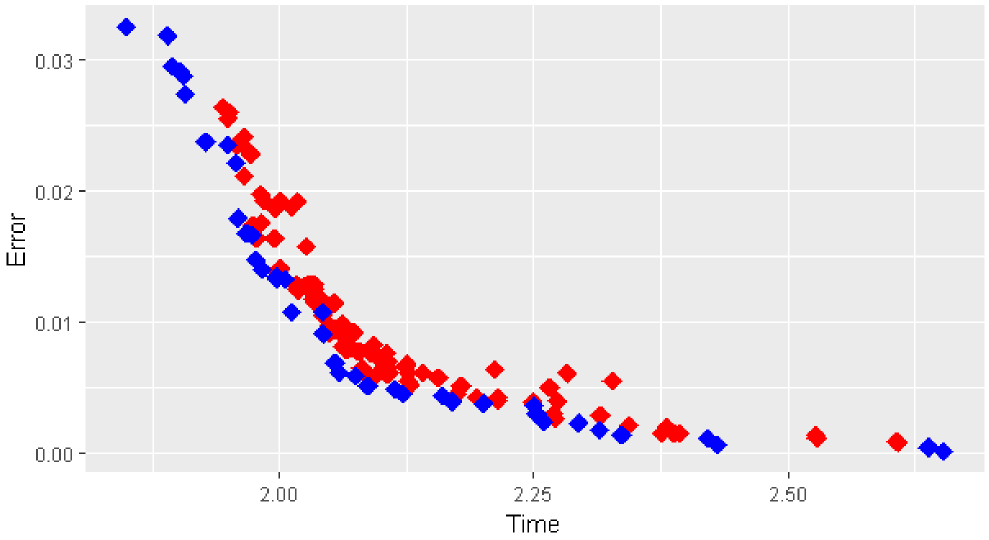

Figure 3.3: Efficiencies of Number of Samples and Samples Size Combinations
Source: Personal collection

Efficient		Number of Samples								
		50	60	80	100	300	400	500	750	1000
Size of Samples	50	▩	🟦	🟦	🟦	🟦			🟦	
	60	🟦	▨		🟦			🟦		🟦
	80			▩				🟦		🟦
	100				▨					🟦
	300					▨	🟦	🟦		
	400						▩	🟦	🟦	
	500							▩	🟦	
	750				🟦		🟦	🟦	▩	🟦
	1000								🟦	▩

Table 3.11: Efficient Combinations of the Number of Samples and the Samples Size
Source: Personal collection

Some of the efficient combinations can be seen in Table 3.11 marked blue. Unfortunately, there is no universal rule which would help one decide which combinations of the number of samples and samples size are optimal. It seems, however, that most of the efficient combinations lie on the diagonal or to the right from the diagonal. This means that the best solutions are obtained using either the same number of samples as the size of the samples or using a slightly higher number of samples than the samples size. Although the diagonal rule does not hold for all the trials, it will be used in this case study and the number of samples and the samples size will be held equal.

The next task is to find optimal combinations of the number of clusters and the number of samples. The combinations can be found using a similar approach to the one described above which was used to find optimal combinations of the number of samples and the size of the samples. However, the third argument, the size of the samples, is not fixed to a certain value as it was the case with the number of clusters but its value depends on the value of the number of samples. Thanks to this, all the three arguments can be tested at the same time and an optimal combination of all of them can be found.

However, there are certain restrictions regarding the values of the arguments. For instance, the samples size should be higher than the number of clusters. Therefore, in the case of the combinations with the number of clusters higher than the number of samples, samples size cannot equal the number of samples but it needs to be higher.

Time		Number of Samples							
		2	5	10	50	100	200	300	500
Number of Clusters	2	0.47	0.47	0.47	0.50	0.52	0.55	0.60	0.61
	5	0.52	0.55	0.57	0.60	0.64	0.64	0.68	0.68
	10	0.76	0.77	0.77	0.81	0.87	0.87	0.94	0.98
	50	2.04	2.10	2.10	2.15	2.28	2.30	2.44	2.54
	100	3.76	3.78	3.79	3.84	3.99	3.99	4.03	4.11
	200	7.03	7.08	7.11	7.21	7.39	7.70	13.85	14.18
	300	11.03	11.08	11.20	11.31	11.74	11.91	11.91	12.53
	500	18.55	18.62	18.88	20.60	44.66	46.63	50.76	57.53

Table 3.12: Computational Time Based on the Number of Samples and Clusters
Source: Personal collection

Error		Number of Samples							
		2	5	10	50	100	200	300	500
Number of Clusters	2	29.119%	17.894%	16.560%	14.278%	7.938%	6.697%	6.340%	5.913%
	5	11.170%	8.095%	6.611%	3.818%	3.219%	2.836%	2.692%	2.558%
	10	6.335%	4.818%	3.899%	3.021%	2.171%	2.056%	1.915%	1.734%
	50	2.286%	2.223%	1.612%	1.119%	0.872%	0.788%	0.632%	0.542%
	100	1.170%	0.588%	0.548%	0.356%	0.274%	0.235%	0.214%	0.124%
	200	0.585%	0.514%	0.429%	0.250%	0.163%	0.108%	0.108%	0.107%
	300	0.231%	0.225%	0.213%	0.180%	0.067%	0.054%	0.049%	0.039%
	500	0.170%	0.154%	0.135%	0.059%	0.014%	0.010%	0.008%	0.008%

Table 3.13: Error Based on the Number of Samples and the Number of Clusters
Source: Personal collection

Table 3.12 and Table 3.13 show the calculation times in minutes and errors of some of the tested combinations. Similarly to the test of the combinations of the number of samples and the samples size, the computational time is increasing and error is decreasing with both increasing number of clusters and number of samples.

The relationship between the computational time and error can be seen in Figure 3.4. The function is convex and decreasing and it seems that the convexity is higher than in the previous tested case. Optimal combinations are marked blue.

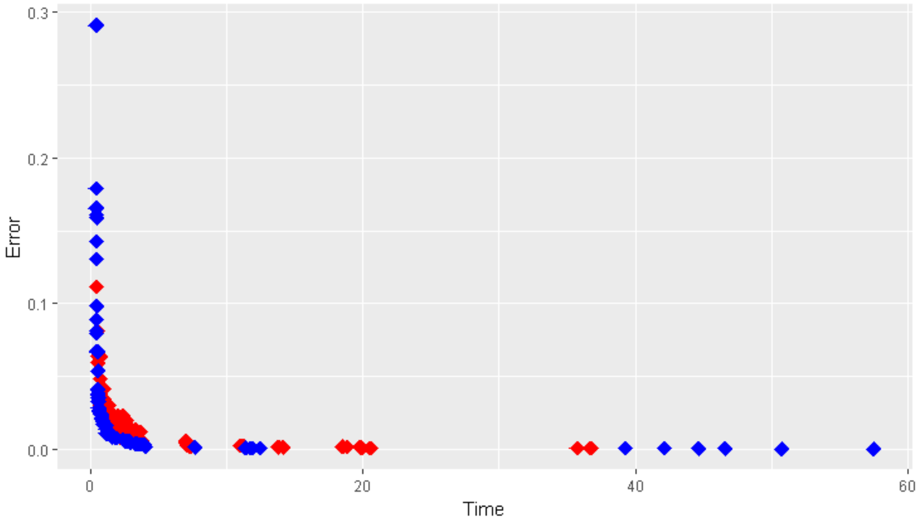

Figure 3.4: Efficiencies of the Number of Samples and Clusters Combinations
Source: Personal collection

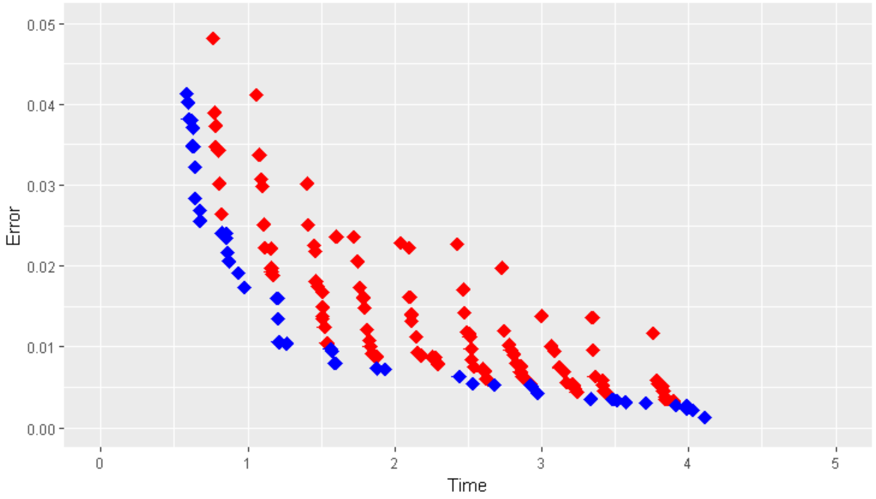

Figure 3.5: Efficiencies of the Number of Samples and Clusters Combinations (Detail)
Source: Personal collection

Since most of the times are lower than 5 minutes and most of the errors are up to 5% a greater detail of the values is shown in Figure 3.5 where one can see the efficient combinations.

Time		Number of Samples							
		2	5	10	50	100	200	300	500
Number of Clusters	2	🟦	🟦	🟦	🟦	🟦	🟦		
	5			🟦	🟦	🟦	🟦	🟦	🟦
	10					🟦	🟦	🟦	🟦
	50							🟦	🟦
	100					🟦	🟦	🟦	🟦
	200					🟦	🟦	🟦	🟦
	300						🟦		
	500					🟦	🟦	🟦	

Table 3.14: Efficient Combinations of the Number of Samples and Clusters
Source: Personal collection

The efficient combinations are marked blue in Table 3.14. All these combinations give the lowest error at the lowest possible computational time. Unfortunately, no generally applicable rule can be derived for the best combinations. Most efficient combinations lie in the upper triangle, though. This suggests that the number of samples should be at least as high as the number of clusters. For this reason, the efficient combinations lying in the lower triangle are ignored and only those in the upper triangle are considered for further modeling.

To sum up, the clustering method used in this case study is CLARA algorithm using Manhattan distance measure where the number of samples is at least as high as the number of clusters and the size of the samples is the same as the number of samples. From the arguments combinations fulfilling the stated rules, only the efficient ones are considered.

Representative Portfolio

Once the clustering method has been executed the model points are grouped into clusters and the original portfolio is represented by a selected number of medoids. Since each medoid represents a different number of model points, the medoids are assigned different weights.

Until now, all trials have been made using the number of model points within clusters as weights. An alternative approach is using the sum of the values of one of the tested variables over all the model points within the cluster as weights. The obvious choice seems to be using the negative sum of the present value of future cash flows from each model point within a cluster as weights. What could be considered a drawback of such approach is that some model points might yield positive expected future cash flow and thus some weights might be negative. In the tested portfolio, this happens with 21,362 model points out of the 106,524 that is with about every fifth model point. For some insurance companies, this might represent a problem because the weights could be interpreted as the number of policies represented by the medoid and there is no such a thing as a negative number of policies. Therefore, some information systems require the weights to be positive. However, mathematically, there is no problem with the weights being negative so in this study, the weights are allowed to be negative.

The accuracy of both approaches is tested with 28 trials each of which uses clustering algorithm with different setting of the number of clusters, the number of samples and samples size fulfilling the rules stated in the previous text.

In each trial, two reduced portfolios are created, the former using the number of model points within clusters as weights and the latter using the present value of future cash flows as weights. A sample reduced portfolio created using 5 clusters and 70 samples of the size 70 can be seen in Table 3.15.

