in co-operation with

Reserving Algorithms and Portfolio Analysis for Life Insurance using Apache Spark.

By: AMIR ANSARI

MSc Media Informatics

Rheinische Friedrich-Wilhelms-Universität Bonn
Institut für Informatik III
Smart Data Analytics

Reserving Algorithms and Portfolio Analysis for Life Insurance using Apache Spark.

MSc. Media Informatics

Submitted by: Amir Ansari
Matriculation Number: 328001
Date of submission: 06^{th} November 2017

First examiner/Erstgutachter: Prof. Dr. Jens Lehmann
Second examiner/Zweitgutachter: Prof. Dr. Jürgen Gall
Supervisor/Betreuer: Dr. Hajira Jabeen & Dr. Sven Ebert

Rheinische Friedrich-Wilhelms-Universität Bonn
Institut für Informatik III
Smart Data Analytics

DECLARATION FOR MASTER-THESIS

I hereby declare and confirm that this thesis is entirely the result of my own original work. Where other sources of information have been used, they are indicated as such and properly acknowledged. I further declare that this or similar work has not been submitted for credit elsewhere.

Bonn, November 06^{th}*, 2017*

Amir Ansari

ACKNOWLEDGEMENTS

I would like to thank Prof. Dr. Jens Lehmann and Prof. Dr. Jürgen Gall for allowing me to do my master thesis under their supervision.

A special thanks to my supervisors Dr. Hajira Jabeen and Dr. Sven Ebert (Actuary) for their useful suggestions and stimulating discussions throughout the master thesis. It would not have been possible without them.

I would like to sincerely thank Mr. Jens Sonnenschein (Director Middle East Life), Mrs. Britta Tausgraf (Team Leader Actuarial & Administration) and Ms. Sophia-Friederike Bremer (Marketing Actuary) at SCOR, Cologne for giving me the opportunity to do my master thesis with SCOR.

Finally, I must express my very profound gratitude to my parents and to my sisters for providing me with unfailing support and continuous encouragement throughout my years of study. Thank you.

'Where shall I begin, please your Majesty' he asked. 'Begin at the beginning,' the King said, gravely, 'and go on till you come to the end then stop.'

-Lewis Carroll

ABSTRACT

For (re-)insurance companies, estimating reserves, known as reserving, is an essential and recurring task. Reserves are money put back by the insurance company to pay future obligations. In general, future obligations are unknown and there are many methods to estimate reserves that have been established in the last decades. Probably the most famous or traditional method is net present value (NPV). Its simplicity and estimation power has made it the standard estimation technique. With the use of NPV, estimation of future claims and hence of reserves can easily and quickly be done with standard software.

We want to introduce the use of advance analytics for doing reserving and go beyond the traditional methods, because data is increasing day by day and standard softwares like Excel is not always capable of handling large data sets. This thesis aims to solve this problem by using clustering algorithms which reduce the data used for reserving in a meaningful way. In the end we have concluded that bisecting K-means clustering algorithm can be used for reserving which will produce good approximation to the traditional methods and would also be capable of handling large data sets. To prove this we have compared the solution of the reserving algorithms which use clustering with the traditional methods on a test data set.

ZUSAMMENFASSUNG

Für (Rück-)Versicherer ist das Schätzen von Reserven eine essentielle und wiederkehrende Aufgaben. Reserven sind Geld, das die Versicherer zurückstellen, um zukünftige Verpflichtungen zu bezahlen. Grundstzlich sind zukünftige Verpflichtungen unbekannt. In den vergangenen Jahrzenten wurden viele Methoden entwickelt, Reserven hier für zu schtzen. Die wahrscheinlich bekannteste und traditionellste Methode ist der ′Net Present Value (NPV)′. Seine Einfachheit und Aussagekraft haben ihn zur Standardschätztechnik gemacht. Mit dem NPV können zukünftige Schäden und daher auch Reserven einfach und schnell mit Standardsoftware geschätzt werden.

Wir wollen die Nutzung von ′Advanced Analytics′ für die Aufgabe der Reservierung einführen und über die traditionellen Methoden hinausgehen, da die Datenmengen täglich zunehmen und Software nicht mehr in der Lage ist, diese Datenmengen zu verarbeiten. Ziel dieser Arbeit ist es dieses Problem durch den Einsatz von Clustering algorithmen zu lösen welche die Datenmenge auf sinnvolle Art und Weise reduzieren. Am Ende haben wir geschlussfolgert, dass der bisecting K-means Clusteringalgorithmus, der für die Reservierung verwendet werden kann und gute Näherungslsungen für die traditionellen Methoden produziert sowie gleichzeitig große Datenmenge verarbeiten kann. Um dies zu beweisen, haben wir die Lösung des Reservierungsalgorithmus, der Clustering verwendet mit den traditionellen Methoden auf einem Testdatensatz vergleichen.

Contents

1 Abbreviations and Definitions

Reinsurance
A transaction in which one party, the "re-insurer", in consideration of a premium paid to it, agrees to indemnify another party, the "re-insured", for part or all of the liability assumed by the re-insured under a policy of insurance that it has issued. The re-insured may also be referred to as the "original" or "primary" insurer or the "ceding company" [1].

Insurance
A contract that exists between two parties, where one party (insurer) agrees to pay to another party (the insured) for loss to a specified subject caused by designated contingencies for example hazards or liability. The term "assurance", commonly used in England, is considered synonymous with "insurance" [29].

Permanent Total Disability (PTD)
The permanent and complete loss of use of both hands, both arms, both feet, both legs, both eyes, or any two such parts, like one leg and one arm or a complete disability that renders the employee permanently unable to do any kind of work for which there is a reasonably stable job market [23].

Net Present Value (NPV)
Net present value is the difference between the present value of cash inflows and the present value of cash outflows. NPV is used in capital budgeting to analyze the profitability of a projected investment or project [2, 25].

Mortality Table
A table that shows the rate of deaths occurring in a defined population during a selected time interval, or survival from birth to any given age [28]. In other words a mortality table is a statistical table that shows the life expectancy of people of each age and consecutively frequent deaths for a given age or occupation [19]. A mortality table is also known as a "life table" or "actuarial table".

Risk-Free Rate of Return
The risk-free rate of return defines the theoretical rate of return of an investment without any risk. The risk-free rate shows the interest of an investor that he/she would expect from an absolutely risk-free investment over a given period of time [9].

Cost of Capital
The cost of capital is simply the return expected by those who provide capital for the business [4, 12].

Cedent
A party to an insurance contract who passes financial obligation for certain potential losses to the insurer. An insurance premium has to be paid by the cedent in return for bearing a particular risk of loss. The term cedent is frequently used in the reinsurance industry, although the term could be applied to any insured party.

Inception Date

The date of inception of the insurance policy refers to the actual date when the insurance policy goes into effect. From the initiation date of the policy until its expiration or cancellation, coverage remains in the effect for the policyholder. Before the date of inception of the insurance policy, the policyholder does not have coverage under the plan [39].

Underwriting

In the insurance world, underwriters determine whether an insurance agency should undertake the risk of insuring a client. They determine the risk and exposure of clients, the needed insurance that should be granted to a client, amount that should be paid for it and whether or not to offer an insurance policy to the client in the first place [21].

Medical Underwriting

Medical underwriting is a health insurance term which refers to the use of medical or health information in the assessment of an applicant for coverage, usually for life or health insurance. As part of the underwriting process, an individual's health information may be used in making two decisions: whether to offer or deny coverage and what premium rate to set for the policy [24]. It is a procedure, wherein an underwriter makes sure of the health conditions of the client who applies for the insurance, keeping in mind certain aspects like age, geographical zone, nature of work and health condition. After looking at all the factors, underwriter suggest whether policy should be given to the person and, if so, what will be the premium [30].

Administrative Expenses

An administrative expense means the expenses related to the everyday operations of a business. Administrative expenses refers to operation expenses that can be directly related to the services such as rent, utilities and salaries for running a business[16]. These expenses are related to the organization as a whole as opposed to an individual department. Salaries of senior executives and costs of general services such as accounting are examples of administrative expenses [27].

2 Overview

In this chapter we will give short introduction about SCOR, explain the reason for choosing this topic of reserving and the motivation behind it. Further, giving a short introduction about the topic and in the end we will outline structure of the thesis.

About SCOR[1]: SCOR is an independent global reinsurance company headquartered in Paris, with an aim to develop its Life and P&C business, to provide its clients with value-added solutions and to pursue an underwriting policy based on profitability, through effective risk management and a cautious investment policy. This way, SCOR provides its clients an optimal level of security (AA− rating from S&P and Fitch and Aa3 rating from Moody's) and generates value for its shareholders. The Group's strategy is based on a development model driven by three entities: the P&C entity, the Life entity and the Asset Management division. I am doing this master thesis in Middle East Life department (Cologne, Germany) which is a part of Life & Health division of the SCOR group.

2.1 Motivation

In single premium credit life insurance, the insurance company gets full premium at the beginning of the contract. This premium has to be reserved in order to pay the claims which might occur over the duration of policy. This process is called reserving.

For example, a life insurance company had one thousand policies issued at the beginning of 2017. Without having any reserve requirements, the company could take all the collected premiums as its revenue by the end of 2017 and then take the profit based on the same revenue. Typically nothing goes wrong at the beginning since the expected frequency of claims is usually small because policies just started. However, as time goes by, the expected frequency of claims usually increases and eventually in two or three years the claims costs are going to exceed the total annual premiums collected in 2017. From then on, the life insurance company would need reserve an amount of cash to offset the extra claims costs and starts to have negative profit [41].

This volatile cycle (huge gains at the beginning and huge losses at the end) is risky and unhealthy for insurance companies. In another words, if all the insurance companies work without reserving, then a question about the functioning of risk management might occur. In reality, insurance companies set aside a part of their collected annual premiums as reserves to pay future claims [41]. If an insurance company does not keep enough reserve, it will face the insolvency crisis when claim payments come at a certain point of time in the near future. Inaccurate reserves, either underestimated or overestimated, will impact the financial health of insurance companies. That's why calculating accurate reserves is an important process of any life insurance company.

For reserving, traditional methods are used by insurance industry which is discussed in detail in this thesis. The results obtained by traditional methods are called "correct solution" or "mathematical correct solution" or "exact mathematical solution".

[1]Retrieved online on September, 29^{th} 2017 from https://www.scor.com/en

In life insurance, treaty duration can be up to 30 years and payment of claims can be very far in future while premium is received today. So insurance companies have to keep a reserve in order to pay the future claims. The reserves required at any time are the resources needed to meet the costs, as they arise, of all claims not finally settled at that time. The insurer must be able to quantify this liability if it is to assess its financial position correctly, both for statutory and for internal purposes [34].

Consolidated balance sheet		
SHAREHOLDERS' EQUITY AND LIABILITIES In EUR million	As of December 31, 2015	As of December 31, 2014
Shareholders' equity - Group share	6,330	5,694
Non-controlling interests	33	35
Financial debt	3,155	2,232
Contingency reserves	300	297
Contract liabilities	27,839	25,839
Other liabilities	3,948	3,309
TOTAL SHAREHOLDERS' EQUITY AND LIABILITIES	41,605	37,406

Table 1: SCOR Annual Financial Statement 2015

As shown in the Table 1, contract liabilities is around 27 billion EUR. Contract liabilities can be understood as amount of reserve SCOR needs to pay all its future claims. While on the other hand SCOR writes a gross premium of around 13 billion EUR (as of 31 December 2015) and 11 billion (as of 31 December 2014). One thing clearly seen here is that the amount of reserve needed is much higher than premium SCOR writes every year. This means that miscalculated reserves cannot be adjusted easily with the premium received during a financial year. This means that accurate reserving is crucial to (re-)insurers to avoid bankruptcy.

For pure life insurance companies, the ratio of reserves to premium can be even higher, e.g. Hannoversche Lebensversicherung AG Balance sheet 2015 as shown below in the Table 2.

Consolidated balance sheet		
SHAREHOLDERS′ EQUITY AND LIABILITIES In EUR million (approx)	2015	2014
Contract liabilities	8,689	8,531
Gross premium written	1,123	1,145

Table 2: Hannoversche Lebensversicherung AG Balance Sheet 2015 [2]

Till date reserving is done via traditional methods. We want to introduce the use of advance analytics in reserving. Since we are already having the results of reserving from traditional methods it will be easy to compare it with original solution. The reason of using analytics in this area is that the amount of data is increasing day by day and it is getting difficult to do reserving via traditional method.

[2]Retrieved online on September, 26^{th} 2017 from https://www.vhv-gruppe.de/grp/files/VHV%20Annual%20Report%202015.pdf

We hope that the results will produce good approximation to the exact/traditional methods which would be able to deal with larger data sets than the traditional methods. The goal of this master thesis is to develop a tool for a single premium credit life insurance - a special type of insurance product - and to investigate reserving pattern through clustering using different algorithms.

2.2 Introduction

As discussed above in section 2.1, we will focus only on single premium credit life insurance. So let us understand the meaning of single premium credit life insurance and its working.

Credit life Insurance: Credit life insurance is a life insurance policy designed to pay off a borrower's debt if borrower dies. The face value of a credit life insurance policy decreases in proportion with the outstanding loan amount as the loan is paid off over time until both reach the value of zero [5].

Life Insurance: Life insurance is a policy that protects against any financial loss that would result from the untimely death of an insured. The beneficiary receives the proceeds and is thereby safeguarded from the financial impact of the death of the insured. The death benefit is paid by a life insurer in consideration for premium payments made by the insured [42].

Single premium credit life Insurance: Single premium credit life insurance is a form of insurance that you purchase when you take out a loan. If a person dies or becomes disabled, the credit life insurance policy makes payments on his/her debt. Single premium credit life insurance is linked with a specific loan and full premium is to be paid at the time of taking loan.

Working: A borrower purchases single premium credit life insurance from the same company; a bank, finance company, credit union, auto dealer that makes the loan. From a consumer's standpoint, single premium credit life insurance provides borrowers and their families with the security of knowing that they will not default on their obligations due to death, disability, or an interruption in their employment. Single premium financing is an affordable way to pay for credit insurance, particularly in connection with long-term home loans [6]. You can usually get single premium credit life insurance for many types of loans, including car loans, student loans, credit cards and mortgages.

2.3 Structure

This thesis is divided into six chapters. Below is detailed description about each chapter.

Chapter 01 is about abbreviations and definitions that one should know before reading this thesis because there are technical and insurance terms frequently used.

Chapter 02 is divided into two parts. First part is about motivation, importance of the chosen topic and its impact on reinsurance companies. Second part is about introduction to credit life insurance, life insurance, single premium credit life insurance and and its method of working.

Chapter 03 is about literature review and background. This chapter is divided into three parts. First part is about the concept and necessity of reserving. Second part is about available methods and techniques used for reserving. This very part also deals with the manner in which we can introduce use of advance analytics as a means of reserving. Last part is about the clustering, its method of working in insurance industry in general and the availability of standard clustering algorithms that could be used to get reserving pattern.

Chapter 04 describes the problem and its solution. This chapter is divided into three parts. First part is about introduction of net present value (NPV) and a short mathematical background about NPV. Second part is about reserving tool in VBA, and in which fashion we can put data into the tool to get the reserving pattern (step by step process). Third part is about the technologies used for clustering with a short description. Last part shows a blueprint which depicts the overall structure of this master's thesis.