Policy number	**Age at Entry**	**Sex**	**Annual Premium**	**Policy Term (Months)**	**Duration (Months)**	**Sum Assured**	**Weights (model points)**	**Weights (PVCF)**
29,271	39	0	1,020	372	4,464	50,000	22,968	23,537.3
75,161	28	1	3,264	504	6,048	200,000	29,035	28,097.3
20,874	32	1	2,160	456	5,472	100,000	37,894	37,401.8
28,102	38	0	10,440	384	4,608	500,000	3,518	5,310.2
37,659	48	0	6,168	264	3,168	100,000	13,109	12,912.3

Table 3.15: Reduced Portfolio
Source: Personal collection

The accuracy of the approaches is tested comparing the results of the reduced portfolio with the results of the whole portfolio. It is obvious that using the number of model points within clusters as weights leads to a certain approximation already when computing best estimate liability whereas using PVCFs as weights leads to zero error. On the other hand, the first approach can be less sensitive to changes in the assumptions and it can give a more reliable picture of the LAT reserve than the second approach.

The discrepancy between the exact value of BEL and LAT and the approximation obtained from both the approaches can be seen in Figure 3.6. One can notice that especially with a lower number of clusters, the PVCF weighed reduced portfolio may be highly sensitive to changes in the assumptions and it does not lead to good results when calculating the LAT reserve. The second approach seems to be a better solution even though it leads to certain inaccuracies when estimating BEL. After all, the purpose of the model is not to calculate

BEL since it needs to be computed before the method can be applied anyway. Therefore, the choice of weights should be based on LAT estimation accuracy rather than BEL estimation accuracy. For this reason, the case study will use the number of model points within clusters as weights.

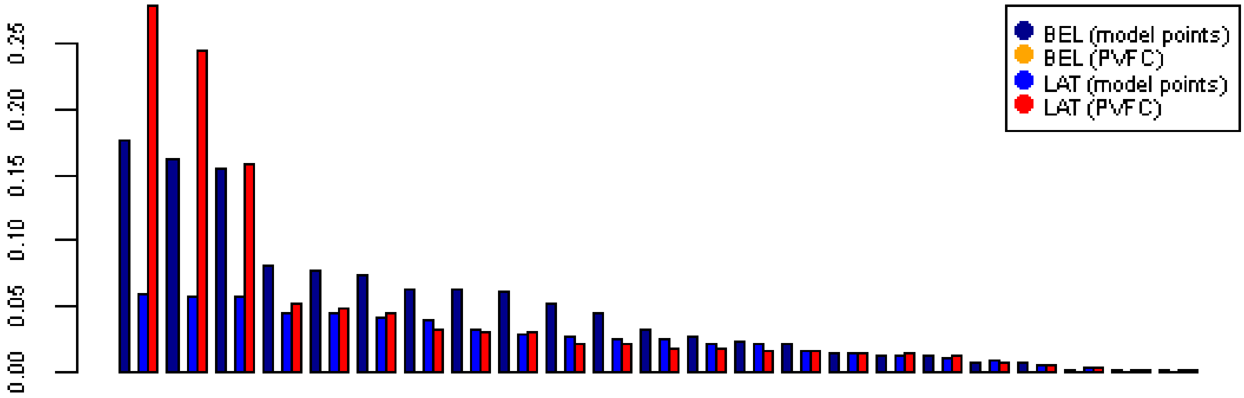

Figure 3.6: Approximation Error Based on Weights
Source: Personal collection

Testing the Model

The previous parts of the chapter have shown that the best results can be achieved using Manhattan distance measure and applying CLARA algorithm where the number of samples is at least as high as the number of clusters and the size of the samples is the same as the number of samples. From these combinations of the number of clusters, the number of samples and the samples size, some have been omitted because their accuracy can be achieved at a lower computational time using different combinations. Altogether, there are 56 combinations considered in this case study.

Applying the clustering algorithm, the original portfolio can be reduced to the portfolio of medoids. As the previous text has shown, the best way to weight these medoids is to use the number of model points which the medoid represents as weights.

The precision of the approximation is measured comparing the approximate LAT reserve with its exact value. Testing the precision of the algorithm based on each of the 56 combinations of arguments and measuring the computational time of each, one can derive a relationship between the precision and computational time of the clustering method. The relationship is shown in Figure 3.7.

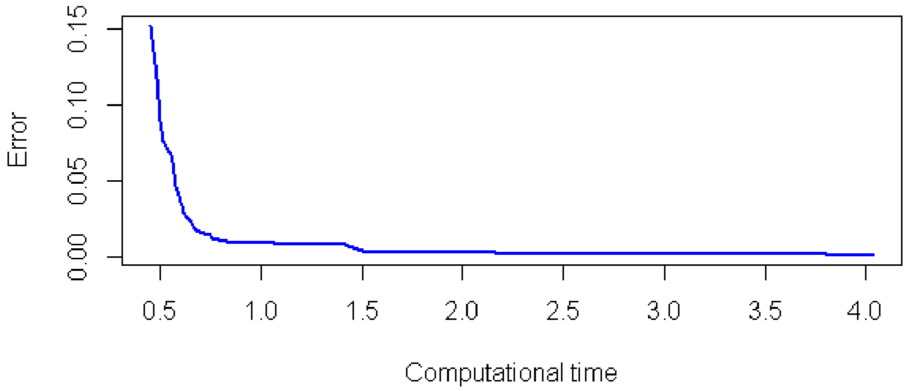

Figure 3.7: Clustering Method Computational Time and Approximation Error
Source: Personal collection

One can see that the approximation error decreases very fast and in less than 4 minutes, one can achieve almost zero error. Based on the preferences of the insurance company, one of the arguments settings can be chosen for calculations.

To illustrate, it is assumed that the insurance company requires the accuracy of at least 99.5%. Under such assumptions, they would use the clustering method with 30 clusters and 30 samples of 30 model points. Such algorithm calculates the LAT reserve with the accuracy of 99.6% and the computational time of about 1 minute and 30 seconds. If they contented themselves with the accuracy of 99%, they could calculate the LAT reserve in less than 50 seconds. Recall that the accurate computation using cash flow analysis lasted more than 2 hours.

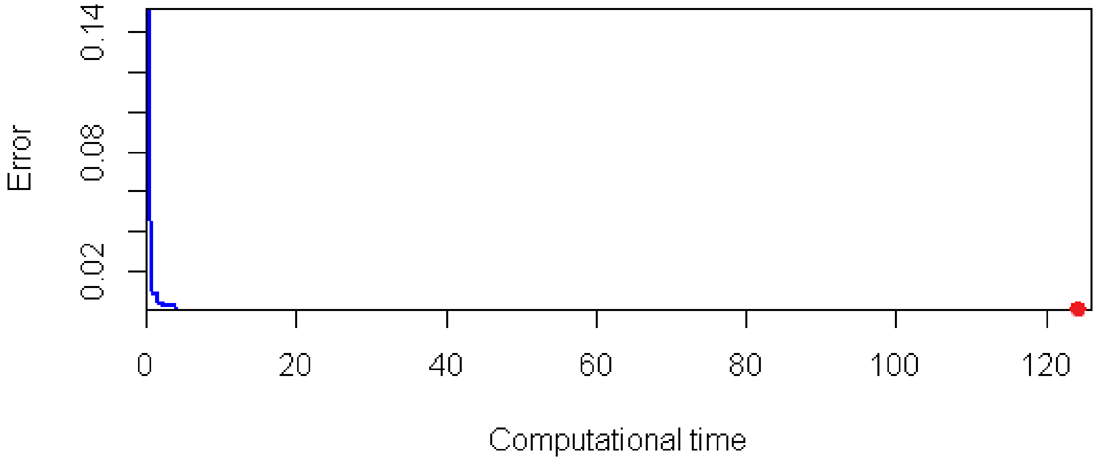

Figure 3.8: Clustering Method and Cash Flow Analysis Comparison
Source: Personal collection

To see the comparison of clustering method and cash flow analysis, the solution of cash flow analysis is added to the plot in Figure 3.8. One can see that the accurate solution is extremely time-consuming. In reality, the difference might be even more striking because insurance companies need to test much more scenarios under different assumptions while in this case study, only five different scenarios are tested before estimating the LAT reserve.

Different Clustering Variables

All the models that have been tested so far used the present value of future cash flows from each model point as the clustering variable. A question arises whether this variable gives the best results. In order to answer this question, different clustering variables and their combinations are tested.

First, a combination of four basic characteristics of each model point is selected as clustering variables: age, annual premium, policy term and sum insured. The advantage of this approach is that the parameters can be derived directly from the dataset and no accurate cash flows need to be calculated since they can be approximated using the clustering method. On the other hand, the impact of such variables is rather ambiguous. To illustrate, the distance between ages 60 and 65 is the same as the distance between ages 30 and 35 but it is obvious that the formal difference would impact the liability much stronger than the latter.

Second, the annual cash flows from each contract in the next three years are used as clustering variables. The advantage of such approach is that it is not necessary to calculate all the expected values of future cash flows and that using three clustering variables instead of one might lead to more accurate results.

Finally, the second approach is combined with the original approach and annual cash flows within the next three years together with the present value of all future cash flows from each contract are used as clustering variables.

The comparison of the three approaches together with the original approach can be seen in Figure 3.9. Not surprisingly, using policy parameters as clustering variables leads to the least accurate results and the highest accuracy which can be achieved under current settings is 99.5%. The second and the third approach lead to better results but neither is able to achieve as high accuracy as the original approach. It is possible, that those two approaches would be more convenient when computing cash flows within a limited time period. However, the original approach using the present value of future cash flows only as the clustering variable seems to be the best one for the calculation of the LAT reserve.

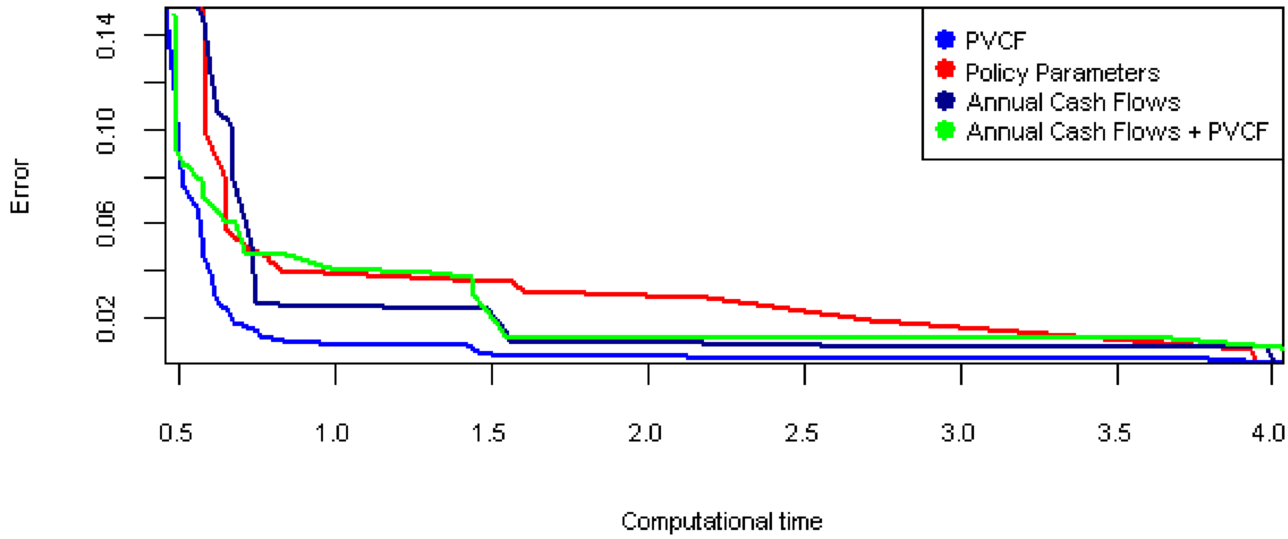

Figure 3.9: Comparison of Clustering Variables
Source: Personal collection

To sum up, the best clustering algorithm for estimating the LAT reserve of a homogeneous portfolio is CLARA algorithm based on Manhattan distance measure. The clustering variable is the present value of future cash flows from each model point. The number of samples is at least as high as the number of clusters and the size of samples is the same as the number of samples. From these combinations of arguments, only 56 optimal combinations have been selected.

Applying the algorithm, one can create a reduced portfolio of weighted medoids. The best way to weight the medoids is to use the number of model points within the cluster which the medoid represents as weights. The reduced portfolio can be then used for the calculation of BEL under various assumptions and for the calculation of the LAT reserve.

3.3.2 Heterogeneous Portfolio Clustering

In this subchapter, the clustering algorithm calibrated in the previous section is applied to the real portfolio of life insurance policies. Figure 3.10 shows the relationship between the computational time and the approximation error of clustering of the heterogeneous portfolio. Similarly to the homogeneous portfolio clustering, the approximation error is exponentially decreasing.

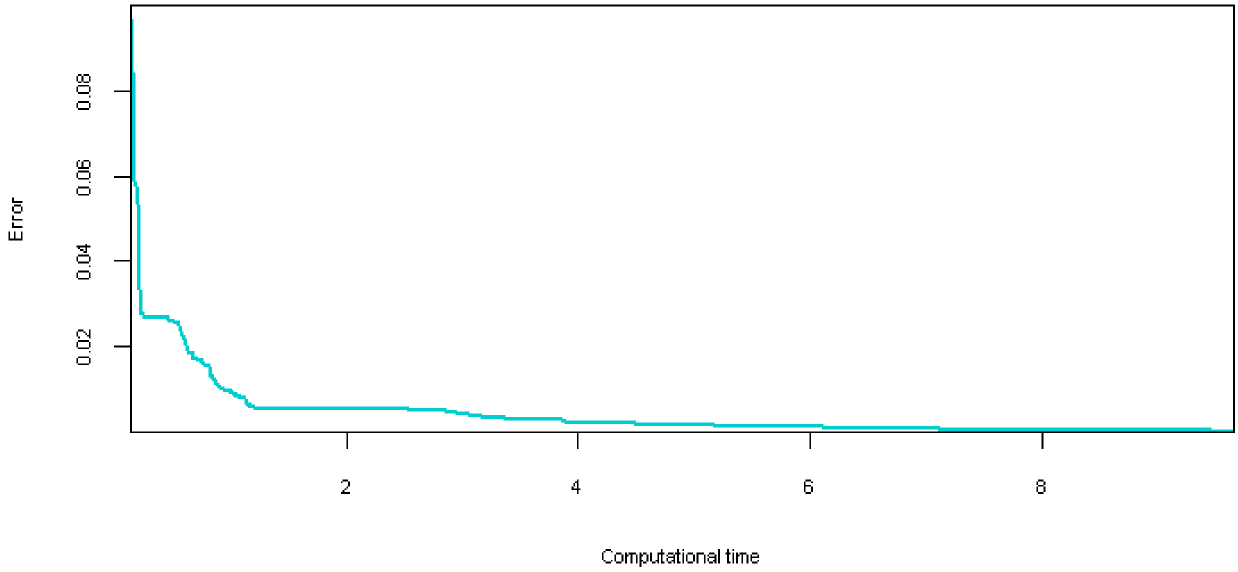

Figure 3.10: Heterogeneous Portfolio Clustering
Source: Personal collection

Assuming the insurance company wants to calculate the LAT reserve with the accuracy of at least 99.5% they would use 200 samples and group the portfolio into 10 clusters. Such combination of argument computes the LAT reserve in 3 minute and 15 seconds with the accuracy of 99.55%. The highest achievable accuracy under the given settings of arguments is 99.99% with the computational time of 9 minutes and a half. Such high accuracy should be more than sufficient for insurance companies and the time saving is significant.

To compare the results of the heterogeneous portfolio with the ones of the homogeneous portfolio, both relationships are put in one plot in Figure 3.11. The heterogeneous portfolio curve is more flat and it enables one to achieve more accurate results at a minimum computational time. The computational time of 38 seconds is the critical point at which it is possible to achieve the accuracy of about 98% with both heterogeneous and homogenous portfolio clustering. After this point, the clustering of homogeneous portfolio is more accurate thought the differences are not significant.

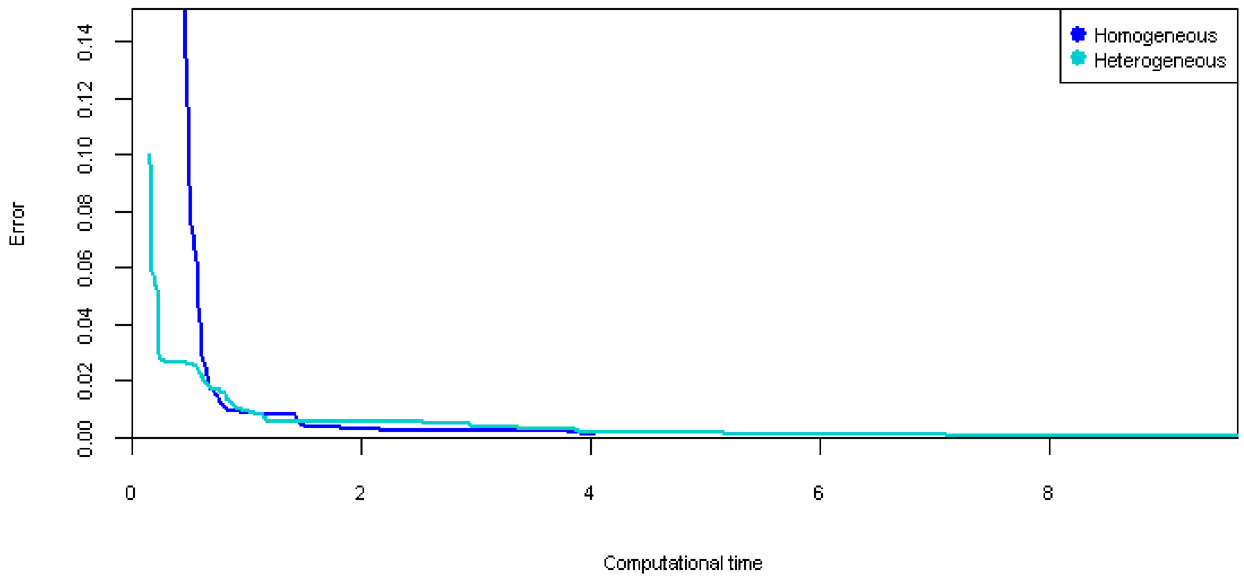

Figure 3.11: Heterogeneous Portfolio versus Homogeneous Portfolio Clustering
Source: Personal collection

Apart from increasing the time efficiency, the algorithm also enables a better insight into the portfolio of policies and the insurance company can see a selected number of representative model points and identify the variables with a distinctive power. An example with five medoids can be seen in Table 3.16. It can be noted that benefit seems to play an important role in BEL development as all four benefit forms are represented in the representative portfolio while quite surprisingly, product type C is not represented at all so it seems that the product type might not play a very important role.

Policy number	Product type	Sex	Age at Entry	Annual Premium	Premium Frequency	Duration (Months)	Sum Assured	Benefit	Weight
653	B	0	18	1,488	12	250	100,000	SA	31,771
39,884	B	1	18	4,800	12	206	300,000	SA_CV	10,496
59,143	D	0	38	612	2	247	30,000	SA_CV	28,310
74,320	A	0	27	3,3648	12	257	200,000	M_SA_CV	16,549
91,251	D	0	23	2,928	12	190	150,000	Decreasing	19,398

Table 3.16: Medoids
Source: Personal collection

It could be concluded that the clustering algorithm works well with both homogeneous and heterogeneous portfolio and it enables an enormous time reduction at a very high accuracy.

3.4 Stratified Random Sampling

The application of stratified random sampling follows the five steps mentioned in chapter 2.3. As it was the case with cash flow analysis and cluster analysis, stratified random sampling is first applied to the homogeneous portfolio so that the stratification variables and other settings of the model can be optimized more easily and only then, the method is applied to the real portfolio.

3.4.1 Homogeneous Portfolio Stratified Sampling

The main advantage of the homogeneous portfolio lies in the fact that there are fewer variables describing the model points and so it is easier to choose the stratification variables. Also, since the model points are more similar, it is possible to divide the portfolio into fewer strata of more model points. This offers better opportunities to test the relationships within strata and to study the results based on various settings of the samples size.

Data Matrix and its Stratification

The data matrix should contain the reference variable and a selected set of stratification variables. The clustering approach has shown that using the present value of future cash flows from each model point as the clustering variable leads to the best results when estimating the development of liabilities. This suggests that PVCF is an appropriate variable to measure similarities between model points. Therefore, it seems a good choice to use PVCF as the reference variable for the stratified sampling. The PVCF values are obtained by applying equation (9).

A more challenging task is to select the stratification variables. The homogeneous portfolio contains following information about each model point: *age at entry, sex, annual premium, policy term, duration* and *sum assured.* Any of these can be used as a stratification variable. Some of the variables need to be adjusted though so that they can be used as categorical variables.

Since there are many stratification variables, the variables should not be allowed to take too many values because the number of strata would be too high and it could be even higher than the number of model points. Generally, the number of strata is given by

$$number\ of\ strata = \prod_{i=1}^{n} k_i$$

where k_i stands for the number of all possible values of the i-th variable and n is the number of used stratification variables.

In order that the stratification has any impact on the computational time, the variables should not take more than 4 different values, because using 5 variables taking 4 different values results in 625 strata. This number even multiplies in the case of the heterogeneous portfolio when more than 5 variables can be used. Therefore the variables are adjusted as follows:

Sex does not require any adjustments since it already is a categorical variable and it only takes two values.

Annual Premium needs to be discretized so that it can be used as a categorical variable. Annual premium values range from 432 to 498,612. These values should be broken into four intervals and the aim is that the intervals are of similar width and each interval contains a similar number of observations. Unfortunately, these aims are mutually exclusive and one needs to find a compromise. This is illustrated in Table 3.17 which shows the impact of fixing the number of model points within intervals and interval width respectively on the other argument.

Interval widths using a fixed number of observations	1,368	960	1,320	494,532
Number of model points within intervals using a fixed interval width	106,517	4	2	1

Table 3.17: Annual Premium Intervals
Source: Personal collection

The final choice of intervals is summarized in Table 3.18.

	(432, 2000]	**(2000, 4000]**	**(4000, 10000]**	**(10000, 498612]**
Interval Width	1,568	2,000	6,000	488,612
Number of model points	31,003	48,637	24,465	2,419

Table 3.18: Annual Premium Intervals after Finding a Compromise
Source: Personal collection

A similar approach is applied to *Sum assured.* The values range from 10,000 to 7,000,000 but the medium is only 100,000 and the 3rd quartile is 200,000 which means that if the values are divided into four equally wide intervals, more than 75% values falls into the lowest interval. Therefore, the chosen intervals cannot be of the same width, nor can they contain the same number of model points. The final choice of the intervals is summarized in Table 3.19.