Chapter 05 is about results and analysis. In the first part, we formulated a hypothesis that the results after clustering will almost be similar to the exact mathematical solution, obtained from the reserving tool in VBA/Microsoft Excel. Then we have explained it's similarity to exact mathematical solution. In the end we have given an error formula that can be used to calculate the deviation of clustered results from the exact mathematical solution. After hypothesis, we have given general information about the data. The results obtained from exact mathematical solution are compared with K-means and bisecting K-means results. Further, we explained the problems faced while using K-means and bisecting K-means and also their solutions. Then we have ranked the algorithms to show which one performs better based on certain standards in insurance industry. At last we have replicated our data ten times and performed clustering using SparkR to show scalability.

Chapter 06 is future work. This chapter is divided into three parts. First we came up with an algorithm that can be used for reserving, which gives good approximation as compared to exact mathematical solution. As we have only touched upon reserving, second part describes about other fields in insurance where big data can be used. At the end, we have shown the challenges that big data is facing in financial world particularly in insurance.

3 Literature review and background

Reserving is an important process for re-insurers, whether it is life insurance or non life insurance and there are some standard methods to calculate reserves. Before going in detail we should understand the concept of reserving and its importance.

3.1 Reserving

An insurance company should be able to pay for claims whenever they occur. This is something clients expect when they sign a contract with the insurance company and also authorities verify this throughout the year. On the other side, insurance companies probably don′t need to put back enough money to pay for all claims of all clients at the same time, since the probability that this event occurs, is vanishingly small. It is the insurer′s task to calculate the right amount of money to put back to pay those claims which are likely to occur. That has to be ensured on the short term as well as on the long term. This process is called reserving [31].

The idea behind reserving is to estimate future obligations of the insurance company, mainly based upon their claims history. This gives the insurance company an estimate of the amount they will have to pay in the next month, next year and eventually in the next couple of decades. The money, an insurance company has to put back then is called reserve.

Reserving is the term used for ′estimating unknown future payments′. Good estimation techniques for reserves are essential because [31]:

1. Too high reserves are unnecessarily bounded capital. That means the insurance company could use a fraction of capital for other investments with higher returns.
2. Too low reserves can lead to serious problems for an insurance company. If not enough money has been reserved, the insurance company might not be able to pay all the obligations and could become insolvent.

An insurance company has to pay claims whenever they occur, meaning for credit life insurance the outstanding amount at the time of death of the insured. On the other hand, the premium for the insurance is often paid in total as a single premium at the beginning of the loan duration. But in order to be able to pay all the future claims the insurance company must distribute the premium over the duration of the loan, i.e. reserve premium for future claims. To calculate the amount of the (yearly) reserve needed, several mathematical techniques are known (find below) and this process is called reserving [38].

Estimation of reserves for every single contract would be a difficult task. On the other hand, estimation of reserve for the whole portfolio would be too loose. A most common and effective way of reserving would be to analyze the claims history [31, 38].

3.2 Methods & techniques for reserving

Now as we know the importance of reserving and why it is done, now let's have a look at its operation. Till date reserving is done via traditional methods, which can be understood by the following given example:

For non-life insurance companies:

1. In order to estimate reserves, insurance companies make use of their own claims history -if possible. The underlying assumption is that future claims will have a similar pattern as historical claims or at least a functional relation exists between historical and future claims. A way of analyzing historical data is to align them in a triangle. Clearly, it is not the only way, but triangles have been used for a long time and proven to be a useful tool in order to estimate reserves[31].

2. The most famous method is the chain ladder method. Originated as a purely deterministic algorithm, it′s simplicity and estimation power made it the standard estimation technique. With the chain ladder method, estimation of ultimate losses and hence of reserves can be easily and quickly done with standard computer software[31].

   The chain ladder method is based on an assumption that the expectations underlying the columns and the rows in the run off triangle are proportional. In practice it seems rational to use the chain-ladder method to estimate the outstanding claims reserve if the data is consistent with the model. The chain ladder method uses cumulative data and assumes the existence of a set of development factors[37]:

$$\left\{ f_j | j = 2, ..., n \right\} \text{,With} \tag{1}$$

$$E[C_{i,k+1}|C_{i1}, ...C_{in}] = C_{ik} f_{k+1}, \quad 1 \leq i \leq n, 1 \leq k \leq n-1 \tag{2}$$

   These factors are estimated by the chain-ladder method as:

$$\hat{f}_j = \frac{\sum_{i=1}^{n-j+1} C_{ij}}{\sum_{i=1}^{n-j+1} C_{i,j-1}}, \qquad 2 \leq j \leq n \tag{3}$$

   To forecast future values of cumulative claims, these factors are applied to the latest cumulative claim in each row:

$$\hat{C}_{i,n-i+2} = C_{i,n-i+1} \hat{f}_{n-i+2}, \qquad 2 \leq i \leq n \tag{4}$$

$$\hat{C}_{i,k} = \hat{C}_{i,k-1} \hat{f}_k, \qquad 2 \leq i \leq n, \quad n-i+3 \leq k \leq n \tag{5}$$

For life insurance companies:

1. Net preset value (NPV) is the present value of the cash flows at the required rate of return of your project compared to your initial investment. In other words, with the help of NPV one can calculate return on investment, or ROI, for a project or expenditure. One can decide the outcome of a project by calculating the expected returns on the investment and converting those returns into Euros[18]. We have discussed in detail about NPV and in particular the tactics behind the estimation of underlying cash flows in section 4.1.

Since we have only traditional or well known methods that can be used for reserving. And all these traditional methods are implemented via some software, example, an NPV function in Excel can make the process effortless after entering the stream of cost and benefits. Many of the financial calculators also include and utilize the benefit of NPV function.

As we know, data is increasing day by day and most of these softwares are not capable of handling large data sets. For example we have an portfolio of 55,000 insureds and it is very difficult to calculates reserves by using standard softwares. That's why we want to introduce the use of advance analytics as a means of reserving with the help of clustering.

3.3 Clustering algorithms

As we have discussed that we will use clustering algorithms for reserving. We would like to see the influence of clustering on reserving pattern and to investigate reserving pattern through clustering using different algorithms.

Clustering is an important step in the process of data analysis with applications to numerous fields. The goal is to partition the data objects into a set of clusters such that objects in the same group are similar, while objects in different groups are dissimilar.

In the cluster analysis, the goal is to identify new groups in the data. These new groups can then be used for many purposes, for example, in automated classification, recognition of patterns (reserving pattern in our case), image processing, market segmentation or in any other methods that rely on such knowledge [14].

Following method is well known in insurance industry:

1. Group all policies with the help of predefined age bands.
2. Split these groups on predefined duration bands.
3. Choose representatives for each group by taking averages on age and duration.

Example 10 tripples of age, duration and sum insured (age, duration, SI).

$$(25, 8, 10), (37, 4, 10), (44, 8, 10), (29, 5, 10), (39, 3, 20)$$
$$(20, 7, 20), (41, 9, 20), (23, 3, 20), (25, 6, 10), (35, 14, 20)$$

The predefined age bands are assumed to be 18-35 and 36-45. The predefined duration bands are 0-10 and 11-20.

With the procedure stated above we get the total sum insured per age and duration band.

Age band / Duration band	0-10	11-20
18-35	70	20
36-45	60	0

Table 3

This yields 4 groups with the following representatives:

$$(26.5, 5, 70), (26.5, 15.5, 20), (40.5, 5, 60), (40.5, 15.5, 0)$$

.

Apart from this naive clustering example discussed above, in general there are two types of clustering algorithms that we will implement to get the reserving pattern. Now we will discuss about partitional clustering and hierarchical clustering algorithms:

3.3.1 Partitional clustering algorithms:

These state the construction of various partitions and then evaluate them by some criterion. Perhaps the most popular class of clustering algorithms is the combinatorial optimization algorithms a.k.a. iterative relocation algorithms.

These algorithms minimize a given clustering criterion by iteratively relocating data points between clusters until a (locally) optimal partition is attained. In a basic iterative algorithm, such as K-means- or K-medoids, convergence is local and the globally optimal solution can not be guaranteed.

Algorithm K-means:

K-means is one of the simplest unsupervised learning algorithms which aims at solving the well known clustering problem. The main idea is to define K centers, one for each cluster. These centers should be placed in a cunning way because different location give different result [15]. So, the better choice is to place them as far as possible, from each other.

The next step involves picking out each point that belongs to a given data set and merge it with the nearest center. When no point is pending, the first step is completed and an early group age is done. At this stage there is a need for re-calculating K new centroids as barycenter of the clusters formed from the previous step. After we have these K new centroids, a new binding has to be done between the same data set points and the nearest new center.

Now, a loop will be generated. After the formation of this loop, we may observe that the K centers shift to other location step by step until no more changes are made. Ultimately, this algorithm targets to minimize an objective function known as ”squared error” function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (||x_i - v_j||)^2 \qquad \text{where,} \tag{6}$$

$||x_i - v_j||$ is the euclidean distance between x_i and v_j
c_i is the number of data points in i^{th} cluster.
$'c'$ is the number of cluster centers.

Working of K-means is shown by the following example, below in the Figure 1:

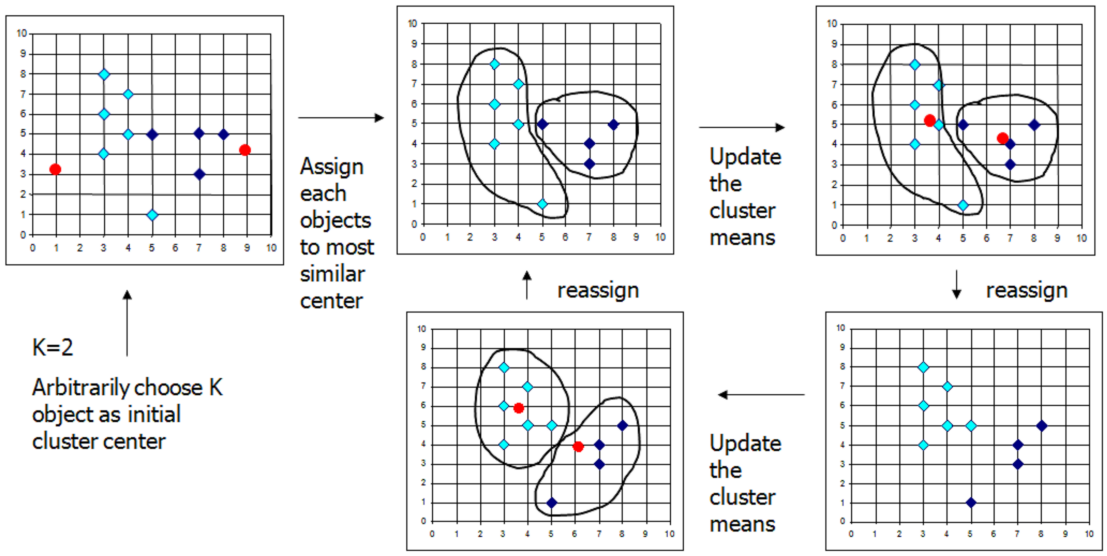

Figure 1: K-means Example [17]

As can be seen from the figure above, we start with a definite number for the number of required cluster (in this case k=2). The algorithm takes 2 random seeds and maps all other data points to these two seeds. The algorithm re-iterates till the overall penalty term is minimized.

3.3.2 Hierarchical clustering algorithms:

Hierarchical techniques produce a nested sequence of partitions, with a single, all inclusive cluster at the top and singleton clusters of individual points at the bottom. Each intermediate level can be estimated by the combination of two clusters from the next lower level (or cleaving a cluster from the next higher level). The result of a hierarchical clustering algorithm can be graphically displayed as tree, called a dendogram as shown below in the Figure 2. This tree graphically displays the merging process and the intermediate clusters [33].

There are two basic approaches to generate a hierarchical clustering:

Agglomerative:

Start with the points as individual clusters and, at each step, merge the most similar or closest pair of clusters. This requires a definition of cluster similarity or distance.

Divisive:

Start with one, all-inclusive cluster and, at each step, split a cluster until only singleton clusters of individual points remain. In this case, we need to decide, at each step, to split which cluster and in which manner.

An example of hierarchical clustering is shown below in the Figure 2:

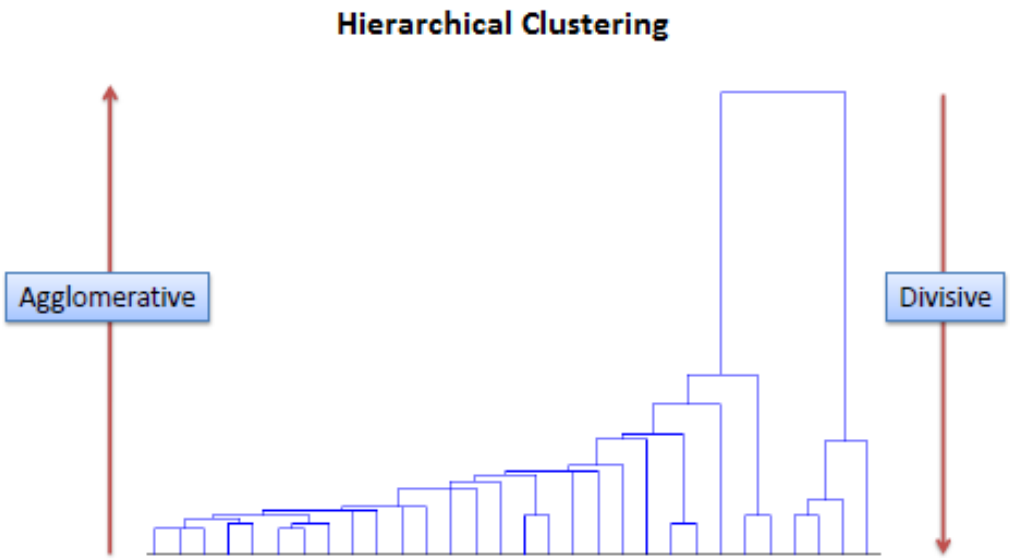

Figure 2: Clustering [8]

Bisecting K-Means:

Bisecting K-Means is the amalgamation of K-Means and hierarchical clustering. It starts with all objects in a single cluster. In strict sense, the bisecting K-means algorithm is a divisive hierarchical clustering algorithm. Bisecting K-means algorithm for finding K clusters [33].

1. Pick a cluster to split.
2. Find 2 sub-clusters using the basic K-means algorithm. This step is also called bisecting step.
3. Repeat above step, the bisecting step, for ITER times and take the split that produces the clustering with the highest overall similarity.
4. Repeat the first three steps until the expected number of clusters are achieved.

It has been noted that bisecting K-means is ahead and sometimes on par with hierarchical clustering in terms of clustering quality metrics such as F-measure and entropy [33]:

1. K-means is not so bad as many expect and sometimes beat hierarchical clustering.
2. Bisecting K-means is much faster comparing to hierarchical clustering - $O(n)$ vs $O(n^2)$.

I believe, based on such good results, bisecting K-means was obvious choice for implementation in famous Apache Spark framework instead of hierarchical clustering.

4 Methodology

Credit life insurance is becoming popular in world and so is the risk associated with it. It is a powerful tool for all the parties involved in it. For the loan providers, it assures proper recovery of loan within specified amount of time in case of unfortunate death of borrower. Also it assists the family of the borrower to pay loan amount within the given period of time without putting financial pressure on them [11].