	(10000, 50000]	(50000, 100000]	(100000, 200000]	(200000, 7M]
Interval Width	40,000	50,000	100,000	6,800,000
Number of model points	30,068	36,758	29,879	9,819

Table 3.19: Sum Assured Intervals
Source: Personal collection

The variable *Age at Entry* could be used as well but it seems more appropriate to calculate current age of the insured person since the liabilities calculation is based on *current age* rather than age at entry. The values are relatively evenly distributed so it is possible to divide the values into intervals of almost the same width. The intervals can be seen in Table 3.20.

	(29, 40]	(40, 50]	(50, 60]	(60, 71]
Interval Width	11	10	10	11
Number of model points	13,381	36,944	33,710	22,489

Table 3.20: Actual Age Intervals
Source: Personal collection

The final variable used in this study is *years to maturity* which is calculated as the difference between *policy term* and *duration*. Similarly to *actual age*, the values of *years to maturity* are relatively evenly distributed and can be divided into intervals of almost the same width. The created intervals are shown in Table 3.21.

	(0, 10]	(10, 20]	(20, 30]	(30, 41]
Interval Width	11	10	10	11
Number of model points	22,784	33,737	36,973	13,030

Table 3.21: Years to Maturity Intervals
Source: Personal collection

The five adjusted stratification variables can be used to define the categories based on which each model point can be classified into a stratum. Altogether, 512 categories are created from the combinations of the variables values. This means that the least random sample drawn from the dataset has 512 observations. To increase the possible time reduction, the number of categories should be reduced. This can be done either by excluding some of the stratification variables or by merging intervals of the variables.

In this study, four different settings of the number of strata are tested using successively 512, 64, 16 and 4 strata.

The task of finding the best selection of 64, 16 and 4 categories is based on their impact on the PVCF. Note that all the tested stratification variables are inputs of the PVCF calculation. This implies that reducing the number of variables or merging some variables values necessarily results in an increased error and intra-group variance. For this reason, one cannot compare different models unless they are based on the same number of variables values. This makes it impossible to use for instance ANOVA for the comparison since ANOVA is only applicable to nested models[3].

To compare the impact of the variables on PVCF, the generalized linear model is used for each variable and for all possible pairs of variables. In the model, PVCF is used as the response variable and the tested stratification variables are used as explanatory variables. In order to stick to the rule that the variables of each model should have the same number of possible values, the number of dummy variables of each trial model is set to four. This means that the explanatory variable can be either one of the variables which can take four different values or a pair of variables which can take two different values.

In the previous part, *annual premium, sum assured, current age* and *years to maturity* have been adjusted to categorical variables taking four different values. Any of these variables can be therefore tested separately using univariate GLM model. To test the variables pairs, the above-mentioned variables need to be merged into two intervals. For example, *current age* is only divided into intervals (0, 20] and (20, 41]. The analogical approach applies to *years to maturity, annual premium* and *sum assured*. Any pair of these variables together with the variable *sex* can be tested using bivariate GLM model.

The relationship between PVCF and each variable or pair of variables is measured using Akaike information criterion which is a measure of the relative ability of the model to estimate the value of PVCF which can be used for comparison of non-nested models. The criterion is defined as

$$AIC = 2k - 2\ln(\hat{L})$$

[3] Two models are said to be nested, if the parameters of one of the models are a subset of the parameters of the other model (Grace-Martin).

where k is the number of used dummies (in this case 4) and $\hat{L}$ is the maximum of the model's likelihood function. The criterion is only used for measuring a relative quality of a model when comparing models. A lower AIC value indicates a better fit (Snipes and Taylor).

The AIC values of all models are summarized in Table 3.22. One can see that *annual premium* is the variable with the strongest discriminatory power followed by *sum assured*. The least significant variable is *actual age*. From the variables combinations, the combination of *premium* and *sum assured* and the combination of *premium* and *sex* are the most significant ones while the least significant are *years to maturity* and *sex* combined with *actual age*.

Univariate GLM (4 Values per Variable)		Bivariate GLM (2 Values per Variable)	
Variable	**AIC**	**Variables**	**AIC**
Annual Premium	2,063,063	Premium + Sum Assured	2,085,692
Sum Assured	2,090,377	Premium + Sex	2,093,887
Years to Maturity	2,133,612	Premium + Years to mat.	2,094,045
Actual age	2,133,617	Premium + Actual Age	2,094,045
		Sum Assured + Sex	2,105,937
		Sum assured + Age	2,109,975
		Sum Assured + Years to Mat.	2,109,978
		Years to Maturity + Sex	2,130,345
		Actual Age + Sex	2,130,346
		Years to Maturity + Age	2,134,492

Table 3.22: Akaike Information Criterion Values
Source: Personal collection

Based on the criterion, the best combinations of stratification variables, as well as the intervals they are divided into can be selected for the chosen numbers of strata. The selections are summarized in Table 3.23.

Once the stratification variables have been defined, the model points can be classified into strata. However, even though the variables have been broken into intervals in such a way that there is a similar number of model points in every interval, the combinations of variables may result in categories with quite different numbers of model points.

In the case of 512 categories, only 188 strata can be created since there are no model points falling into other categories. In addition, there are some strata with very few observations among the 188 and some only contain one model point. It is obvious that such strata cannot be used for random sampling since either all or none of the model points could be selected which conflicts with random sampling principles. Even the strata with more than one but not enough observations are problematic because they restrict the possibilities of random sample sizes since the size is supposed to be an integer. For this reason, all strata with fewer than 20 model points are joint to the strata which are most similar to them. After such adjustment, only 122 strata are left for the stratified random sampling.

Number of Strata (full)	**Number of Strata (Adjusted)**	**Variables**	**Number of intervals**
512	122	Sex Actual age Annual Premium Years to Maturity Sum Assured	- 4 4 4 4
64	56	Sex Annual Premium Sum Assured Years to Maturity	- 4 4 2
16	15	Annual Premium Sum Assured Sex	2 4 -
4	4	Annual Premium Sum Assured	2 2

Table 3.23: Variables Selection
Source: Personal collection

The situation is similar with 64 categories where only 58 strata created out of the categories contain model points. Since there are fewer strata, they are required to contain more model points. This is why not 20 but 40 model points within strata are used as the required minimum and strata with fewer observations are joined to their closest stratum. After such adjustment, only 56 strata are left.

In the case of 16 categories, there are no empty strata created. However, there is one stratum with only 36 model points which needs to be joined to another category. The smallest stratum then contains 600 model points.

In the final case using 4 categories, no adjustment is needed since there are enough model points in each stratum created based on the categories. The stratum with the least number of model points contains 7,522 model points.

Random Sample Selection

Once the stratification variables have been determined, it is possible to select a stratified random sample. The allocation used in this paper is the proportional allocation. This means that the proportion of the sample taken from any stratum is equal to the proportion of the stratum in the entire portfolio.

The selection of the proportion is a crucial element of the whole study as it may influence both computational time and accuracy of the method. For this reason, multiple settings of the proportion are tested and the accuracy of the results, as well as the computational time, are used to compare the efficiency of each setting. The error is calculated using equation (7) comparing accurate calculation of LAT reserve and the calculation based on the random sample.

Since the selection is random, a different sample is drawn from the dataset every time and thus a different reduced portfolio is created giving different results. Therefore, any setting of the method needs to be tested multiple times and the properties of the each can be measured based on the maximum, minimum, medium or average value. In this study, ten trials are performed for each choice of proportion.

Proportion	Number of Model Points	122 Strata		56 Strata		15 Strata		4 Strata	
		Time	Error	Time	Error	Time	Error	Time	Error
0.005%	5	0.01	92.21%	0.01	29.28%	0.01	28.92%	0.01	91.27%
0.01%	11	0.02	17.36%	0.02	18.58%	0.01	24.94%	0.02	19.63%
0.05%	53	0.04	8.60%	0.04	5.04%	0.04	6.34%	0.04	6.48%
0.1%	107	0.15	2.97%	0.16	2.60%	0.15	4.17%	0.14	4.72%
0.5%	533	0.44	1.21%	0.5	1.91%	0.4	2.42%	0.4	3.15%
1%	1,065	6.2	0.41%	8.6	0.69%	6.3	0.59%	5.8	1.43%
5%	5,326	9.9	0.27%	12.8	0.41%	10.0	0.28%	9.3	1.24%
10%	10,652	16.9	0.11%	20.2	0.33%	17.0	0.27%	15.6	0.52%
20%	21,305	22.4	0.09%	25.1	0.13%	22.4	0.27%	21.2	0.37%

Table 3.24: Results Based on the Proportion of Model Points Selected
Source: Personal collection

Table 3.24 summarizes the average results of some of the tested settings of the proportion of model points selected. The trial was performed on the portfolio divided into 122, 56, 15 and 4 strata respectively using the clustering variables summarized in Table 3.23.

All the applications give comparable results and not surprisingly, the computational time is increasing while the approximation error is decreasing with the increasing proportion.

One can notice that with low proportion, there would be no random sample drawn from small strata. For instance, if the portfolio is divided into 122 strata, some strata only contain 20 model points. However, strata of the size 20 only enter into the sample if the proportion is at least 0.05%. If a lower proportion is used, some model points are automatically ignored based on their properties and they never enter into the computation of life liabilities. This represents a chief drawback of the method and particularly, if more strata are used, it causes very a high error with low proportion.

The problem could be solved either by using only higher proportion values for a high number of strata or by merging strata with fewer model points. Another option is to divide the model points into strata in a way that the strata are of a similar size. On the one hand, such approach does not allow using some combinations of stratification variables and in most cases, it can only be achieved with as few as one stratification variable. On the other hand, it ensures that every model point is equally likely to be selected. Both approaches are tested at the end of this section.

To compare the quality of each of the four stratifications, the relationships between their computational time and error are plotted in Figure 3.12. The plot is based on average values of the 10 trials and it only shows a detail for error lower than 10%. One can see that the best results can be achieved with 122 strata because it enables one to get the most accurate results at the lowest computational time. On the other hand, using 4 strata leads to the worst results.

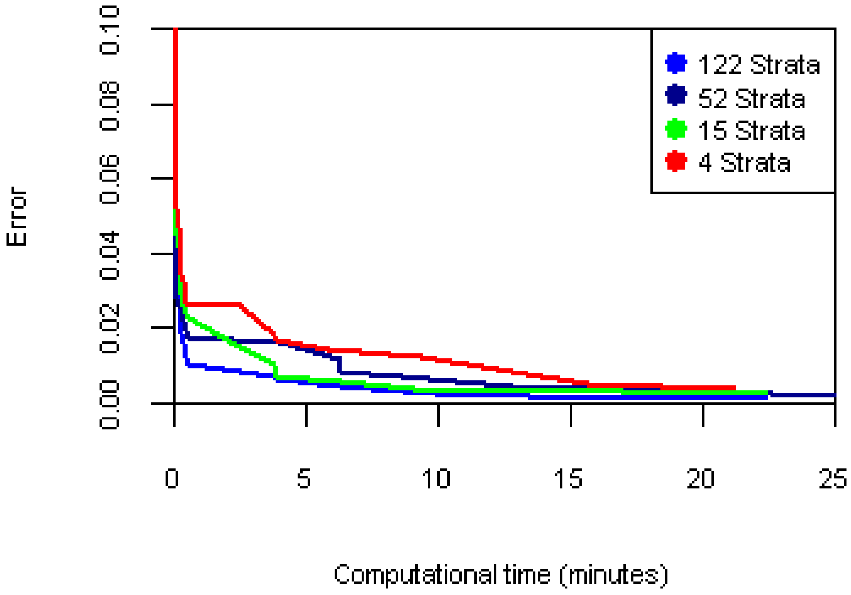

Figure 3.12: Efficiencies Based on the Number of Strata
Source: Personal collection

Strata with not enough Model Points

In the previous section, it has been mentioned that choosing the combination of stratification variables by measuring their combined impact on the reference variable may cause that some strata have very few model points. This makes it impossible to select a random sample of a very low size because there would be no sample selected from the smallest strata.

The aim of this subchapter is to present an alternative approach to the selection of stratification variables whose objective is to create strata of equal sizes. In order to achieve this, only one clustering variable is chosen and it is broken into intervals based on quantiles. The previous part has measured Akaike information criterion for every variable which could be chosen as a stratification variable. The measure is lowest for annual premium so it seems appropriate to use annual premium as the stratification variable. Using annual premium has also the advantage that insurance companies set the premium based on the properties of the client which means that other variables such as age and sex may be to a certain extent reflected in the premium value.

The annual premium needs to be discretized so that it can be used as a categorical variable. Since it is not clear how many intervals are optimal, the variable is discretized twice and the results are compared. First, it is broken into 10 categories of equal sizes and then, it is broken into 100 categories. The selection of the endpoints of the intervals is based on finding the deciles and percentiles of the variable.

It is not possible to create strata of exactly the same size, though, because in some cases, the endpoints of the intervals represent a frequent value and all model points with the same value of premium are classified into the same group regardless of whether it increases the size of one stratum and decreases the size of the other. With 100 intervals, some intervals are merged completely because there are values of premium which occur with more than 1% of the model points. However, most strata are of similar size and if some strata are larger it does not cause any real problem.

Proportion	Number of Model Points	10 Strata		100 Strata	
		Time	Error	Time	Error
0.01%	11	0.09	34.10%	0.08	80.08%
0.05%	53	0.29	3.82%	0.38	8.85%
0.1%	107	0.63	2.56%	0.81	1.90%
0.5%	533	2.58	1.38%	3.3	1.14%
1%	1,065	22.7	0.80%	58.1	0.45%
5%	5,326	42.1	0.77%	75.0	0.40%
10%	10,652	65.0	0.50%	93.8	0.27%
20%	21,305	114.6	0.20%	136.0	0.12%

Table 3.25: Results Using Annual Premium as the only Stratification Variable
Source: Personal collection

The results for both 10 and 100 strata are summarized in Table 3.25. Both applications give comparable results to the previous ones. Even though the approach should work better for low proportion than the previous one, the error is extremely high and it is obvious that the method cannot be applied with low proportion either. Therefore, it seems that the approach does not really solve the problem of unequal sizes of the strata.

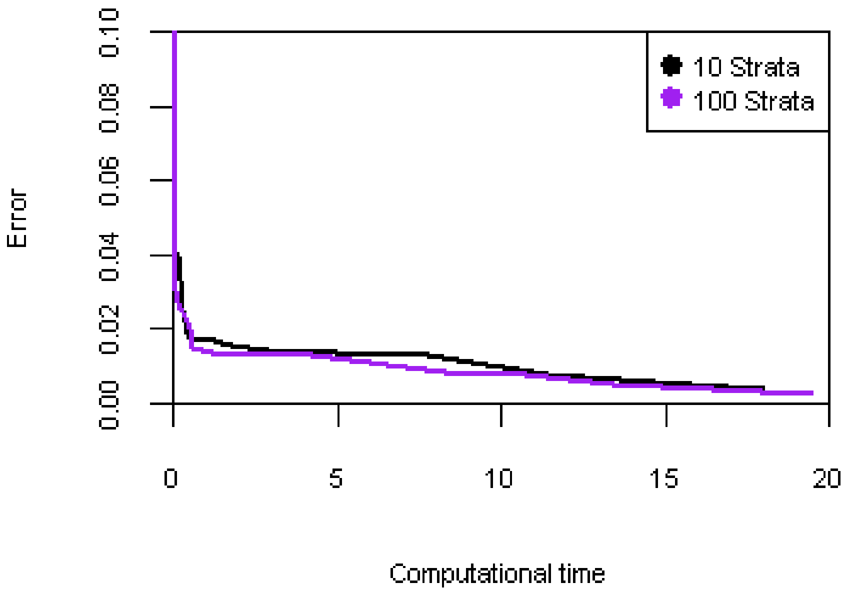

Figure 3.13: Efficiencies Using Annual Premium as the only Stratification Variable
Source: Personal collection

The relationship between the computational time and approximation error of both applications can be seen in Figure 3.13. One can see that using 100 strata leads to more accurate results at the same computational time than using 10 strata.

So far, six different selections of stratification variables have been tested. In order to compare each selection, all of them are put in one plot in Figure 3.14. One can see that from all the six combinations, the combination resulting in 122 strata is the most efficient one. This means that the best combination of stratification variables is *sex*, *annual premium*, *years to maturity* and *sum assured*. All the mentioned variables except *sex* are divided into four intervals.

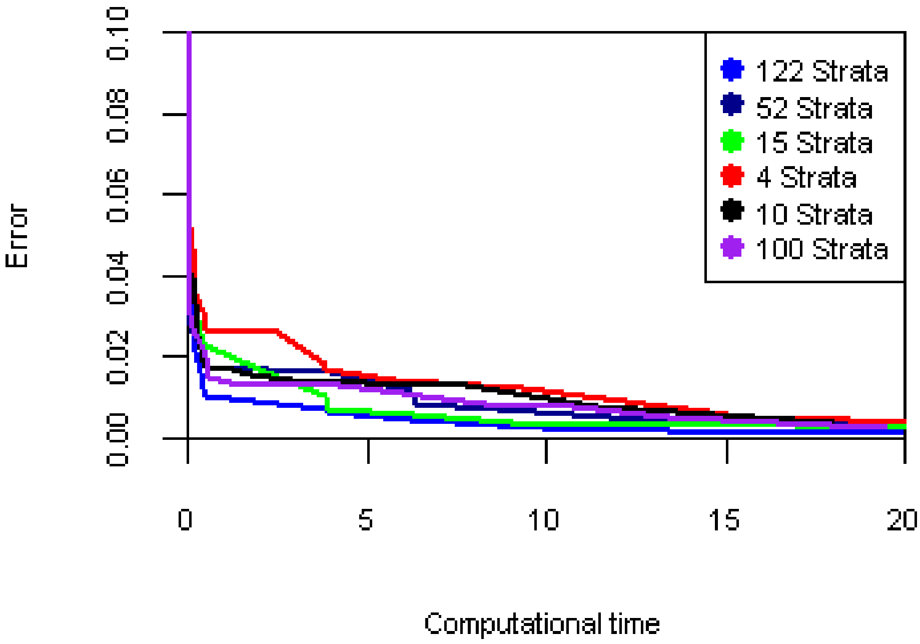

Figure 3.14: Efficiencies of all the Tested Stratification Variables Combinations
Source: Personal collection

Another approach to solving the problem of some strata being too small is to merge the categories with not enough observations. In this study, all strata with fewer than 100 model points are joined to their closest category. Applying this approach to the most efficient model with 122 strata reduces the number of strata to 78.

Figure 3.15 shows the efficiencies of the original setting which divides the portfolio into 122 strata and the new setting where the strata with fewer than 100 model points are joined to other strata. The results are quite similar but in the major part of the time interval, slightly more accurate results can be obtained with 78 strata. In addition, the model with 78 strata is more appropriate because it does not ignore any model points provided that the proportion of model points selected is higher than 0.09% so it seems to be the best model.

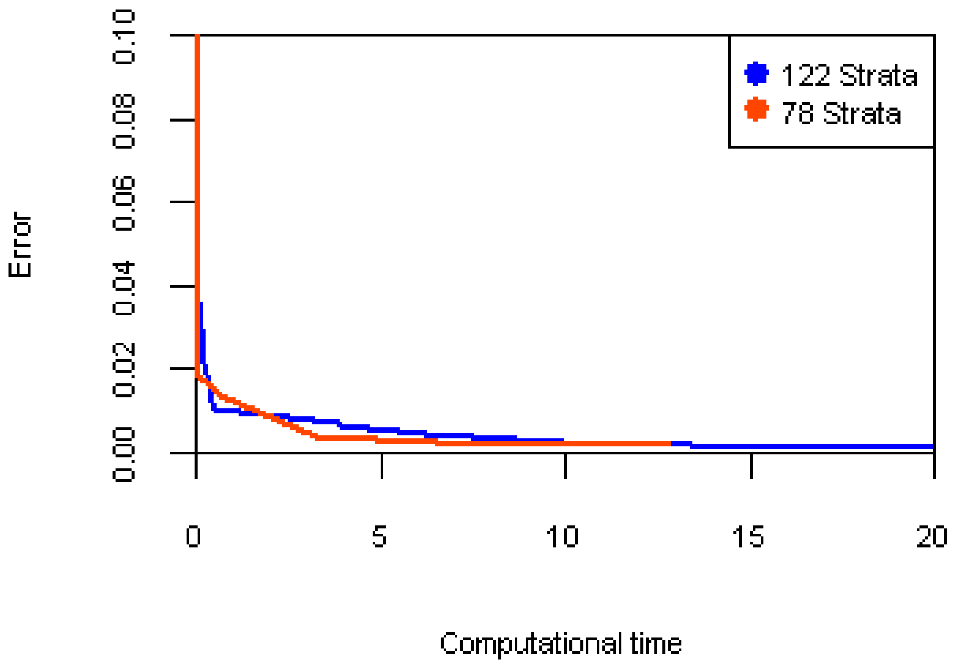

Figure 3.15: Efficiencies after Merging the Smallest Strata
Source: Personal collection

It could be concluded that the final model with 78 strata gives relatively good results and if the insurance company requires a minimum accuracy of 99.5% it can obtain the results within 3 minutes selecting 5 % of the model points.

3.4.2 Heterogeneous Portfolio Stratified Sampling

In this subchapter, the stratified random sampling model optimized in the previous part is applied to the real portfolio of life policies. The application to the heterogeneous portfolio is slightly complicated by the fact that new variables are added which can be used as stratification variables. The variables that seem most likely to have a discriminatory power in the model are *product type* and *benefit.* It might be advisable to combine these two variables with the ones used for homogeneous portfolio or to replace some of the used variables with the new ones.

Both *product type* and *benefit* are categorical variables taking four different values so no adjustment of the variables is needed. The discriminatory power of each variable is measured using Akaike information criterion of the univariate general regression model for the heterogeneous portfolio. However, the combinations of variables cannot be tested for either *product type* or *benefit* because the combinations are supposed to take four different values altogether but there is no reasonable way of merging the values of *product type* or *benefit.*

Akaike criterion of each applicable variable and variables combination can be seen in Table 3.26. Looking at the variables separately through the results of the univariate GLM one can see that the newly added variables are those with the highest AIC and therefore also the ones with the lowest likelihood function. This means that adding the variables to the model may not necessarily bring any improvement.

Univariate GLM (4 Values per Variable)		Bivariate GLM (2 Values per Variable)	
Variable (4 values)	**AIC**	**Variables (2 values)**	**AIC**
Annual Premium	1,993,076	Premium + Sum Assured	2,001,363
Sum Assured	1,998,755	Premium + Years to mat.	2,006,117
Years to Maturity	2,031,395	Premium + Actual Age	2,006,130
Actual age	2,031,434	Premium + Sex	2,008,388
Benefit	2,031,986	Sum Assured + Sex	2,010,989
Product Type	2,036,288	Sum Assured + Years to Mat.	2,013,823
		Sum assured + Age	2,013,834
		Years to Maturity + Sex	2,028,968
		Actual Age + Sex	2,028,979
		Years to Maturity + Age	2,032,537

Table 3.26: Akaike Information Criterion Values (Heterogeneous Portfolio)
Source: Personal collection

For this reason, the first setting of stratification variables for heterogeneous portfolio used in this study is the one with 78 strata marked as the best for homogeneous stratification. It uses *sex* together with *annual premium, sum assured, years to maturity* and *actual age* divided into four intervals as stratification variables and it joins all strata with fewer than 100 variables to the stratum which is most similar to them.