The main role played in this whole process is by the insurance industry. As insurance company ensures the smooth recovery of loan in case if borrower dies, insurance company has to keep reserve to make sure that they can pay to loan provider if any unwanted condition appears. In other words, they have to manage the risk associated with it till the whole amount of loan is paid. The aim is to solve above mentioned problem as a part of master thesis by developing a tool which ensures risk adequate reserving and at the same time check profitability during the contract duration [32].

4.1 Mathematical background of net present value (NPV)

There are several methods that can be used for reserving, as discussed in section 3.2. The method that we will use in our reserving tool is based on the Net Present Value (NPV).

In general the Net Present Value (NPV) is defined as the difference between the present value of cash inflows and the present value of cash outflows [26]. For our purpose of reserving, we concentrate on the cash outflows, i.e. the expected losses of a policy. In single premium insurance, the insurance company should charge this amount plus a cost margin as premium.

First, we will give an introduction of net present value (NPV) starting with an example:

Suppose the death benefit is payable at the end of the year of death. Let $G > 0$ (the "age at death") be the random variable that models the age. $T(G, x) := ceiling(G - x)$ is the number of "whole years" (rounded upwards) lived by (x) beyond age x [25].

Suppose that there is a 90% chance of an individual of surviving in any given year (i.e. T has a geometric distribution with parameter p = 0.9, i.e. a one year survival period of 90% independent of the age x and the set 1, 2, 3 ... for its support). Then the probability P_r of dying in the 1^{st}, 2^{nd} and 3^{rd} year is:

$$P_r[T(G, x) = 1] = 0.1 \tag{7}$$

$$P_r[T(G, x) = 2] = 0.9(0.1) = 0.09 \tag{8}$$

$$P_r[T(G, x) = 3] = 0.9^2(0.1) = 0.081 \tag{9}$$

To determine the net present value of the benefit, we need to calculate the expected value $E(Z)$ of the random variable $Z = Losses$. And at interest rate 6% the net present value of one unit of the three year term insurance is:

$$P(T = 1) * (1.06)^{-1} + P(T = 2) * (1.06)^{-2} + P(T = 3) * (1.06)^{-3}, \tag{10}$$

which is just the expected claim amount per year discounted with 6%.

$$A_{\frac{1}{x}:3} = 0.1(1.06)^{-1} + 0.09(1.06)^{-2} + 0.081(1.06)^{-3} = 0.24244846 \tag{11}$$

So the net present value of the $100,000 insurance is $24,244.85. Now we use the net present value for reserving in the following manner: As seen above we need a total premium of $24,244.85 to be able to pay the expected future claims for this policy.

Assume, we get this premium at the beginning of the insurance period as a single premium, we now have to distribute it over the duration of the policy i.e. for 3 years for above mentioned example. This can be done again with the NPV method:

For 1^{st} year: The net present value of the $100,000 insurance is $24,244.85.

For 2^{nd} year:

$$A_{\frac{1}{x}:3} = 0.09(1.06)^{-1} + 0.0081(1.06^{-2}) = 0.156995372 \tag{12}$$

So the remaining net present value of the expected claim amount after the first year of the $100,000 insurance is $15,699.53. Therefore this amount is equal to the reserve which has to be kept for future claims. The difference between $A_{\frac{1}{x}:3}$ and $A_{\frac{1}{x+1}:2}$ is the so called earned premium.

For 3^{rd} year:

$$A_{\frac{1}{x+1}:2} = 0.081(1.06)^{-1} = 0.076415094 \tag{13}$$

So the net present value after the second year of the \$100,000 insurance is \$7,641.15.

A n year term life insurance of a x year old insured with a benefit of \$100,000 payable at the end of the year in case of death has net present value

$$100,000A_{\frac{1}{x}:n} = 100,000\sum_{t=1}^{n} v^t Pr[T(G,x) = t] \tag{14}$$

Now we would like to give the general formula for the net present value (NPV) for our purpose of credit life insurance. Net present value (NPV) of one unit of insurance is given by :

$$\begin{aligned}
A_x &= E(Z) && (15)\\
&= E(V^T) && (16)\\
&= \sum_{t=1}^{\infty} v^t Pr(T = t) && (17)\\
&= \sum_{t=0}^{\infty} v^{t+1} Pr[T(G,x) = t+1] && (18)\\
&= \sum_{t=0}^{\infty} v^{t+1} Pr[t < G - xt + 1 | G > x] && (19)\\
&= \sum_{t=0}^{\infty} v^{t+1} (\frac{Pr[G > x+t]}{Pr[G > x]})(\frac{Pr[x+t < G \leq x+t+1]}{Pr[G > x+t]}) && (20)\\
&= \sum_{t=0}^{\infty} v^{t+1} {}_tp_x \cdot q_{x+t} && (21)
\end{aligned}$$

Where ${}_tp_x$ is the probability that (x) survives to age $x+t$, and q_{x+t} is the probability that $(x+t)$ dies within one year.

4.2 Reserving tool in VBA

Now we will explain the tool, which was newly developed and used for single premium credit life business. Below we describe the tool worksheet by worksheet.

In general, all petrol blue fields have to be filled. All other fields are non-editable. It is recommended to go through the tool step by step and follow instructions to achieve accurate reserving.

4.2.1 General information

First we have general information sheet in tool. This sheet is used to enter general information like company's name (cedent), policy name, risk free rate, type of cover, percentage of mortality table for death any cause, PTD, underwriting effect 1^{st} and 2^{nd} year, additive mortality in per mile and product (specified, if available).

1. Cedent name have to be selected from a drop down menu, market country will be selected automatically as soon as cedent is selected and policy name have to be entered manually as shown in figure below:

Policy and group information

Cedent	TRUST INTL, BAHRAIN
Market country	Bahrain
Policy name	Please enter the policy name

Figure 3: Screen shot of the sheet "General Information"

2. Only risk free rate as mentioned in section 4.1 is to be filled if known, otherwise it should be left blank.

Brokerage	0.00%
RI Commission	0.00%
RI Tax	0.00%
Risk free rate/internal rate of return	0.00%

Figure 4: Screen shot of the sheet "General Information"

3. Type of cover should be selected from drop down menu, first inception year will be selected automatically as soon as next sheet (census) is filled. This determines the values of the q_x as described in section 4.1 .

Type of cover	Death any cause
First Inception year	2007

Figure 5: Screen shot of the sheet "General Information"

4. Only petrol blue fields have to be filled, if known, otherwise leave it. As we can see below in the Figure 6 that most of the values are in percentage because there is an underlying base table of which a percentage is taken.

Percentage of Mortality table for Death any cause	100.00%
PTD	20.00%
Underwriting effect 1st year	20.00%
Underwriting effect 2nd year	10.00%
Additive Mortality in per mile	0.00
Cost of capital & AE	0.00%
Admin Expenses	0.00%

Figure 6: Screen shot of the sheet "General Information"

5. These are optional fields, should be filled only if known.

Optional

Product	Credit Life
Product (specified, if available)	
Currency	USD

Figure 7: Screen shot of the sheet "General Information"

4.2.2 Input

On this sheet all parameters per single policy like age, sum insured, interest rate of loan and duration and gender are entered which have an impact on reserving and can vary for each an every loan.

This sheet known as ”Credit Life Census Sheet” is used to enter exposure data such as gender, date of birth, year, duration of loan in months, type of loan, interest rate of loan, rate in per mill per month and sum insured. Some of the fields mentioned below in the figure 8 are mandatory for reserving others can be left empty if not known.

The mandatory fields are as follows:

1. Gender (enter F for female and M for male)
2. Date of birth (DOB) or age (Date of birth as DD.MM.YYYY or two digits age)
3. Duration of loan (in months only)
4. Year (YYYY)
5. Interest rate of loan (in percentage)
6. Total sum Insured (in digits only)

You can click on ’Clear contents’ button any time as shown below in the Figure 8, to re-enter the data (For e.g. if previously entered data was incorrect). Once the required data is filled, you can click ’Next step’ button to proceed.

Clear contents

Next step

Key	Enter M or F Gender	Date of Birth (DD.MM.YYYY)	Age	Techn. Age	Year	Duration of Loan in months	Type of Loan	Interest Rate of Loan	Rate in per mille per month	Sum Insured	Initial Sum Insured
6240679167	M		28	28	2016	139	Personal	5.4%		6,240,679	6,240,679
153276236133	F		47	47	2015	86	Personal	3.4%		153,276,236	153,276,236
1149981668	M		29	29	2016	39	Personal	3.6%		11,499,816	11,499,816
3610904282	M		41	41	2012	241	Mortgage	5.4%		3,610,904	3,610,904
15243394308	M		45	45	2011	263	Mortgage	5.4%		15,243,394	15,243,394
6613516163	M		24	24	2016	139	Personal	5.4%		6,613,516	6,613,516
8300469200	F		42	42	2016	158	Personal	5.4%		8,300,469	8,300,469
19015315233	M		35	35	2016	198	Personal	3.6%		19,015,315	19,015,315
6563314203	M		42	42	2016	161	Personal	3.6%		6,563,314	6,563,314
1327954115	M		44	44	2017	71	Personal	5.4%		1,327,954	1,327,954
13003251207	M		27	27	2017	180	Personal	3.6%		13,003,251	13,003,251
1232030278	M		54	54	2017	224	Personal	5.4%		1,232,030	1,232,030
15664211304	M		32	32	2014	272	Personal	3.6%		15,664,211	15,664,211
42618366309	F		44	44	2011	265	Personal	3.7%		42,618,366	42,618,366
1142910307	F		39	39	2011	268	Personal	5.4%		1,142,910	1,142,910
20413470220	F		50	50	2012	170	Personal	5.4%		20,413,470	20,413,470
7897512306	M		54	54	2014	252	Personal	5.4%		7,897,512	7,897,512
4013335207	M		44	44	2013	163	Personal	5.4%		4,013,335	4,013,335
161458780	M		31	31	2016	49	Personal	5.4%		1,614,587	1,614,587
3107315215	M		57	57	2016	158	Personal	5.4%		3,107,315	3,107,315
289716264	M		44	44	2016	220	Personal	3.6%		289,716	289,716
231819152	M		32	32	2016	120	Personal	3.6%		231,819	231,819
3437282207	M		31	31	2016	176	Personal	3.6%		3,437,282	3,437,282
2817161228	M		45	45	2016	183	Personal	3.6%		2,817,161	2,817,161
43762246	M		21	21	2016	225	Personal	3.6%		43,762	43,762
23679302	M		22	22	2016	280	Personal	3.6%		23,679	23,679
3262772153	M		55	55	2016	98	Personal	3.6%		3,262,772	3,262,772
235644397	M		22	22	2016	75	Personal	3.6%		2,356,443	2,356,443
2637392241	M		21	21	2013	220	Mortgage	3.6%		2,637,392	2,637,392
4653732286	M		32	32	2014	254	Mortgage	5.4%		4,653,732	4,653,732
363637323	M		33	33	2016	290	Mortgage	5.4%		363,637	363,637
236373223	F		43	43	2013	180	Mortgage	5.4%		236,373	236,373
3637485156	F		36	36	2015	120	Mortgage	5.4%		3,637,485	3,637,485
2726272129	M		19	19	2015	110	Mortgage	5.4%		2,726,272	2,726,272
32672216	M		29	29	2015	187	Mortgage	5.4%		32,672	32,672
3673632202	M		37	37	2015	165	Mortgage	5.4%		3,673,632	3,673,632
565872208	M		32	32	2015	176	Personal	5.4%		565,872	565,872
526272214	M		27	27	2015	187	Personal	5.4%		526,272	526,272
326272233	F		43	43	2016	190	Personal	5.4%		326,272	326,272
294848223	M		23	23	2014	200	Personal	5.4%		294,848	294,848
2363783153	M		43	43	2012	110	Personal	5.4%		2,363,783	2,363,783
328282253	M		43	43	2013	210	Personal	5.4%		328,282	328,282
373738249	M		29	29	2016	220	Personal	5.4%		373,738	373,738
3224774207	M		22	22	2016	185	Personal	5.4%		3,224,774	3,224,774

Figure 8: Screen shot of the sheet ”Census”

4.2.3 Outstanding loan amount

Outstanding amount is the principal amount of loan plus interest and other charges minus already paid installments on the loan due from the borrower as on a particular month as shown below in the figure 9.

$$OL_t = TLB + TI_t - I_t + C, \text{ with} \tag{22}$$

OL_t is the outstanding loan amount after t months.
TLB is total loan disbursed
TI_t is total interest after t months also known as initial loan amount.
I_t is installments or amount repaid after t months.
C is charges if any.

The tool calculates the outstanding loan amount per month on this sheet. For this standard interest rate formulas are used.

Months	0	1	2	3	4	5
Outstanding Loan Balance						
	525,250	523,404	521,548	519,684	517,811	515,929
	233,900	232,995	232,086	231,171	230,251	229,326
	1,087,583	1,085,387	1,083,182	1,080,970	1,078,749	1,076,519
	199,850	199,335	198,817	198,297	197,774	197,249
	2,646,000	2,635,595	2,625,144	2,614,646	2,604,101	2,593,508

Figure 9: Screen shot of the sheet "Outstanding Loan Amount"

4.2.4 Expected claim for loan

To calculate the expected claim amount per single policy as described in section 4.1, the following formula is programmed on this sheet. Due to the performance reasons, we switch from months to years on this sheet onwards, as shown in Figure 10. Expected claim amount for loan in the year t can be defined as follows:

$$EC_t = OLT_{12t} * q_x * p_x * (1+i)^t, \text{ with} \tag{23}$$

EC_t is expected claim in year t
q_x is probability that the borrower dies in year t.

p_x is probability that borrower has survived $(t-1)$ years.
i is the internal rate of return

Years	0	1	2	3	4	5
Expected claim for Loan						
	68	138	140	130	121	115
	111	262	303	314	324	331
	133	306	348	358	370	384
	341	832	997	1,070	1,143	1,215
	380	857	908	838	750	657

Figure 10: Screen shot of the sheet "Expected claim for Loan"

4.2.5 Reserves

Reserves is basically sum of all future claims. For example there is a 8 year policy. After 3rd year reserves will be expected claim amount after 3rd year till 8th year and is given by the formula.

$$R_t = \sum_{i=t}^{40} EC_i,\ t \geq 0, \text{with} \tag{24}$$

R_t is reserves after t year.

Here, the tool calculates the reserves per year on this sheet as shown below in the Figure 11. For this, we use net present value (NPV) formulas as described in section 4.1.

Years	0	1	2	3	4	5
Reserves						
	1,387	1,319	1,181	1,041	911	789
	3,588	3,477	3,215	2,911	2,597	2,273
	11,529	11,395	11,090	10,742	10,384	10,013
	20,622	20,281	19,449	18,452	17,382	16,239
	6,969	6,589	5,731	4,824	3,986	3,236

Figure 11: Screen shot of the sheet "Reserves"

4.2.6 Output

A reserving pattern to ensure risk adequate reserving and to check profitability during contract duration, per single policy and per portfolio (sum of all single policies).

Therefore, the above mentioned example in section 4.1 means:

Input: age (x), sum insured \$100,000, interest rate (6%), duration 3 years output will be reserving pattern for 1^{st} year is \$24,244.85, for 2^{nd} year is \$15,699.53 and for 3^{rd} year is \$7,641.15.

The above mentioned example is for a single policy only. For portfolio consisting of more than one policy, the mathematical correct reserving pattern is the sum of the reserves of the single policy.