The second setting adds *benefit* to the above-mentioned variables. This means that the number of strata should increase by four. Theoretically, it should be possible to create 2048 strata because the combined clustering variables can take 2048 different values altogether. However, most strata are empty or contain very few model points. After joining strata with fewer than 50 model points to their closest stratum, only 260 strata are left.

The final setting uses *annual premium, sum assured, years to maturity* and *benefit* as clustering variables and it sets the minimum of model points within strata to 100. Such setting results in 137 strata. The variable *product type* is not used in this study at all because it seems to have very little impact on the PVFC value.

The results are summarized and compared in Table 3.27 showing the computational time and approximation error for each of the defined settings of stratification variables and for some of the tested proportion values. The comparison is even better visible in Figure 3.16 showing the relationship between the computational time and error for each variables combination. All three models give similar results and it is not easy to decide which model is the best because each model gives the best results in at least one time interval. However, the model with 78 strata gives the best results in the longest time interval and it also enables one to achieve the lowest error so it could be considered the best. This is also consistent with the performed regression analysis which showed that the variables used in the model with 78 strata are those with the strongest impact on the PVFC development.

Proportion	Number of Model Points	78 Strata		260 Strata		137 Strata	
		Time	Error	Time	Error	Time	Error
0.05%	53	1.22	17.84%	1.20	68.08%	1.20	18.06%
0.1%	107	3.64	5.92%	1.64	20.13%	3.61	12.18%
0.5%	533	6.03	1.32%	6.0	1.61%	6.0	0.89%
1%	1,065	7.2	0.37%	6.2	0.99%	6.7	0.55%
5%	5,326	12.1	0.28%	12.2	0.11%	12.3	0.31%
10%	10,652	24.2	0.18%	24.4	0.06%	23.8	0.22%
20%	21,305	36.5	0.02%	36.6	0.04%	36.7	0.09%

Table 3.27: Heterogeneous Portfolio Results Based on Proportion and Number of Strata
Source: Personal collection

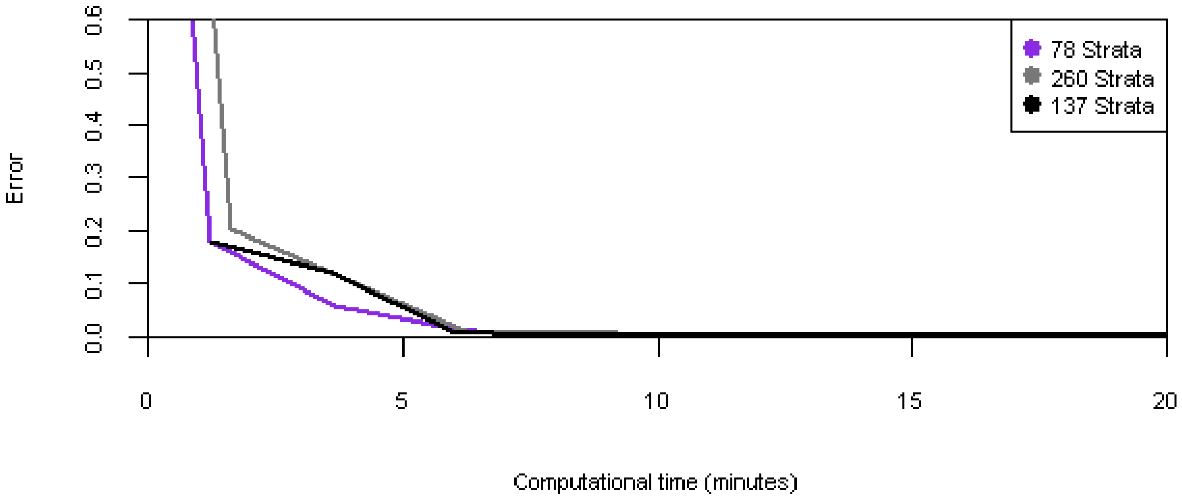

Figure 3.16: Efficiencies of the Stratification Variables Combinations for Het. Portfolio
Source: Personal collection

It could be concluded that the best model for stratified random sampling of both the homogeneous and the heterogeneous portfolio is the model using *sex, annual premium, sum assured*, *years to maturity* and *actual age* as stratification variables. In order to see whether stratified random sampling works equally well for homogeneous and heterogeneous portfolio, the results are put in one plot in Figure 3.17.

One can see that the results for the heterogeneous portfolio are not as good as those for the homogeneous one, especially with a low proportion of model points selected. This is caused primarily by the increased computational time of all computations performed on the heterogeneous portfolio. Nevertheless, it could be said that stratified random sampling works relatively well and if the insurance company requires the minimum accuracy of 99.5% it can get the results in 7.2 minutes which is more than 18 times faster than with the accurate method.

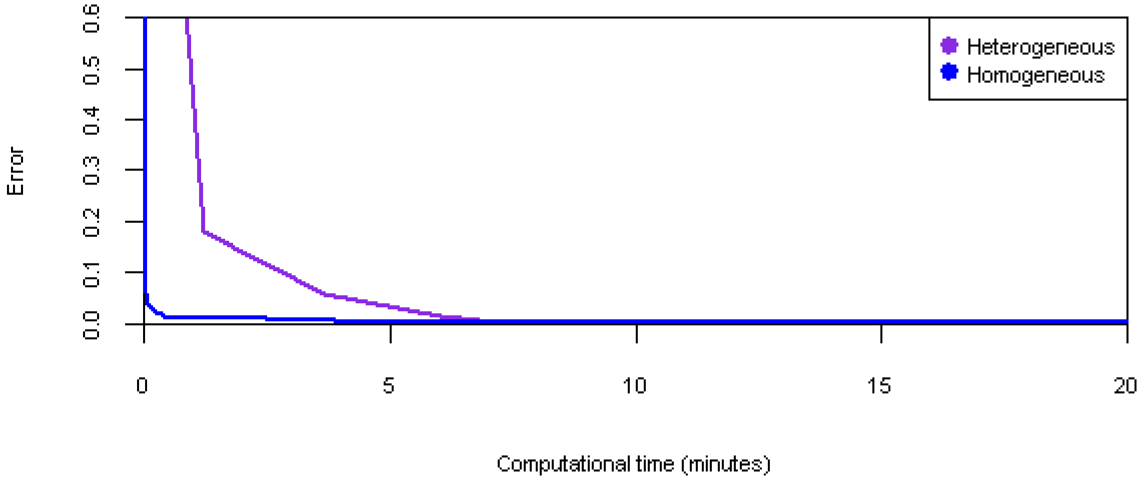

Figure 3.17: Heterogeneous and Homogeneous Portfolio Comparison
Source: Personal collection

4 Comparison of the Methods

The aim of the final chapter of the thesis is to compare the tested methods and find the relative strengths and weaknesses of each one. Insurance companies need to be able to obtain the results under given assumptions at the highest possible accuracy. However, the accuracy should not be achieved at the cost of an extremely high computational time because it would either prevent the insurance company from testing enough scenarios or it might happen that the liabilities values will have changed by the time the insurance company obtains the results. The computational time and accuracy seem to be crucial aspects of the computation of life liabilities. This is why the comparison between the methods is based primarily on these two properties of each method.

The first method is cash flow analysis which is based on calculating the projected cash flows from each model point for each year in which the policy may be in force. The main strength of the method lies in the fact that it enables one to get a relatively accurate estimation of liabilities under given assumptions. Supposing the insurance company knew the values of all the unknown variables including the time of the death of the insured person, they could calculate the liabilities at a 100 percent accuracy. The main weakness of the method is a very high computational time and memory usage, which makes it difficult for the insurance company to test enough scenarios under different assumptions.

The second method is cluster analysis. The method is based on selecting a given number of model points from the whole dataset which could be considered the best representatives of the whole portfolio based on their distances from other model points. The process of finding the representative model points is slightly time-consuming and it requires prior calculation of the present value of future cash flows from each model point. On the other hand, once the representative model points have been found, it is possible to reduce the time of every scenario computation at the expense of not getting completely accurate results.

The final method is stratified random sampling which selects a given number of model points in such a way that the selection should behave in a similar way to the whole dataset. The method does not require any prior computation and is, therefore, the fastest one. On the other hand, the samples selected from the dataset are random and they may not be the best representatives of the portfolio, which may lead to a higher computational error.

Since all the methods were tested for both the homogeneous and the heterogeneous portfolio, the comparison of the methods is performed separately for either type of the portfolio.

4.1.1 Comparison of the Methods for the Homogeneous Portfolio

In order to compare the computational times per scenario of each method, it is necessary to take into consideration the total time of all the operations. This is slightly problematic because the time of the creation of the representative portfolio in the case of cluster analysis and stratified random sampling should be included in the computational time too. However, since the creation of the representative portfolio is only performed once, only a part of its computational time should enter into the computational time of one scenario. This means that in the case of cluster analysis and stratified random sampling, the computational time per scenario is decreasing with the increasing number of scenarios. What is more, since it is faster to create the representative portfolio using stratified random sampling, the computational time is decreasing faster with cluster analysis.

As the computational time per scenario decreases in inverse proportion to the number of scenarios, two different situations are considered. In the first comparison, the insurance company only tests 6 different scenarios assuming successively the best estimate scenario and the scenarios with one of the five assumptions (mortality, lapses, commissions, investment return and expenses) increased by 1%. This number of scenarios is the bare minimum to calculate the LAT reserve and in reality, insurance companies need to test much more scenarios than that.

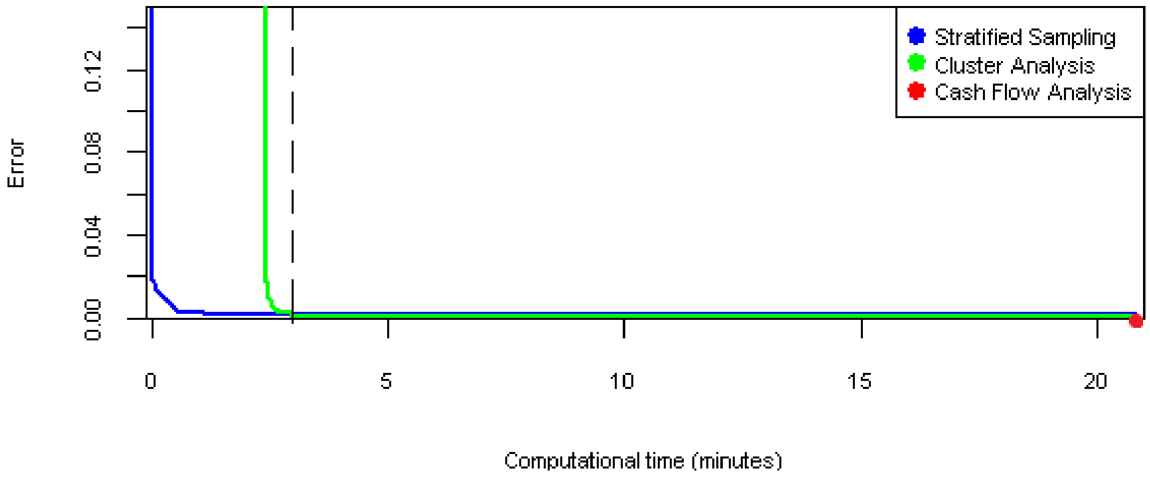

Figure 4.1: Methods Comparison with 6 Scenarios
Source: Personal collection

Figure 4.1. shows the relationship between the computational time and error for each method. The computational time is the time of the calculation of one scenario and in the case of cluster analysis and stratified sampling, it also includes one-sixth of the time of the representative portfolio creation.

Looking only at the efficiency of the methods, which means the ability to achieve the highest accuracy at the lowest time, stratified random sampling is the best method for the computational time per scenario lower than 3 minutes and 10 seconds. In 3:10 minutes it is possible to achieve the accuracy of about 99.9% with both cluster analysis and stratified random sampling. After this point, higher accuracy can be achieved at any time using cluster analysis until the time of 21 minutes. In 21 minutes, one can achieve the accuracy of 100% using cash flow analysis. Such accuracy can never be achieved with either cluster analysis or stratified random sampling unless the number of clusters is the same as the number of model points or the proportion of model points selected from each stratum equals 1. In such cases, however, the computational time is higher than 21 minutes. The accuracy achievable within 21 minutes using cluster analysis is 99.98%.

It could, therefore, be concluded that if the insurance company has a homogeneous portfolio of model points and only tests 6 scenarios to evaluate the liabilities, the selection of the method for liabilities calculation would be as follows:

- If the insurance company requires the accuracy of at least 99.98%, they would use cash flow analysis and they would be able to calculate the results at 100% accuracy at the computational time of 21 minutes per scenario.
- If they require a higher accuracy than 99.9% but not necessarily as high as 99.98% they would use cluster analysis and depending on the required accuracy, they would be able to compute the results of one scenario between 3:10 and 21 minutes.
- If they need to obtain the results in less than 3:10 minutes per scenario, they would use stratified random sampling and they would have to content themselves with the accuracy of 99.9% or lower depending on the required computational time.

However, the results of the comparison are biased by the fact that the insurance company only tests 6 scenarios. Assuming they test 100 scenarios, the impact of the time spent creating the clusters and strata is smaller and the results change as it can be seen in Figure 4.2.

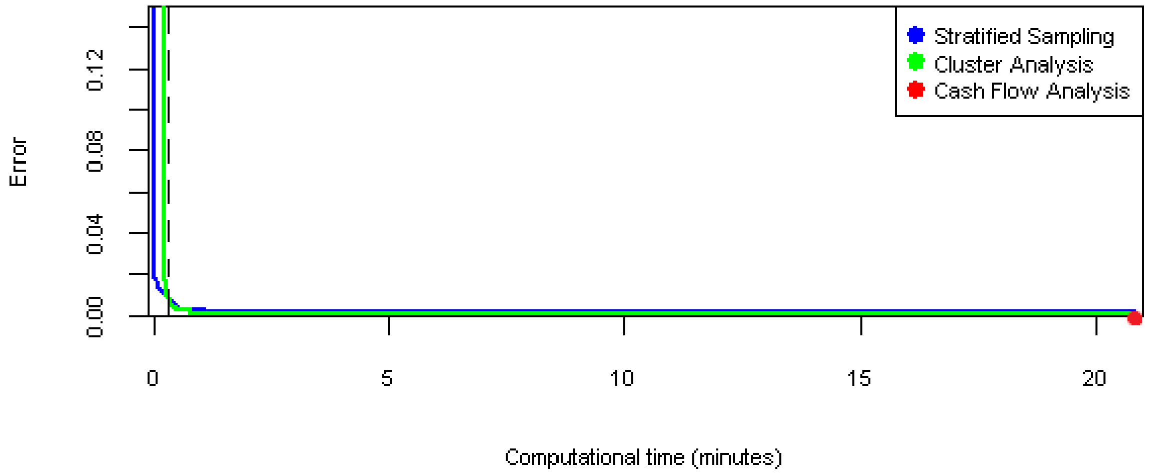

Figure 4.2: Methods Comparison with 100 Scenarios
Source: Personal collection

When testing 100 scenarios, the critical point for switching between the methods is the computational time of 32 seconds per scenario. In such time, it is possible to obtain the results at the accuracy of 99.1% with both cluster analysis and stratified random sampling. If the insurance company requires higher accuracy than 99.1% but not necessarily higher than 99.98% they would use cluster analysis. If they content themselves with lower accuracy than 99.1% but require a lower computational time than 32 seconds per scenario, they should use stratified random sampling.

It could be concluded that the more scenarios are tested, the more likely the insurance company is to prefer cluster analysis to stratified random sampling.

There is one more aspect which should be taken into consideration and that is the stability of the results. All the results which have been presented so far are based on the average value of ten trials. The comparison should not be based solely on the average value, though. One should consider the variability of the results as well.

Table 4.1 and Table 4.2 show the average, minimum, maximum and the spread between the minimum and the maximum error of cluster analysis and stratified sampling results. Taking into account the approximation error only and not the computational time, the average error is generally lower with cluster analysis. However, looking at the maximum error and the spread between the maximum and minimum error, stratified random sampling results are slightly better since for a sufficiently high proportion of model points selected, the spread is less than 1%. If the insurance company selects the method on the basis of maximum error rather than its average value, they would always prefer stratified random sampling to cluster analysis because it computes the results with a lower maximum error at a lower computational time.

Cluster Analysis - Error				
Number of Clusters	**Average**	**Minimum**	**Maximum**	**Spread**
2	11.72%	8.45%	13.85%	5.40%
5	8.29%	7.43%	10.07%	2.63%
10	7.56%	6.95%	9.31%	2.36%
50	3.59%	2.73%	6.04%	3.31%
100	2.54%	1.48%	3.83%	2.35%
200	1.41%	0.86%	2.13%	1.27%
300	0.29%	0.17%	1.28%	1.11%
500	0.09%	0.07%	0.44%	0.37%

Table 4.1:Cluster Analysis Results Variability
Source: Personal collection

Stratified Random Sampling - Error				
Proportion of Model Points	**Average**	**Minimum**	**Maximum**	**Spread**
0.01%	90.58%	90.41%	91.36%	0.95%
0.05%	2.57%	3.47%	16.93%	13.46%
0.1%	1.84%	1.65%	5.81%	4.17%
0.5%	1.71%	1.24%	3.72%	2.47%
1%	1.34%	0.95%	1.37%	0.42%
5%	0.36%	0.35%	0.95%	0.61%
10%	0.23%	0.13%	0.86%	0.73%
20%	0.16%	0.04%	0.26%	0.21%

Table 4.2: Stratified Sampling Results Variability
Source: Personal collection

4.1.2 Comparison of the Methods for the Heterogeneous Portfolio

The comparison of the methods for the heterogeneous portfolio is performed in the same way as the comparison for the homogeneous portfolio. Again, the computational time of one scenario should include a part of the time of the creation of strata or clusters apart from the time of the computation of liabilities itself. As the computational time per scenario decreases in inverse proportion to the number of scenarios, two different situations are considered.

First, it is assumed that the insurance company only tests 6 different scenarios. The efficiency of each of the three methods in such a case is summarized in Figure 4.3. One can see that using stratified random sampling enables the insurance company to achieve relatively accurate results at the lowest time. The critical point is the computational time of 3:30 minutes. At such time both stratified sampling and cluster analysis calculate the results with the average error of 0.2%. For a higher computational time than 3:30 minutes, cluster analysis gives more accurate results.

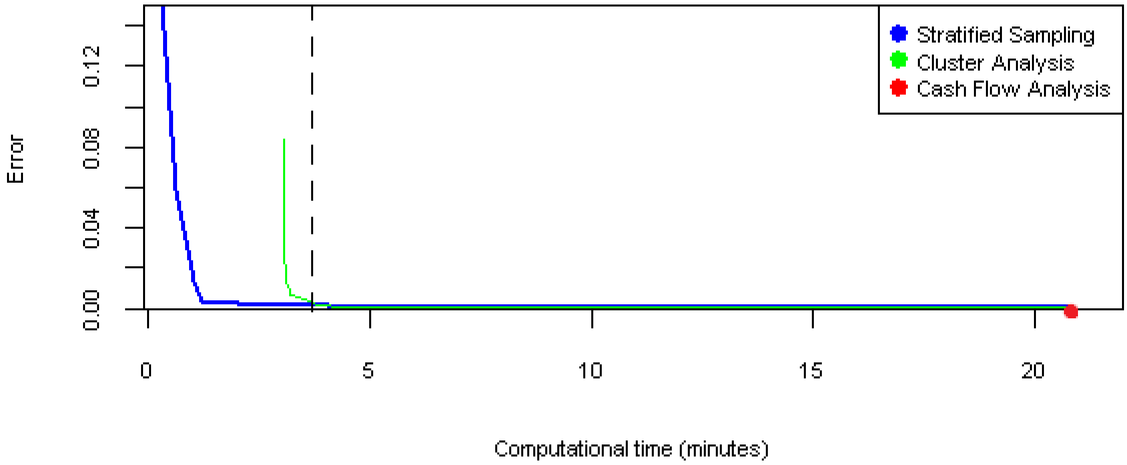

Figure 4.3: Methods Comparison for the Heterogeneous Portfolio with 6 Scenarios
Source: Personal collection

The selection of the method could be summarized as follows:

- If the insurance company needs to calculate the results in less than 3:30 minutes and the accuracy of 99.8% or lower is acceptable for them, they would use stratified random sampling.
- If the insurance company requires a higher accuracy than 99.8% but accept the error of 0.01% or slightly more, they would use cluster analysis.
- If the insurance company requires the accuracy of more than 99.99%, they would calculate the accurate results using cash flow analysis.

An interesting case is the second situation where it is assumed that the insurance company tests 100 scenarios. In such a case, the effect of the increased computational time of cluster analysis due to the rather time-consuming creation of clusters is almost eliminated and cluster analysis always enables the insurance company to achieve more accurate results at a lower computational time.

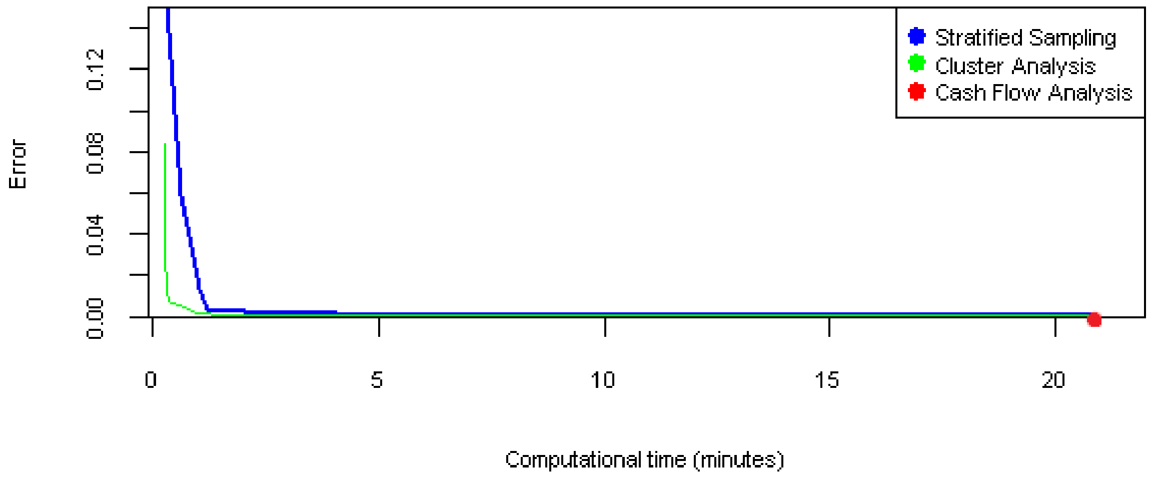

Figure 4.4: Methods Comparison for the Heterogeneous Portfolio with 100 Scenarios
Source: Personal collection

Therefore, it could be summarized that if the insurance company needs to get the results at a lowest possible time and only tests few scenarios, they would prefer stratified random sampling. With the increasing number of tested scenarios or with the increasing accuracy demand they are more likely to prefer cluster analysis.