This states, if portfolio consist of only two policies:

$$\text{Reserving pattern}(P1, P2) = \text{Reserving pattern}(P1) + \text{Reserving pattern}(P2) \quad \text{where,} \tag{25}$$

$P1$ is policy of insured number one.
$P2$ is policy of insured number two.
And reserving pattern is the algorithm giving a mathematical correct reserving pattern for a given set of policies.

Basically the reserving algorithm is additive i.e. reserving pattern obtained from one data set ($P1$) and reserving pattern obtained from another data set ($P2$) can be added together and final reserving pattern will look something like as shown below in the Figure 12.

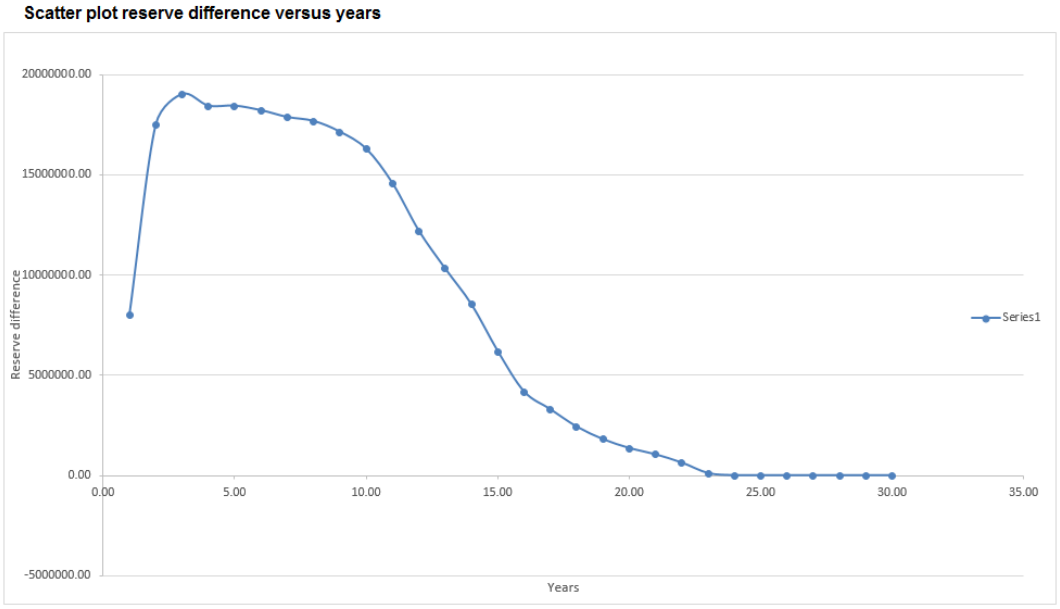

Figure 12: Screen shot of the sheet "Analysis"

4.2.7 Reserving pattern

Here, Figure 13 shows a general form of reserving pattern (not related to above mentioned example in section 4.2.6). We have explained in detail about this graph in section 5.3.

Here,
x axis denotes years which means years of insurance.
y axis denotes reserves difference which is earned premium.

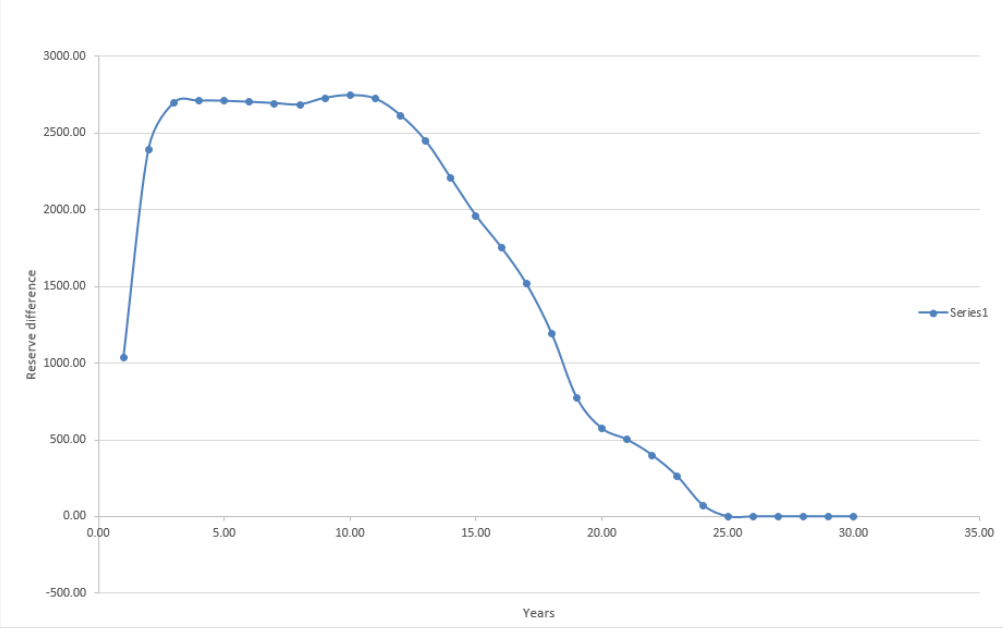

Figure 13: Screen shot of Reserving Pattern

4.2.8 Technologies

We have seen that computer programming is an art, because it applies accumulated knowledge to the world, because it requires skill and ingenuity, and especially because it produces objects of beauty. -Donald E. Knuth [20]

Following technologies have been utilized for the purpose of development, implementation and integration of the elements and components.

VBA

VBA is a powerful scripting language and is one of the most common way to create programs based on Microsoft Office applications (Excel, Word, Access, etc) [40]. We have

used VBA/Microsoft Excel for development of basic reserving tool that we have discussed in section 4.2.

R

R is a free software for statistical computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis [36].

RStudio

We have used RStudio to use R because it provides an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management [35]. RStudio is available in open source and commercial editions and runs on the desktop (Windows, Mac, and Linux). We have used RStudio on Windows 10.

Apache Spark

Apache Spark is a fast and general purpose cluster computing system. It provides high-level APIs in R and other languages, and an optimized engine that supports general execution graphs [22]. Below Figure 14 shows apache spark ecosystem which empowers the spark functionality.

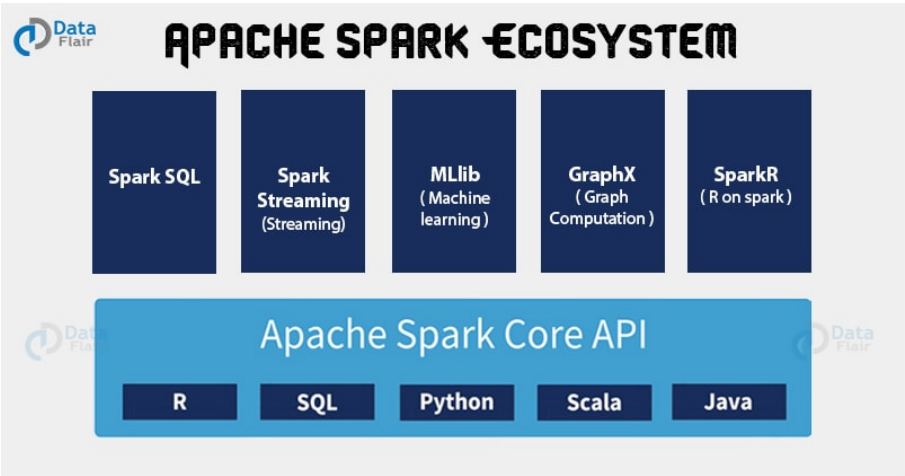

Figure 14: Apache spark components[3]

[3]Retrieved online on August, 08^{th} 2017 from http://data-flair.training/blogs/what-is-apache-spark/

Advantages and few reasons for using apache spark are:

1. Speed
2. Re-usability
3. Fault tolerance
4. High-level Analytics
5. Supports Many Languages

SparkR

SparkR is an R package that provides a light-weight frontend to use Apache Spark from R. In Spark, SparkR provides a distributed data frame implementation that supports operations like selection, filtering, aggregation etc. but on large datasets.

A SparkDataFrame is a distributed collection of data organized into named columns. It is conceptually equivalent to a data frame in R, but with richer optimizations under the hood. SparkDataFrames can be constructed from a wide array of sources such as: structured data files, tables in Hive, external databases, or existing local R data frames [22].

We have used SparkR for clustering and data analysis. In the next steps we will see the use of SparkR from RStudio.

Starting Up: SparkSession

The starting point in SparkR is the SparkSession which connects an R program to a Spark cluster. You can create a SparkSession using sparkR.session and pass in options such as the application name, any spark packages depended on, etc.

sparkR.session()

Starting Up from RStudio

We connected R program to a Spark cluster from RStudio as shown below. To start, make sure SPARK_HOME and JAVA_HOME is set in the environment (you can check whether envoirnment is set properly or not using ”Sys.getenv” as shown below), load the rJava and SparkR package, and call sparkR.session as shown below. It will check for the Spark installation. Alternatively, you can also run install.spark manually.

We have used SparkR in RStudio with the help of the steps shown below [35].

```
Sys.setenv(SPARK_HOME = "C:/Users/amir/Desktop/spark-2.2.0
-bin-hadoop2.7")

Sys.setenv(HADOOP_HOME = "C:/winutils")

.libPaths(c(file.path(Sys.getenv("SPARK_HOME"), "R", "lib"),
.libPaths()))

library(rJava)

library(SparkR)

Sys.setenv('SPARKR_SUBMIT_ARGS'='"--packages""com.databricks
:spark-csv_2.10:1.2.0""sparkr-shell"')

Sys.getenv("JAVA_HOME")

sc <- sparkR.session(master="local",sparkHome=SPARK_HOME)
```

After this, we did clustering on data set shown in section 5.2 and results obtained are discussed in section 5.4 and 5.5.

4.2.9 Linear Interpolation:

Due to best practice standards in accounting, founded e.g. on the "Vorsichtsprinzip" as described in the German trade law "Handelsgesetzbuch" in §252 insurance companies have an obligation to keep the reserves until there is a risk [43]. In other words insurance company must keep the reserves till the treaty expires. A situation may arise after clustering where reserves are consumed before the risk ends because we don't know which cluster representative we will get after clustering. In that case we will do linear interpolation.

Here, is an example to help us better understand the need of linear interpolation: The maximum duration of a single policy is 15 years but cluster representative has a duration of 12 years only.

Linear interpolation is a method of curve fitting using linear polynomials. Linear interpolation works by effectively drawing a straight line between two neighboring points and returning the appropriate point along that line [3].

A Linear Interpolate function calculates an output value y, for the input value x using linear interpolation of the input values $x_0,\ x_1$ (nearest input values) and the output values y_0 and y_1(nearest output values) as shown below in the Figure 15.

$$y = y_0 + (x - x_0) * (\frac{y_1 - y_0}{x_1 - x_0}) \tag{26}$$

where x_0, x_1 are nearest values of input x and y_0, y_1 are nearest values to output y.

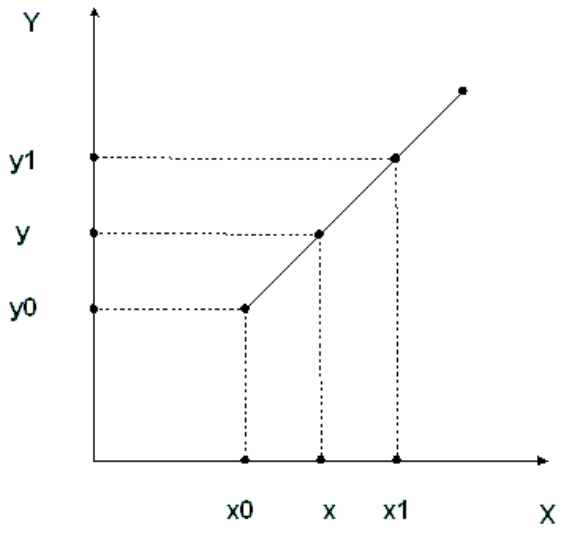

Figure 15: Linear Interpolation [3]

4.3 Blueprint

Below in the Figure 16, a flow chart shows working of the whole process:

1. Raw data is given as input.
2. If data points is less than 5000, we will get reserving pattern using reserving tool in VBA.
3. If it is more than 5000 data points then, we will first do clustering and thereafter get the reserving pattern using reserving tool in VBA as discussed in section 4.2.

4. Once we get the reserving pattern, in the next step, we will check if maximum clustered duration is equal to the maximum original duration. If it is equal, then we will stop and get reserving pattern as output, otherwise will do linear interpolation and get reserving pattern subsequently.

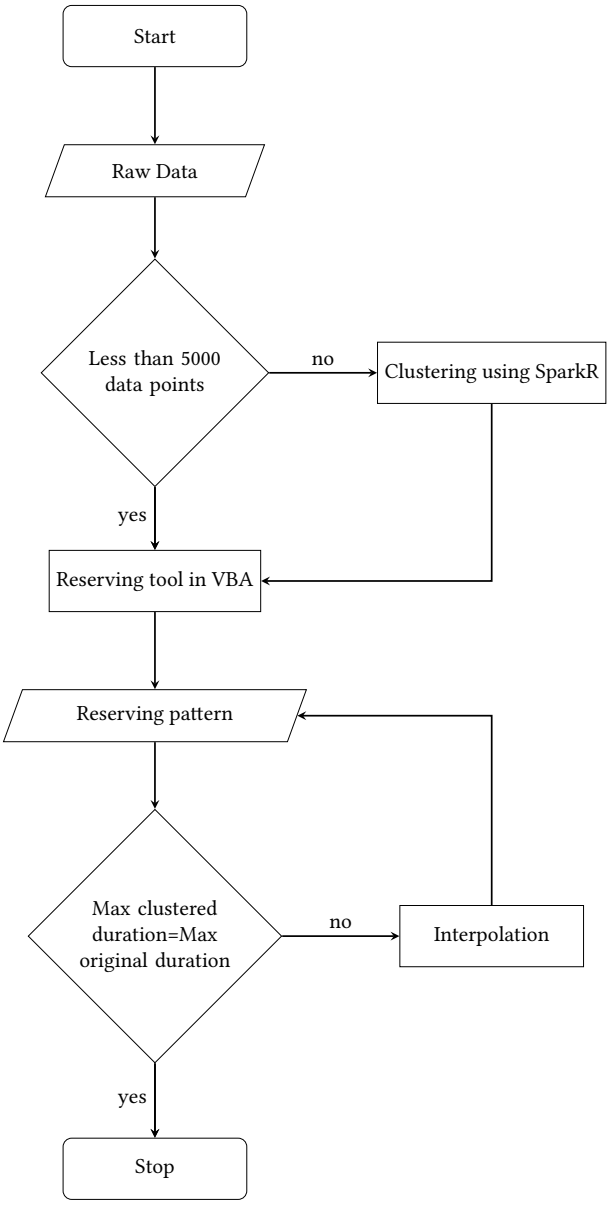

Figure 16: Sequence Flow

5 Results and Analysis

"The goal is to turn data into information, and information into insight."

- Carly Fiorina

5.1 Hypothesis

"The projections of the reserve based on clustered sample data will still be exact enough compared to the projections done on the single policies to be used for calculating the reserving patterns."

We have tested the above hypothesis on a data set for single premium credit life business. The term exact enough has been given a more precise meaning in this chapter.

We have compared the result of the reserve based on raw data i.e. mathematical correct solution (since our data set is still so small that it can be done) and four aggregated methods which operate on clustered data. In the clustering approach insureds are treated which are similar regarding a combination of several attributes, e.g. have similar age and duration and same gender as one insured.

In our case the original data set contains about 55,000 insureds. These insureds have been clustered by the algorithms into at most 20 clusters[4], with one representative for each cluster. We have shown that with 20 clusters, i.e. 20 lines of data instead of 55000, we will still get approximations that can be used for reserving. This is just 0.05 per cent of the original number of lines.

The projection/reserve calculation itself is done on the (aggregated) data by the NPV method as described in section 4.1 above using our standard Excel/VBA tool.

The error should be measured by the sum of overall years of duration where the summand is the squared difference between the reserve calculated on unclustered and clustered data, i.e, the most common goodness of fit measure is 'mean square error' [13]. Here the error is measured by the sum of over all years:

[4]With the help of scree plot we came up with 20 clusters. A Scree Plot is a simple line segment plot that shows the fraction of total variance in the data.It is a plot, in descending order of magnitude, of the eigenvalues of a correlation matrix. In the context of factor analysis or principal components analysis, a scree plot helps the analyst visualize the relative importance of the factors, a sharp drop in the plot signals that subsequent factors are ignorable.[7]

$$Error = 10000 \cdot \sum_{k=1}^{30} (R_k - R_k^{sp})^2, \tag{27}$$

where R_k is the reserve in year k after usage of a clustering, R_k^{sp} is the reserve in year k derived on single policy basis. "Exact enough" as defined above is rather vague that is why we have given certain criteria according to insurance industry which is exact enough, as we advance in this chapter. This measure gives us a first approach to quantify the rather vague expression "exact enough".

5.2 Statistical analytics of the data

According to the Oxford dictionary, analytics is defined as "the systematic computational analysis of data or statistics"[5] or a better definition, also according to the Oxford dictionary: "information resulting from the systematic analysis of data or statistics".

We would like to begin by introducing about the data used for clustering.. We have used a randomly generated portfolio[6] where we choose parameters (like age, sum insured, duration of loan etc.) in a way that it could be a real portfolio. The data set contains 55,000 data points. One data point resembles one single policy i.e. a vector containing all information about the insureds and the loan.

Below in the figures from 17 to 21, we have shown the relation between the different attributes of data points. From a statistical point of view, the following points seem noteworthy:

1. In the Figure 17, we can see that number of insureds are between 18-60 years of age, which are the standard boundaries for such type of insurance.
2. Maximum sum insured is in between 35 to 45 years of age.
3. Maximum duration of loan is 300 months as shown below in the Figure 19.
4. As we can see a slight gap in the Figure 18 around 180-190 months which is due to the fact that we basically simulated two types of loans namely personal loans which have a maximum duration of 15 years and mortgages of duration 25 years and merged them.
5. In general the higher the sum insured, the lesser the number of issued loans.

[5] Retrieved from the online version of Oxford dictionary on July, 29^{th} 2017 from http:// oxforddictionaries.com/definition/analytics?q=analytics

[6] Out of confidentiality reasons we cannot present actual data from the clients of SCOR

6. There are a lot of insureds of 18 years of age.
7. The duration of loan is relatively less for insureds of 50 to 60 years of age as compared to others.

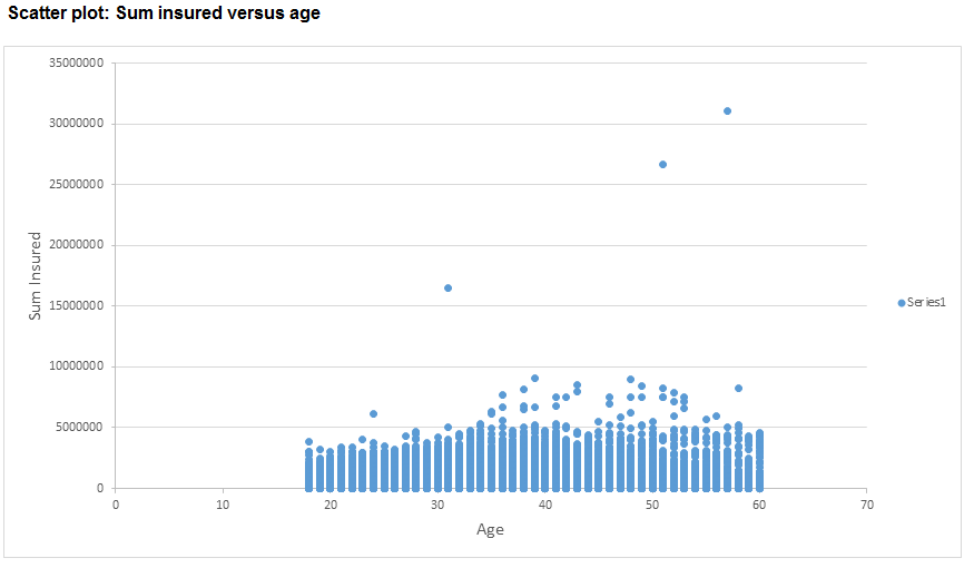

Figure 17

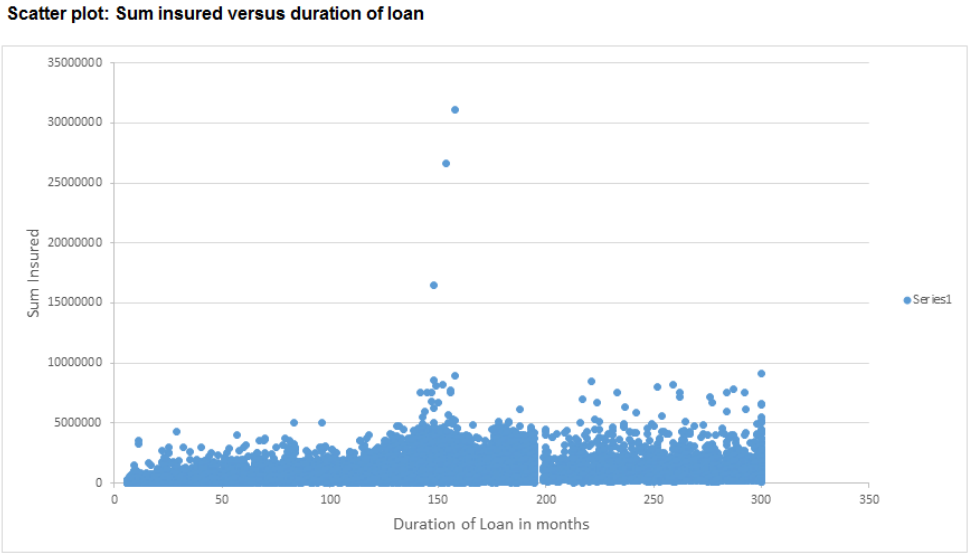

Figure 18

8. There are 53,691 insureds of maximum 200 months duration of loan.
9. There are approximately 2,100 insureds from 200 to 300 months.

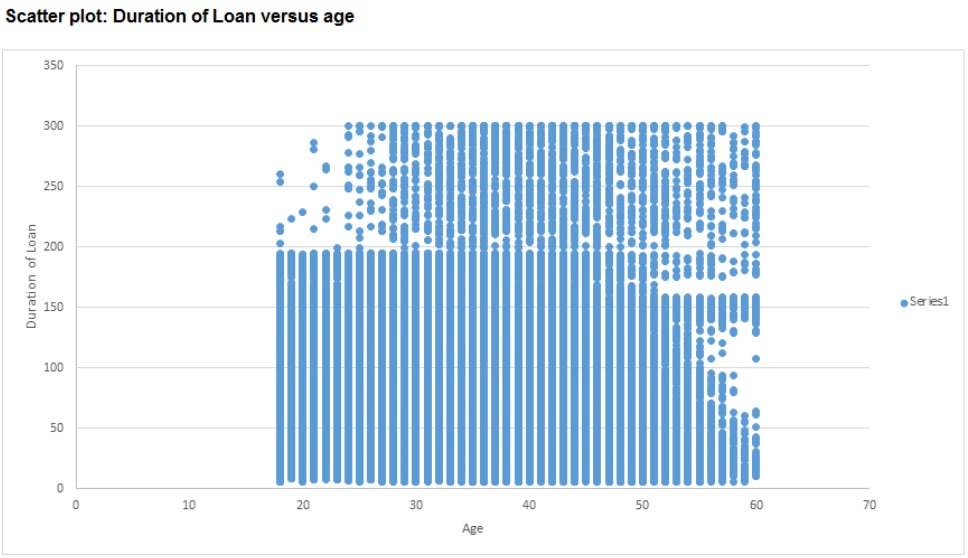

Figure 19: [7]

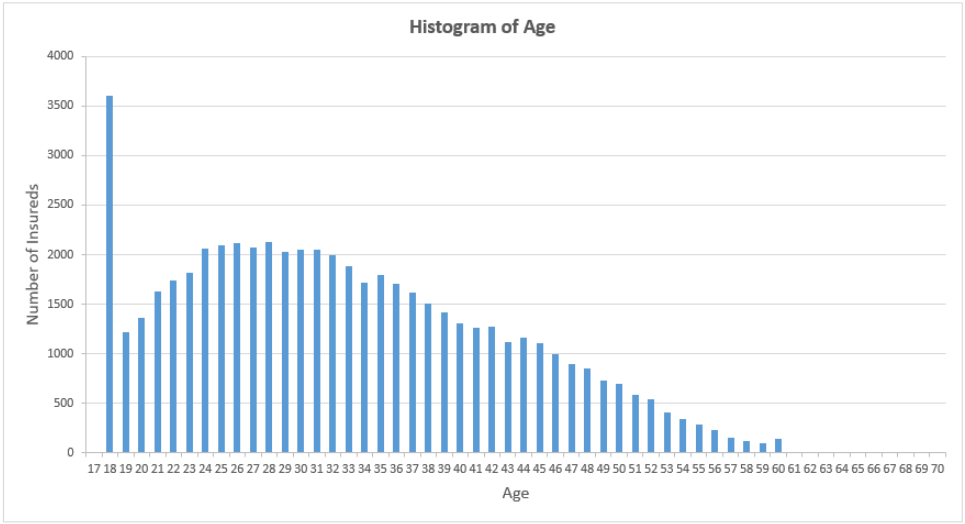

Figure 20

[7] y axis denotes duration of loan in months

10. There are 45,260 insureds in between 18 to 45 years of age group (Figure 20).
11. There are only 144 insureds of 60 years of age (Figure 20).
12. Male dominates the portfolio (male=47,631 and female =8,272) as shown below in the Figure 21.

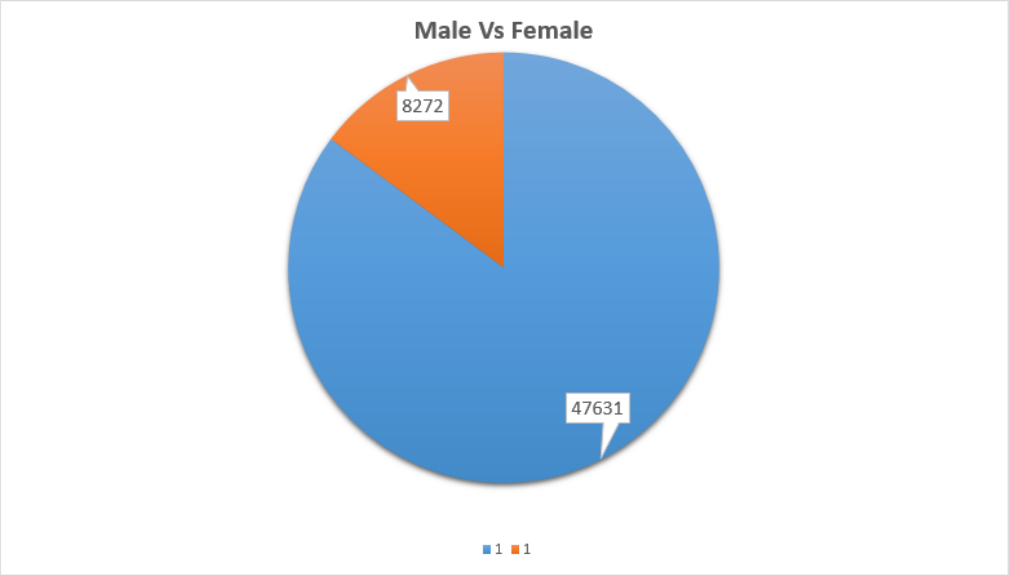

Figure 21: Male versus Female

5.3 Exact mathematical solution

First, we would like to present an exact mathematical solution. Exact mathematical solution is obtained by running data eleven times in the tool (discussed in detail in basic reserving tool in VBA section 4.2) because the tool cannot deal with large data set, so we create eleven subsets (each subset contains 5000 data points) and estimated the value for each subset through reserving tool.

Since the mathematical correct solution is additive (section 4.2.6), with regard to the single policies, therefore we can add up the eleven reserving patterns to get the exact mathematical solution for the full dataset. This is one of the reasons to do clustering, so that we can put cluster representative in the tool and get reserving pattern as discussed later in this section in detail.

Figure 22 and 23 shows yearly consumption and reserves remaining (for terms "yearly consumption" and "reserves remaining" refer to section 4.1) respectively. Since in principle the yearly consumption is the product of the amount to be paid and the mortality (section 4.1) curve shows the following effects:

1. In the first two years the underwriting effect (for definition of underwriting refer to section 1) can be seen, i.e. the reduction of the qx by 20% and respectively 10%.
2. In the following years the increase in mortality (due to aging of the insured) and the decreasing of the sum insured (due to loan repayment) level out.
3. In the end the sum insureds are decreasing faster towards zero than the mortality increases so that premium consumption decreases to zero as well.

Since, sum insured is decreasing faster than mortality is increasing we need less premium. The effect of the decreasing of outstanding loan amount dominates the effect of increase probability of dying. After 15^{th} year there is slight increase before it decreases due to the fact that we have quite some loans that are more than 25 years and our simulation was done in two parts namely one with maximum duration of 15 years and one with duration above 15 years.