Similarly to the comparison for the homogeneous portfolio, the comparison for the heterogeneous portfolio should not be based only on the average error since the insurance company may be more concerned with the maximum error or other measures.

Cluster Analysis - Error				
Number of Clusters	**Average**	**Minimum**	**Maximum**	**Spread**
2	7.15%	4.29%	7.83%	3.54%
5	3.33%	2.27%	4.38%	2.11%
10	2.57%	1.35%	3.99%	2.64%
50	2.49%	1.31%	3.88%	2.57%
100	1.67%	1.20%	2.28%	1.08%
200	0.80%	0.60%	1.52%	0.92%
300	0.08%	0.07%	1.12%	1.05%
500	0.01%	0.01%	0.95%	0.94%

Table 4.3:Cluster Analysis Results Variability - Heterogeneous Portfolio
Source: Personal collection

Stratified Random Sampling - Error				
Proportion	**Average**	**Minimum**	**Maximum**	**Spread**
0.01%	90.97%	90.72%	91.33%	0.60%
0.05%	17.84%	17.07%	18.52%	1.45%
0.1%	5.92%	4.41%	8.15%	3.74%
0.5%	1.32%	0.74%	1.69%	0.96%
1%	0.37%	0.31%	0.41%	0.10%
5%	0.28%	0.25%	0.30%	0.05%
10%	0.18%	0.15%	0.20%	0.04%
20%	0.02%	0.00%	0.05%	0.05%

Table 4.4: Stratified Sampling Results Variability - Heterogeneous Portfolio
Source: Personal collection

Table 4.3 and Table 4.4 show the average, minimum and maximum error and the spread between the maximum and the minimum error for either tested method. Looking at the approximation error only, one can conclude that cluster analysis gives more accurate results, on the average. However, the results obtained using cluster analysis are generally more unstable. The maximum error is higher and so is the spread between the maximum and the minimum error. This means that even though the results are usually more accurate, it might occur that the error is higher than it would be with stratified sampling. This is why some insurance companies may prefer stratified random sampling to cluster analysis because they are more concerned with the maximum error than with the average error.

Conclusion

Estimating the liabilities of a life insurance company is a demanding task which involves predicting the future development of cash flows from all the policies in the company's portfolio. The traditionally used method, cash flow analysis, seeks to obtain the results with the highest possible precision and thus it models the path of each model point separately. However, such approach is extremely time-consuming and it both limits the number of scenarios that can be tested and causes a delay in obtaining the results. Therefore, the accuracy of the results may not always be considered the top priority and insurance companies may prefer a faster method even at the cost of a lower precision.

The aim of the thesis was to find an alternative method for faster estimation of life liabilities. To accomplish this task, two approximation methods were tested, cluster analysis and stratified random sampling. Both methods are based on selecting a limited number of model points from the portfolio, assigning them certain weights and calculating future cash flows from the weighted model points. Both the methods start with dividing all the model points in the portfolio into a selected number of groups in such a way that the model points in each group are as similar to each other and as dissimilar to all other model points as possible. In the case of cluster analysis, the most centrally located model point is selected from each group as a representative model point while in the case of stratified random sampling, a random sample of a given size is drawn from each group.

Both methods were first applied to a modified portfolio of homogeneous contracts in order that the settings of the methods could be optimized. Only then, the applicability of either method was tested using the real portfolio of heterogeneous life insurance contracts. The optimization of the methods settings was based on measuring the relationship between the computational time and accuracy of the results. In the case of cluster analysis, the highest efficiency can be achieved using CLARA as a clustering algorithm, Manhattan distance as a distance measure and the present value of future cash flows from individual model points as a clustering variable. The number of samples should be at least as high as the size of the clusters, which should be the same as the number of samples. In the case of stratified random sampling, the most accurate results at the lowest computational time can be obtained using present value of future cash flows from individual model points as a reference variable, the combination of annual premium, sum assured, years to maturity and actual age, each of which is divided into four intervals, as a stratification variable and joining the strata with fewer than 100 model points to their closest neighbor.

Applying the methods to homogeneous and heterogeneous portfolio respectively, it has been shown that both the methods enable insurance companies to calculate their liabilities at a relatively high accuracy with a significant time-reduction. Comparing the methods with each other and with cash flow analysis was slightly complicated by the fact that both cluster analysis and stratified random sampling require a prior construction of the representative portfolio, whose computational time should be divided among the scenarios. This causes that the computational time per scenario is decreasing with the increasing number of scenarios and what is more, it is decreasing at a different rate for each method. Therefore, the comparison needs to be performed for different numbers of scenarios.

Comparing the methods' application to a homogeneous portfolio, it could be concluded that if the insurance company only tests 6 scenarios, they are likely to prefer stratified random sampling which enables them to calculate the results with the accuracy of up to 99.98% within 3:10 minutes, which is about 7 times faster than using cash flow analysis. The more scenarios are tested, the more likely the insurance company is to prefer cluster analysis and if they test 100 scenarios they would only use stratified random sampling if they content themselves with the accuracy of up to 99.1% and for a higher accuracy, they would use cluster analysis.

In the case of a heterogeneous portfolio, the results of the comparison are slightly more favorable to cluster analysis. If only 6 scenarios are tested, the insurance company is more likely to prefer stratified random sampling which enables them to obtain the results with the accuracy of up to 99.8% within 3:30 minutes and only if they require a higher accuracy than that, they would choose cluster analysis. However, if they test 100 scenarios or more, the insurance company would always prefer cluster analysis to stratified random sampling, because it leads to a higher accuracy at a lower computational time per scenario.

With both the homogeneous and heterogeneous portfolio, the traditional cash flow analysis is only used if the insurance company requires the accuracy of more than 99.99%. If they content themselves with a lower accuracy, they can calculate the results with a significant time reduction using either of the approximation methods.

Since the development of some of the input variables of the life liabilities calculation is unknown, it is necessary to consider their deviation from the predicted path and measure the impact of the deviation on the liabilities development. Therefore, it might be considered more important to test multiple scenarios than obtaining the results at the 100 percent accuracy. Cluster analysis and Stratified random sampling increase the possibilities of the insurance company to test multiple scenarios. For example, at the cost of 1% error, the insurance company can test 20 times more scenarios using stratified random sampling than if they used the traditional method. It could, therefore, be concluded that both cluster analysis and stratified random sampling represent convenient tools for life liabilities estimation and they improve the possibilities of the life insurance company to evaluate the possible changes in the liabilities due to the changes in the input variables.

Finally, it should be mentioned that the study was slightly limited by the fact that the portfolio of policies used to test the methods only contains contracts which combine life insurance with some form of an investment and there are no survival benefits contracts or other types of insurance contracts in the portfolio. It could, therefore, be a subject of a further study, to test the method's applicability to a more diverse portfolio of a life insurance company.

Works Cited

1. Aczel, D Amir and Jayavel Sounderpandian. *Complete Business Statistics*. 7th ed. McGraw-Hill, 2008.

2. Allianz. *Penzijní plan.* Allianz transformovaný fond, 2010. https://www.allianz.cz/file/17096/Penzijni_plan_D.pdf. Accessed 17 November 2017.

3. Camilli, Stephen J., Ian Duncan, and Richard L. London. *Models For Quantifying Risk*. 6th ed. Actex Publications, 2012.

4. Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services. OJ L 373/37.

5. Czech National Bank. *CNB current forecast.* 2 November 2017. https://www.cnb.cz/en/monetary_policy/forecast/index.html. Accessed 2 December 2017.

6. Czech Statistical Office. *Consumer price index according to COICOP - basic index*. https://vdb.czso.cz/vdbvo2/faces/en/index.jsf?page=vystup-objekt&pvo=CEN080&z=T&f=TABULKA&skupId=43&katalog=31779&pvo=CEN080&evo=v2300_!_CEN-SPO-BAZIC2005-R2_1. Accessed 2 December 2017.

7. Czech Statistical Office. *Life Tables*. Life tables for the CR since 1920, 2016. https://www.czso.cz/csu/czso/life_tables. Accessed 17 November 2017.

8. Czech Statistical Office. *Wages - time series*. Table 4 Average gross monthly wages by sphere in 2000 through 2016, 05.09.2017. https://www.czso.cz/csu/czso/pmz_ts. Accessed 2 December 2017.

9. Česká Asociace Pojišťoven. *Investiční Životní Pojištění.* 2014. http://www.cap.cz/vse-o-pojisteni/pojisteni-osob/investicni-zp. Accessed 7 May 2018.

10. Česká Asociace Pojišťoven. *Pojistné Produkty.* 2014. http://www.cap.cz/pojistne-produkty. Accessed 7 May 2018.

11. Fojtík, Jan, et al. Method for Fast Estimation of Life Insurance Liabilities with Respect to Different Investment Strategies. AMSE, 2017.

12. Freedman, Avi. Cluster Analysis: A Spatial Approach to Actuarial Modeling. Milliman, 2008.

13. Grace-Martin, Karen. "What Are Nested Models?" The Analysis Factor, August 2017, https://www.theanalysisfactor.com/what-are-nested-models/. Accessed 7 April 2018.

14. Hennig, Christian, et al. *Handbook of Cluster Analysis*. CRC Press, 2016.

15. Janeček, Martin. Valuation Techniques of Life Insurance Liabilities: Valuation Techniques and Formula Derivation. LAP Lambert Academic Publishing, 2012.

16. Kassambara, Alboukadel. Practical Guide To Cluster Analysis in R. STHDA, 2017.

17. Kim, Y. J., Oh, Y., Park, S., Cho, S., & Park, H. "Stratified Sampling Design Based on Data Mining." Healthcare Informatics Research, 30 September 2013, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3810526/. Accessed 30 March 2018.

18. Lombardi, Louis J. *Valuation of Life Insurance Liabilities.* 4th ed. Actex Publications, 2006.

19. Matoušek, Jan. *Kvalita produkce v životním pojištění*, 2015. https://efpa.cz/files/matousek.0.pdf

20. Perna, Cira and Marilena Sibillo. Mathematical and Statistical Methods for Insurance and Finance. Springer Verlag, 2008.

21. "Pure endowment." *Merriam-webster.com*. Merriam-Webster. Accessed 8 November 2017.

22. Ray, Ajoy Kumar and Tinku Acharya. *Information Technology: Principles and Applications*. PHI Learning, 2004.

23. Rockford, Thomas. "Life Insurance Commissions- How Life Insurance Agents Are Paid.“ LifeAnt, 20 February 2015, http://www.lifeant.com/life-insurance-commissions-how-life-insurance-agents-are-paid/. Accessed 2 December 2017.

24. Romesburg, Charles. *Cluster Analysis for Researchers.* Lulu Press, 2004.

25. Snipes, Michael and D. Christopher Taylor. "Model selection and Akaike Information Criteria: An example from wine ratings and prices." ScienceDirect, 13 December 2013, https://www.sciencedirect.com/journal/wine-economics-and-policy. Accessed 8 April 2018.

26. Thompson, Steven K. *Sampling.* 3rd ed. Wiley, 2012.

27. Whelehan, David D. *International Life Insurance.* Chancellor Publications Limited, 2002.

List of Figures

List of Tables