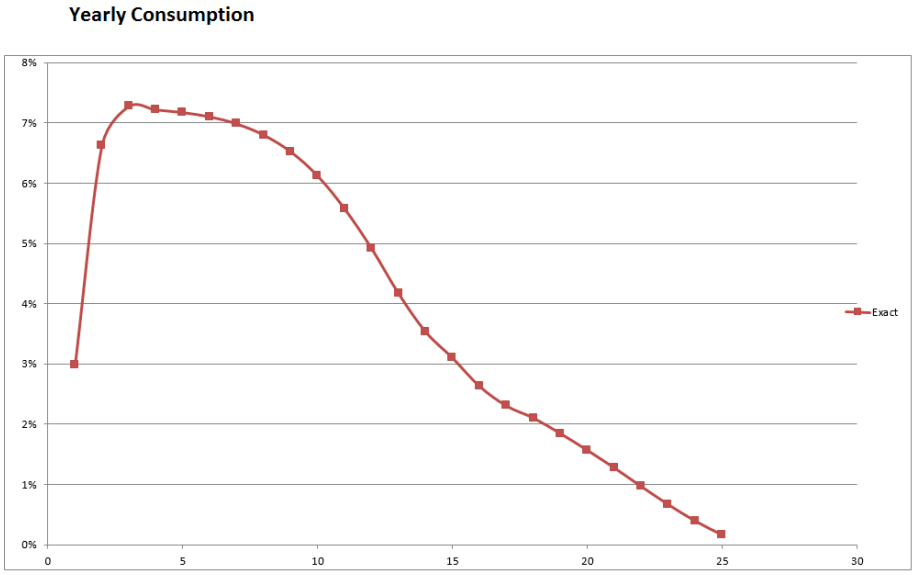

Figure 22: Exact Mathematical Solution[8]

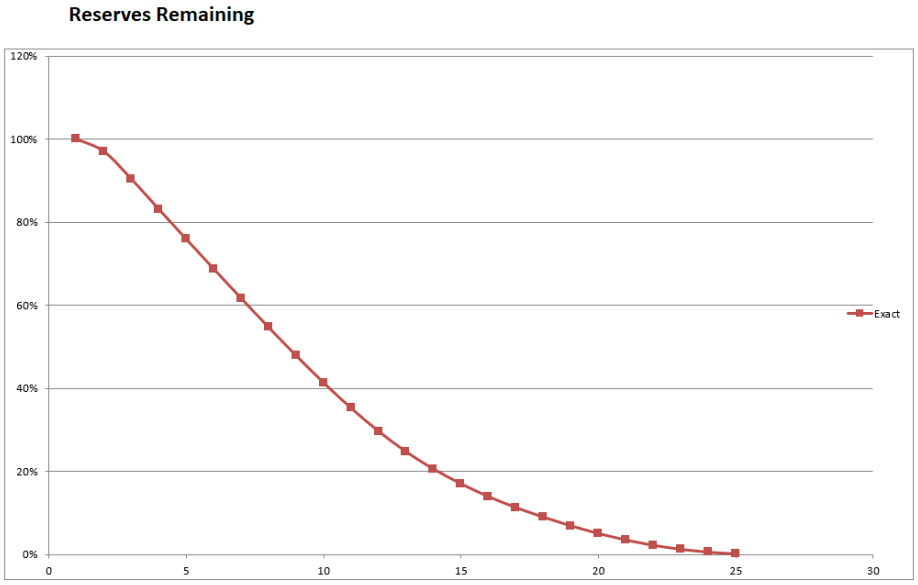

Figure 23: Exact Mathematical Solution[9]

[8] x axis denotes years and y axis denotes yearly consumption in percentage
[9] x axis denotes years and y axis denotes reserves remaining in percentage

In other words yearly consumption is the negative derivative of reserves remaining. As we can see below in the Table 4 that in 1^{st} year, we have 100% reserves (reserves remaining) out of that 3% (yearly consumption) is consumed. In 2^{nd} year there will be 97% (100%-3%) reserves remaining. And in 2^{nd} year we have 7% (yearly consumption) consumption, then in 3^{rd} year we will have 90% (97%-7%) reserves remaining and so on it goes till 25^{th} year.

Pattern: Exact mathematical solution		
Years	Yearly Consumption	Reserves remaining
1	3%	100%
2	7%	97%
3	7%	90%
4	7%	83%
5	7%	76%
6	7%	69%
7	7%	62%
8	7%	55%
9	7%	48%
10	6%	41%
11	6%	35%
12	5%	30%
13	4%	25%
14	4%	21%
15	3%	17%
16	3%	14%
17	2%	11%
18	2%	9%
19	2%	7%
20	2%	5%
21	1%	3%
22	1%	2%
23	1%	1%
24	0%	1%
25	0%	0%

Table 4: Exact Mathematical Solution

5.4 Comparison of exact and K-means clustered approach

Now we start comparing the exact mathematical solution that we have seen in section 5.3 with the solution derived by clustering with the K-means algorithm as described in section 3.3.1. Figure 24 and 25 below show following reserving patterns respectively:

1. The mathematical correct one.
2. The result after applying K-means algorithm.

First we will see list of cluster representative in Table 5 after applying K-means algorithm where K=20 using SparkR as discussed in section 4.2.8.

K-means cluster representative		
Age	Duration of Loan in months	Sum Insured
33	145	1,645,936,000
46	34	556,939,149
29	40	564,611,328
22	128	1,800,780,080
22	43	395,719,860
38	175	2,589,620,250
31	169	3,754,283,417
22	124	1,910,463,541
25	41	560,281,061
38	178	2,737,705,047
45	131	1,491,417,060
32	147	1,745,327,220
46	134	1,626,557,796
47	35	591,864,550
33	143	1,816,785,840
46	200	1,092,588,825
31	170	3,274,992,607
22	127	2,133,016,906
35	39	574,851,354
42	266	1,034,643,974

Table 5

Now we use these cluster representatives to get reserving pattern. But the problem is that all the reserves are consumed by 21^{st} year i.e. no reserve left for last 4 years as we can see in figure 24 and 25. In other words we can say that reserves are consumed very fast before 25^{th} year. This problem occurs due to the fact that we have not got a single cluster of 300 months (25 years) of duration after applying K-means algorithm as we can see above in the Table 5.

But we have seen in section 5.3 that exact mathematical solution keeps reserves till last year i.e. 25^{th} year which is necessary by law that insurance company must keep the reserve till the treaty expires or contract ends.

As we can see clearly in the Figure 24 that K-means keeps more reserves than the exact mathematical solution in initial years and later on after 14^{th} year starts consuming more reserves than the exact mathematical solution.

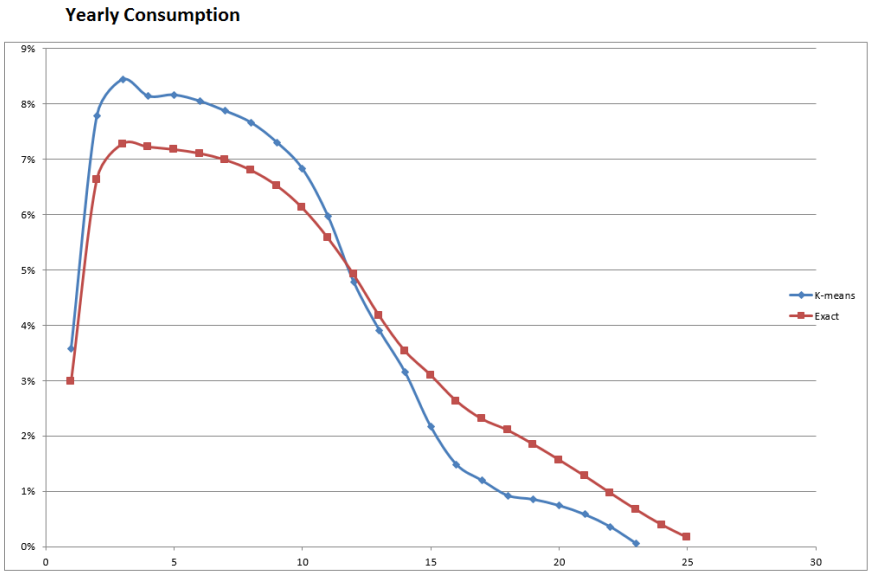

Figure 24: Pattern: Exact and K-means[10]

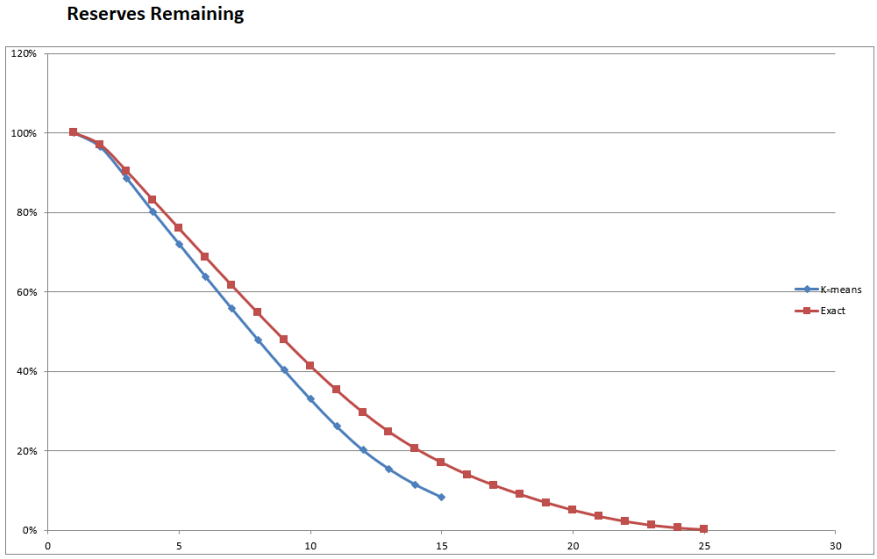

Figure 25: Pattern: Exact and K-means[11]

[10] x axis denotes years and y axis denotes yearly consumption in percentage
[11] x axis denotes years and y axis denotes reserves remaining in percentage

In hypothesis section 5.1 we coined a term ”exact enough” and argued that we will give meaning to it as we proceed in this chapter. That's why we would like to discuss another measure of fit to compare results i.e. conservative and aggressive.

Conservative
A reserving pattern consumes less premium i.e. keeps more reserves in the beginning, which might be at the cost of profitability although the advantage would be, no shortage of reserves.

For example below in Table 6 we can see that exact mathematical solution and K-means have 3% and 4% yearly consumption of reserves respectively in 1^{st} year. And K-means keeps more reserves till 11^{th} year. So we can say that K-means is conservative in initial years i.e. keeps more reserves.

Aggressive
A reserving pattern consumes more premium i.e keeps less reserves but suffers a disadvantage of lack of reserves before the treaty expires or contract ends.

For example below in Table 6 we can see that exact mathematical solution and K-means have 4% and 3% yearly consumption of reserves respectively in 14^{th} year. And K-means keeps less reserves from 14^{th} till last year. We can clearly say that K-means is aggressive in last years i.e. keeps less reserves.

As we can see below in the Table 6 that for K-means in 1^{st} year, we have 100% reserves (reserves remaining) out of that 4% (yearly consumption) is consumed . So in the 2^{nd} year there will be 96% (100%-4%) reserves remaining. And in 2^{nd} year we have 8% (yearly consumption) consumption, then in 3^{rd} year we will have 89% (96%-8%)[12] reserves remaining and so on it goes till 25^{th} year.

Here, we have calculated the square error in the Table 6 and it is 16.46 for K-means compared to exact mathematical solution.

[12]Since the values are rounded thats why difference is 89% otherwise difference should be 88%

Pattern: Exact mathematical solution			Pattern: K-means		
Years	Yearly Consumption	Reserves remaining	Yearly Consumption	Reserves remaining	Square error
1	3%	100%	4%	100%	0.36
2	7%	97%	8%	96%	1.32
3	7%	90%	8%	89%	1.36
4	7%	83%	8%	80%	0.85
5	7%	76%	8%	72%	0.98
6	7%	69%	8%	64%	0.90
7	7%	62%	8%	56%	0.78
8	7%	55%	8%	48%	0.75
9	7%	48%	7%	40%	0.62
10	6%	41%	7%	33%	0.50
11	6%	35%	6%	26%	0.16
12	5%	30%	5%	20%	0.02
13	4%	25%	4%	15%	0.07
14	4%	21%	3%	12%	0.15
15	3%	17%	2%	8%	0.86
16	3%	14%	1%	6%	1.31
17	2%	11%	1%	5%	1.23
18	2%	9%	1%	4%	1.39
19	2%	7%	1%	3%	0.98
20	2%	5%	1%	2%	0.68
21	1%	3%	1%	1%	0.48
22	1%	2%	0%	0%	0.37
23	1%	1%	0%	0%	0.37
24	0%	1%			
25	0%	0%			
					Sum = 16.46

Table 6

5.5 Comparison of exact, K-means and bisecting K-means clustered approach

Now we will compare the exact mathematical and K-means solution with the solution derived from bisecting K-means clustering algorithm as described in section 3.3.2. In the Figure 26 and 27 below, following reserving patterns are displayed:

1. The mathematical correct one.
2. The result after applying K-means algorithm.
3. The result after applying bisecting K-means algorithm.

Below in the Table 7, we have shown the list of cluster representatives after applying bisecting K-means algorithm with 20 clusters using SparkR as discussed in section 4.2.8.

Bisecting K-means cluster representative		
Age	Duration of Loan in months	Sum Insured
28	139	6,240,679,380
47	86	1,532,762,360
29	39	1,149,981,651
41	241	361,090,440
45	263	152,433,944
24	139	661,351,680
42	158	830,046,924
35	198	190,153,152
42	161	656,331,456
44	71	132,795,492
27	180	1,300,325,106
54	224	123,203,016
32	272	156,642,114
44	265	426,183,660
39	268	114,291,000
50	170	204,134,706
54	252	78,975,123
44	163	40,133,355
31	49	16,145,875
57	158	31,073,152

Table 7

Now we use these cluster representatives to get reserving pattern. But again we come across the same problem, as in section 5.4, that all the reserves are consumed before 25^{th} year (Figure 26 and 27). This problem occurs due to the same fact that we have not got a single cluster of 300 months (25 years) of duration after applying bisecting K-means algorithm as we can see above in the Table 7.

As depicted by the Figure 26, it is absolutely clear that bisecting K-means keeps more reserves than the exact mathematical solution for each and every year. Wherein for K-means, the reserves are consumed very fast in the last years as we have discussed in section 5.4.

5.5.1 Shortcoming

Since for both K-means and bisecting K-means algorithm, reserves are consumed before the reserve of the exact mathematical solution. Which by law is wrong because insurance companies have an obligation to keep the reserves till the end of the treaty expires as discussed in section 4.2.9. In our case, treaty ends at 25^{th} year. For K-means, all the reserves are consumed till 21^{st} year and for bisecting K-means also till 22^{nd} year. So, there would be no reserve left for the last 3 to 4 years.

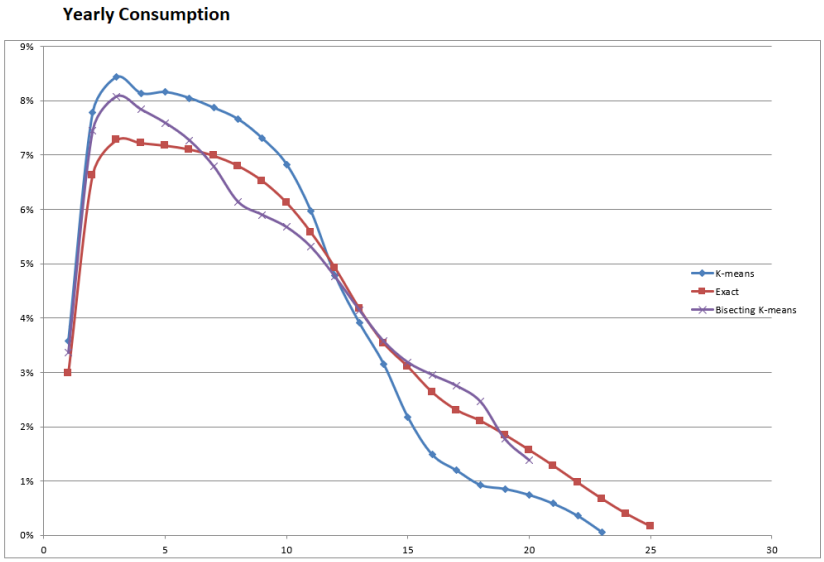

Figure 26: Pattern: Exact, K-means and Bisecting K-means[13]

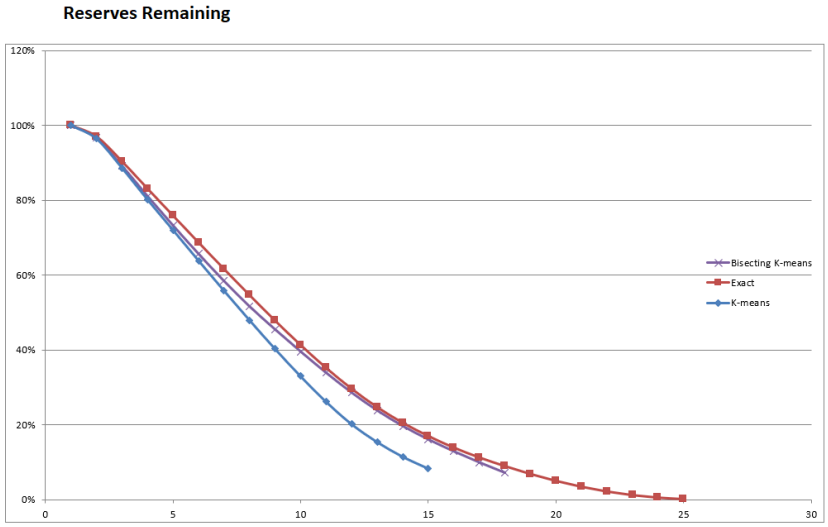

Figure 27: Pattern: Exact, K-means and Bisecting K-means[14]

[13] x axis denotes years and y axis denotes yearly consumption in percentage
[14] x axis denotes years and y axis denotes reserves remaining in percentage

As we can see below in the Table 8 that for bisecting K-means in 1^{st} year, we have 100% reserves (reserves remaining) out of that 3% (yearly consumption) is consumed . So in the 2^{nd} year there will be 97% (100%-3%) reserves remaining. And in 2^{nd} year we have 7% (yearly consumption) consumption, then in 3^{rd} year we will have 89% (97%-7%)[15] reserves remaining and so on, till 25^{th} year.

As we can see below in the Table 8 that square error for bisecting K-means algorithm is 4.30, which is less than the K-means (16.46). So, we can clearly say that bisecting K-means performs better than K-means.

Pattern: Exact mathematical solution			Pattern: Bisecting K-means clustering		
Years	Yearly Consumption	Reserves remaining	Yearly Consumption	Reserves remaining	Square error
1	3%	100%	3%	100%	0.15
2	7%	97%	7%	97%	0.66
3	7%	90%	8%	89%	0.65
4	7%	83%	8%	81%	0.38
5	7%	76%	8%	73%	0.18
6	7%	69%	7%	66%	0.03
7	7%	62%	7%	58%	0.04
8	7%	55%	6%	52%	0.04
9	7%	48%	6%	45%	0.39
10	6%	41%	6%	40%	0.20
11	6%	35%	5%	34%	0.07
12	5%	30%	5%	29%	0.02
13	4%	25%	4%	24%	0.00
14	4%	21%	4%	20%	0.00
15	3%	17%	3%	16%	0.01
16	3%	14%	3%	13%	1.11
17	2%	11%	3%	10%	1.20
18	2%	9%	2%	7%	0.13
19	2%	7%	2%	5%	0.01
20	2%	5%	1%	3%	0.03
21	1%	3%	1%	2%	0.06
22	1%	2%	1%	1%	0.19
23	1%	1%	0%	0%	0.36
24	0%	1%			
25	0%	0%			
					Sum = 4.30

Table 8

[15] Since the values are rounded thats why difference is 89% otherwise difference should be 90%

5.5.2 Solution

Because of the shortcoming explained above in section 5.5.1, we will do linear interpolation on both K-means and bisecting K-means in next step to make sure that reserves last till 25^{th} year.

5.6 Comparison of exact, K-means and K-means with interpolation

Now we will compare the result of K-means interpolated algorithm with exact mathematical and K-means algorithm solution. In the Figure 28 and 29 below, we can find reserving patterns :

1. The mathematical correct one.
2. The result after applying K-means algorithm.
3. The result after applying K-means with interpolation.

Here in the Figure 28 and 29 below we can see that K-means with interpolation keeps the reserves till the 25^{th} year same as for exact solution. Wherein for K-means all reserves are consumed by the end of 21^{st} year.

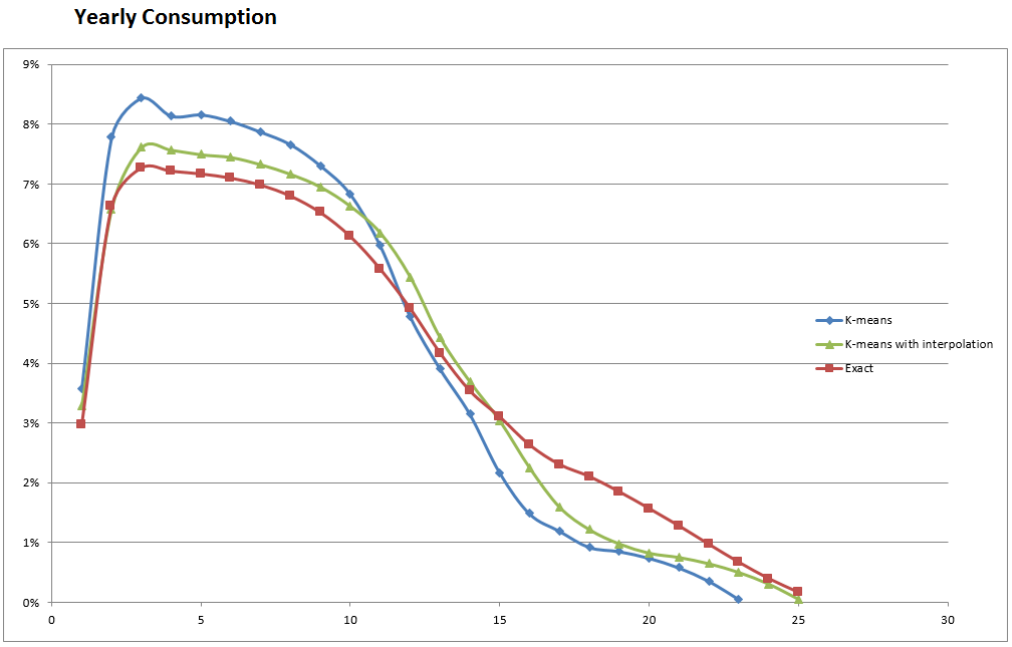

Figure 28: Pattern: Exact, K-means and K-means with Interpolation[16]

[16] x axis denotes years and y axis denotes yearly consumption in percentage

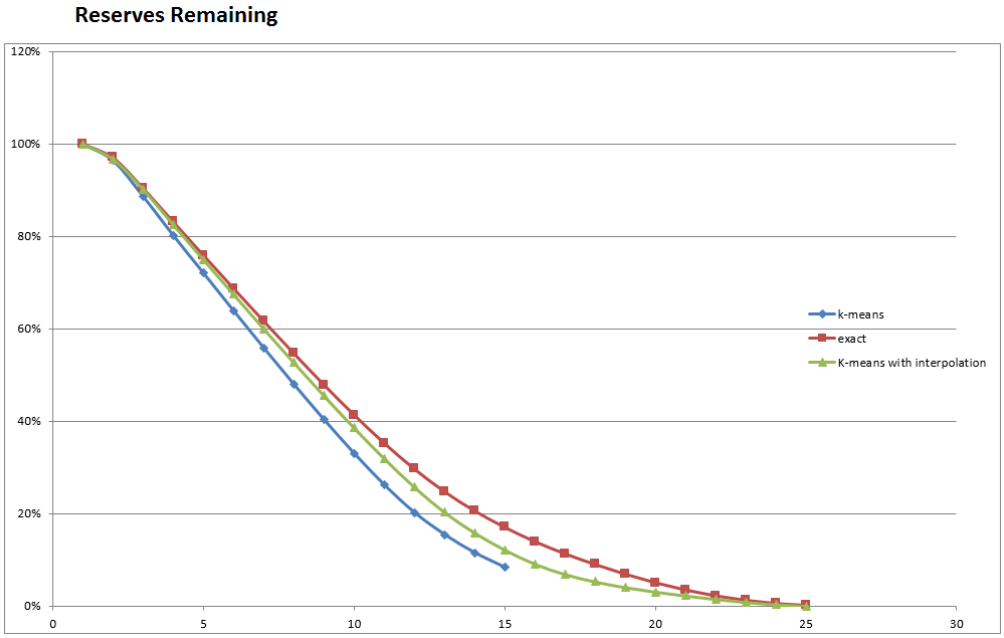

Figure 29: Pattern: Exact, K-means and K-means with Interpolation[17]

Below in the Table 9, for K-means with interpolation in 1^{st} year, we have 100% reserves (reserves remaining) out of that 3% (yearly consumption) is consumed . In 2^{nd} year there will be 97% (100%-3%) reserves remaining. And in 2^{nd} year we have 7% (yearly consumption) consumption, then in 3^{rd} year we will have 90% (97%-7%). And in 3^{rd} year we have 8% (yearly consumption) consumption, then in 4^{th} year we will have 83% (90%-8%) [18] reserves remaining and so on it goes till 25^{th} year.

In the Table 9 below exact mathematical solution is compared with K-means with interpolation. As we can see in the Table 9, square error for K-means with interpolation is 5.18 which is very less than the square error of K-means without interpolation i.e. 16.46. So, we can say that K-means with interpolation performs better than K-means without interpolation but still lags behind bisecting K-means algorithm in terms of mean square error.

[17] x axis denotes years and y axis denotes reserves remaining in percentage
[18] Since the values are rounded thats why difference is 83% otherwise difference should be 82%

Pattern: Exact mathematical solution			Pattern: K-means with interpolation		
Years	Consumption	Reserves remaining	Consumption	Reserves remaining	Square error
1	3%	100%	3%	100%	0.10
2	7%	97%	7%	97%	0.00
3	7%	90%	8%	90%	0.12
4	7%	83%	8%	83%	0.13
5	7%	76%	7%	75%	0.10
6	7%	69%	7%	67%	0.12
7	7%	62%	7%	60%	0.12
8	7%	55%	7%	53%	0.14
9	7%	48%	7%	46%	0.19
10	6%	41%	7%	39%	0.26
11	6%	35%	6%	32%	0.37
12	5%	30%	5%	26%	0.28
13	4%	25%	4%	20%	0.07
14	4%	21%	4%	16%	0.02
15	3%	17%	3%	12%	0.00
16	3%	14%	2%	9%	0.14
17	2%	11%	2%	7%	0.51
18	2%	9%	1%	5%	0.78
19	2%	7%	1%	4%	0.75
20	2%	5%	1%	3%	0.55
21	1%	3%	1%	2%	0.27
22	1%	2%	1%	2%	0.10
23	1%	1%	1%	1%	0.03
24	0%	1%	0%	0%	0.01
25	0%	0%	0%	0%	0.01
					Sum = 5.18

Table 9

5.7 Comparison of exact, bisecting K-means and bisecting K-means with interpolation

Here we will compare the result of bisecting K-means interpolated with exact mathematical solution and bisecting K-means algorithm solution. Below in the Figure 30 and 31 following reserving patterns shown are:

1. The mathematical correct one.
2. The result after applying bisecting K-means algorithm.
3. The result after applying bisecting K-means with Interpolation.

Here in the Figure 30 and 31 below we can see that bisecting K-means with interpolation keeps the reserves till the 25^{th} year same as for exact solution. Wherein for bisecting K-means algorithm all reserves get consumed by the end of 22^{nd} year.

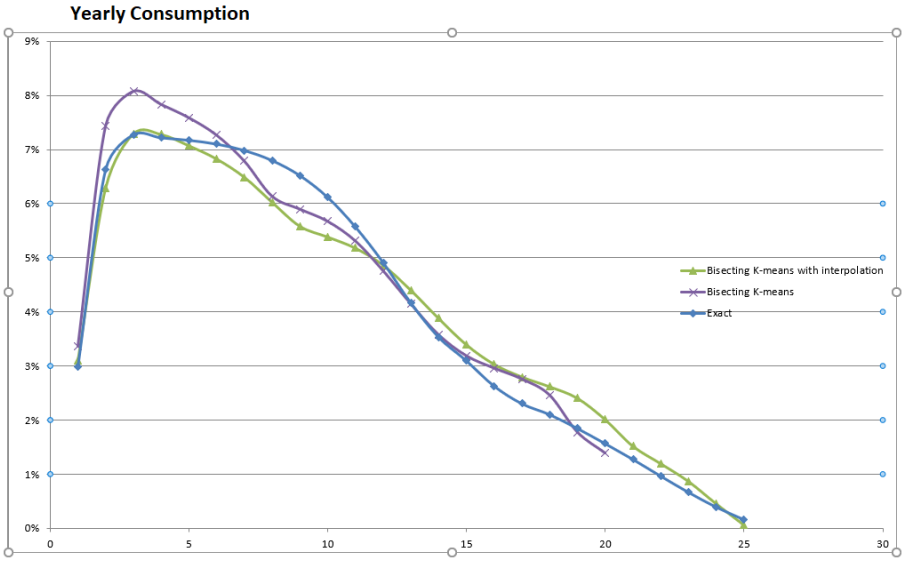

Figure 30: Pattern: Exact, Bisecting K-means and Bisecting K-means with Interpolation[19]

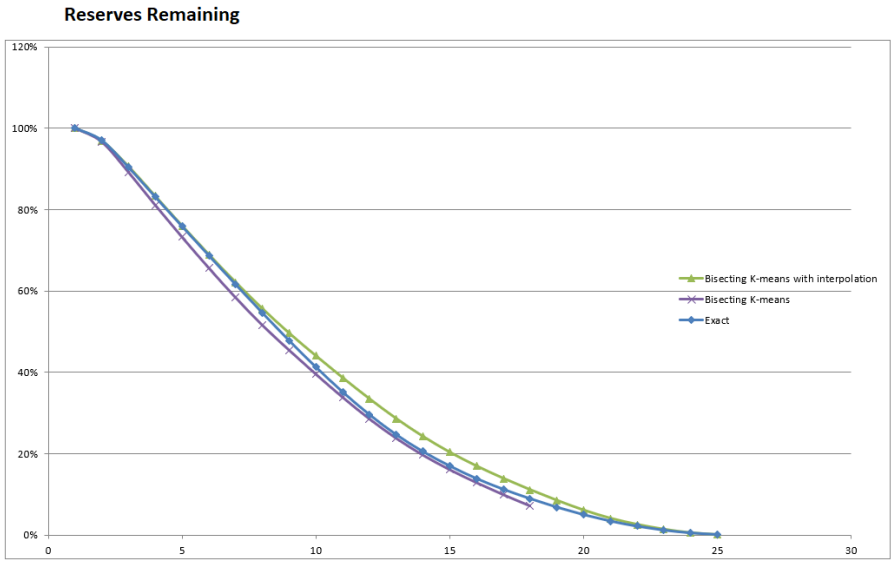

Figure 31: Pattern: Exact, Bisecting K-means and Bisecting K-means with Interpolation[20]

[19] x axis denotes years and y axis denotes yearly consumption in percentage
[20] x axis denotes years and y axis denotes reserves remaining in percentage

As we can see below in the Table 10 that for bisecting K-means with interpolation in 1^{st} year, we have 100% reserves (reserves remaining) out of that 3% (yearly consumption) is consumed . In 2^{nd} year there will be 97% (100%-3%) reserves remaining. And in 2^{nd} year we have 6% (yearly consumption) consumption, then in 3^{rd} year we will have 91% (97%-6%). And in 3^{rd} year we have 7% (yearly consumption) consumption, then in 4^{th} year we will have 83% (91%-7%) [21] reserves remaining and so on it goes till 25^{th} year.

In the Table 10 below exact mathematical solution is compared with bisecting K-means with interpolation. As we can see in the Table 10 that square error for bisecting K-means algorithm with interpolation is 4.22 which is less as compared to K-means with interpolation i.e. 5.18. This leads to conclusion that bisecting K-means with interpolation performs better than K-means with interpolation.

Pattern: Exact mathematical solution			Pattern: Bisecting K-means with Interpolation		
Years	Consumption	Reserves remaining	Consumption	Reserves remaining	Square error
1	3%	100%	3%	100%	0.01
2	7%	97%	6%	97%	0.12
3	7%	90%	7%	91%	0.00
4	7%	83%	7%	83%	0.00
5	7%	76%	7%	76%	0.01
6	7%	69%	7%	69%	0.07
7	7%	62%	6%	62%	0.24
8	7%	55%	6%	56%	0.59
9	7%	48%	6%	50%	0.88
10	6%	41%	5%	44%	0.54
11	6%	35%	5%	39%	0.16
12	5%	30%	5%	33%	0.00
13	4%	25%	4%	29%	0.05
14	4%	21%	4%	24%	0.12
15	3%	17%	3%	20%	0.09
16	3%	14%	3%	17%	0.16
17	2%	11%	3%	14%	0.24
18	2%	9%	3%	11%	0.26
19	2%	7%	2%	9%	0.31
20	2%	5%	2%	6%	0.20
21	1%	3%	2%	4%	0.06
22	1%	2%	1%	3%	0.05
23	1%	1%	1%	1%	0.04
24	0%	1%	0%	1%	0.00
25	0%	0%	0%	0%	0.01
					Sum = 4.22

Table 10

[21] Since the values are rounded thats why difference is 83% otherwise difference should be 84%

5.8 Summary and observation

Now we would like to discuss some important facts that were encountered during the implementation of the above mentioned algorithms and the effectiveness of the obtained results. In the Figure 32 and 33 below following reserving patterns are shown:

1. The mathematical correct one.
2. The result after applying K-means algorithm with interpolation.
3. The result after applying bisecting K-means algorithm with interpolation.

As seen in the previous chapters, the algorithms which use interpolation, performs much better than the ones without interpolation. In addition, they do not automatically comply with the law and can therefore not be used in reserving. Therefore, we concentrate our summary on the ones using interpolation. K-means with interpolation gives good approximation compared to exact mathematical solution. Under the mean square error as a measure of fit, the interpolated bisecting K-means algorithm performs better than interpolated K-means. Due to the fact that the clusters under bisecting K-means clustering show a stronger differentiation by age. Bisecting K-means with interpolation gives better results.

Below in the figure, one can find the reserve needed per year (Figure 32) for all the methods and the overall remaining reserve (Figure 33). Both interpolated algorithms are a bit more conservative regarding the reserves in the first years than the exact mathematical solution but becomes more aggressive in the later years. In my opinion, the interpolated bisecting K-means algorithm has some advantage here because it gives more conservative approach as compared to K-means interpolated (conservative here means keeping more reserves).

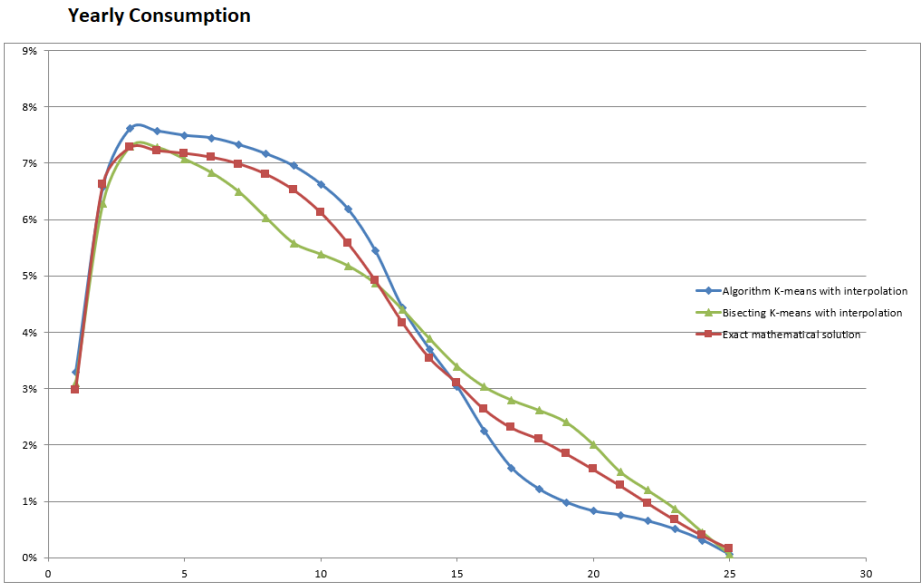

Figure 32: Pattern: Exact, K-means and Bisecting K-means with Interpolation[22]

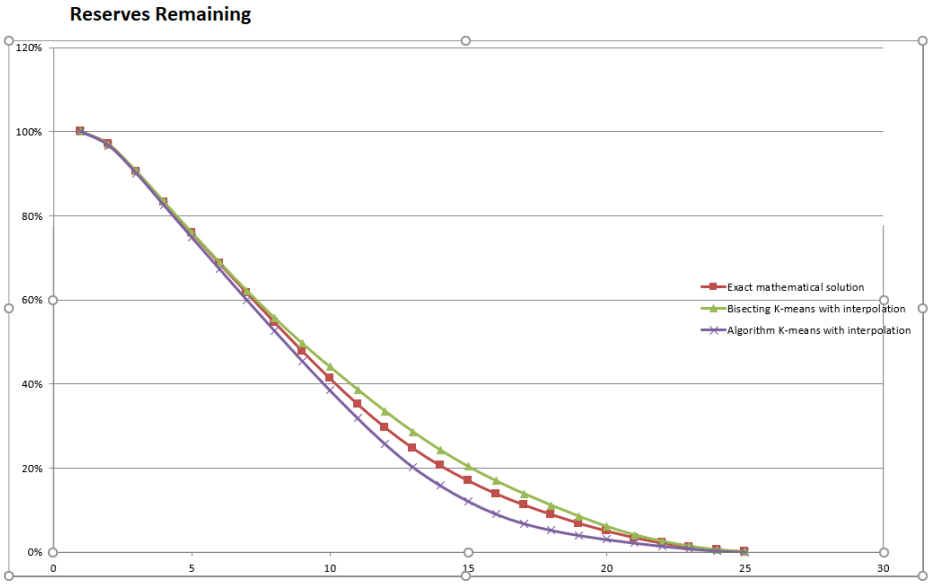

Figure 33: Pattern: Exact, K-means and Bisecting K-means with Interpolation[23]

[22] x axis denotes years and y axis denotes yearly consumption in percentage
[23] x axis denotes years and y axis denotes reserves remaining in percentage

5.9 Detailed discussion of interpolated algorithms

As discussed in our hypothesis (section 5.1) about the term "exact enough", we will give more precise meaning as we proceed in this chapter. Previously, we have discussed about conservative/aggressive in section 5.4. Now we would like to discuss another measure of fit i.e. maximum absolute error and it can be defined as:

$$D_i = ||{S_i}^{EMS} - {S_i}^{AL}|| \tag{28}$$

$$\Delta_{max} = maxD_i \quad , \quad 1 \leq i \leq 30 \quad \text{where,} \tag{29}$$

S_i is the reserves remaining in year i.
EMS is exact mathematical solution.
AL denotes algorithm it can be K-means or bisecting K-means.

As we have seen above in section 5.6 and 5.7, in detail that interpolated algorithm performs better than without interpolation. Now here also, we can see below in the Table 9 there is a absolute maximum difference of 4% in the reserves remaining.

For example, if there is reserves of one million Euros then error is about 40,000 (4% of one million). 4% might look much, but it is after 10 years before that there is almost no difference. From 11^{th} year to 17^{th}, the algorithm is a bit more conservative (keeps more reserves).

Here are few observations about interpolated bisecting K-means algorithm:

1. Interpolated bisecting K-means algorithm is clear winner by mean square error as we have seen above.
2. It is non negative throughout (always keeps reserves).
3. In the initial years (till 10^{th} year), the difference is almost zero.
4. If we take maximum difference of K-means interpolated algorithm it will not perform better (as discussed in detail below in Table 12).

As we can see below in the Table 11 that in 1^{st} year, we have 100% (reserves remaining) reserves for both exact mathematical solution and bisecting K-means with interpolation. So the difference will be 0% (100%-100%). In 2^{nd} year also it is equal (difference is 0%).

In 3^{rd} year we have 91% and 90% reserve remaining for bisecting K-means and exact mathematical solution respectively. So, the difference will be 1% (91%-90%). And it goes on, in 11^{th} year we have 39% and 35% reserve remaining for bisecting K-means and exact mathematical solution respectively. So, the difference will be 4% (39%-35%). In the end we have taken maximum absolute difference i.e. 4%. One important thing to be noticed here is that interpolated bisecting K-means is conservative (keeps more reserves) in each and every year.

Pattern: Exact mathematical solution		Pattern: Bisecting K-means with Interpolation	
Years	Reserves remaining	Reserves remaining	Difference
1	100%	100%	0%
2	97%	97%	0%
3	90%	91%	1%
4	83%	83%	0%
5	76%	76%	0%
6	69%	69%	0%
7	62%	62%	0%
8	55%	56%	1%
9	48%	50%	2%
10	41%	44%	3%
11	35%	39%	4%
12	30%	33%	3%
13	25%	29%	4%
14	21%	24%	3%
15	17%	20%	3%
16	14%	17%	3%
17	11%	14%	3%
18	9%	11%	2%
19	7%	9%	2%
20	5%	6%	1%
21	3%	4%	1%
22	2%	3%	1%
23	1%	1%	0%
24	1%	1%	0%
25	0%	0%	0%
			Max = 4%

Table 11

As we can see below in the Table 12 that in 1^{st} year, we have 100% (reserves remaining) reserves for both exact mathematical solution and K-means with interpolation. So the difference will be 0% (100%-100%). In 2^{nd}, 3^{rd} and 4^{th} year also it is equal (difference is 0%). In 5^{th} year we have 75% and 76% reserve remaining for K-means and exact mathematical solution respectively. So, the difference will be -1% (75%-76%). And it goes on to calculate difference like this. In the end we have taken maximum absolute difference i.e.

5%.It is worth mentioning that this algorithm is in terms of maximum difference almost equal to the other, but is more aggressive than the exact mathematical solution which is a clear disadvantage.

For K-means absolute maximum difference is 5% which is more than bisecting K-means and interpolated bisecting K-means algorithm. Interpolated K-means algorithm performs good in the initial years (till 5^{th} year) but in general performs worse than the interpolated bisecting K-means algorithm as we can see below in the Table 12. It is more aggressive (keeps less reserves) compared to the bisecting K-means and interpolated bisecting K-means algorithm.

But another thing to be noticed here is that the difference is negative from 6^{th} to 21^{st} year. This clearly shows that interpolated K-means is aggressive (keeps less reserves) as discussed above.

Pattern: Exact mathematical solution		Pattern: K-means with Interpolation	
Years	Reserves remaining	Reserves remaining	Difference
1	100%	100%	0%
2	97%	97%	0%
3	90%	90%	0%
4	83%	83%	0%
5	76%	75%	-1%
6	69%	67%	-2%
7	62%	60%	-2%
8	55%	53%	-2%
9	48%	46%	-2%
10	41%	39%	-2%
11	35%	32%	-3%
12	30%	26%	-4%
13	25%	20%	-5%
14	21%	16%	-5%
15	17%	12%	-5%
16	14%	9%	-5%
17	11%	7%	-4%
18	9%	5%	-4%
19	7%	4%	-3%
20	5%	3%	-2%
21	3%	2%	-1%
22	2%	2%	0%
23	1%	1%	0%
24	1%	0%	1%
25	0%	0%	0%
			Max = 5%

Table 12

5.10 Final chart

In the Table 13 below ranking is given to the algorithms based on their performance (1 = best, 4 = worst). Interpolated bisecting K-means performs best among all other algorithms as a measure of mean square error and absolute maximum difference. Interpolated bisecting K-means is also very conservative (keeps more reserves) as compared to others.

Comparison chart				
	K-means	K-means Interpolated	Bisecting K-means	Bisecting K-means Interpolated
Mean square error	16.46	5.18	4.30	4.22
Absolute Max.	10%	5%	4%	4%
Conservative/Aggressive	Aggressive	Less aggressive	Less conservative	Conservative
Ranking	4	3	2	1

Table 13: Ranking

We think that the interpolated bisecting K-means algorithm suffices our hypothesis because it performs better on all the parameters as compared to others, as we can see above in the Table 13. Following points justify that the interpolated bisecting K-means suffices our hypothesis in detail:

1. Bisecting K-means interpolated has mean square error of 4.22. If we assume that reserve difference to be equal for each year i.e. $\epsilon = R_k - R_k^{sp},\ k = 1, \ldots, 30$ and put this into the error formula discussed in our hypothesis section 5.1 we get:

$$Error = 10000 \cdot \sum_{k=1}^{30} (R_k - R_k^{sp})^2 \tag{30}$$

$$\frac{4.22}{10000} = 30 \cdot \epsilon^2 \tag{31}$$

$$\frac{4.22}{10000 \cdot 30} = \epsilon^2 \tag{32}$$

$$\epsilon = 0.3\% \tag{33}$$

   So the percentage of error is marginal i.e. only 0.3%.

2. The maximum absolute error is only 4% which is very less as compared to other algorithms (Table 13).

3. Last but not least interpolated bisecting K-means is conservative i.e. it always keeps reserves. So we never ran out of reserves as discussed in section 5.4.

5.11 Scalability

The data is increasing day by day and it is quite challenging to manage. Our findings will prove fruitful in the future scenario of big data. SparkR can handle big data and in order to demonstrate this, we replicated our data ten times that we have shown in section 5.2. We have proved earlier that bisecting K-means algorithm is best amongst the others (section 5.10), therefore we will apply it here also, in performing clustering on the replicated data.

Below in the Table 14, we have shown the list of cluster representative after applying bisecting K-means algorithm with 20 clusters using SparkR as discussed in section 4.2.8.

Bisecting K-means cluster representative		
Age	Duration of Loan in months	Sum Insured
30	113	8,698,200,000
34	167	37,876,000,000
57	158	776,828,800,000
31	139	15,959,925,000
33	155	24,764,650,000
35	180	71,477,600,000
39	177	84,261,675,000
45	207	208,062,675,000
31	103	5,954,500,000
36	177	60,419,475,000
31	129	12,379,000,000
35	33	1,285,850,000
45	186	101,622,125,000
35	167	31,045,400,000
45	236	179,884,100,000
32	145	19,733,225,000
34	171	44,834,575,000
32	76	3,463,725,000
35	177	51,562,575,000
44	209	121,702,625,000

Table 14

Now we will compare the exact mathematical solution with the solution derived from the bisecting K-means clustering algorithm as described in section 3.3.2. In the Figure 34 and 35 below, following reserving patterns are displayed:

1. The mathematical correct one.
2. The result after applying bisecting K-means with interpolation.

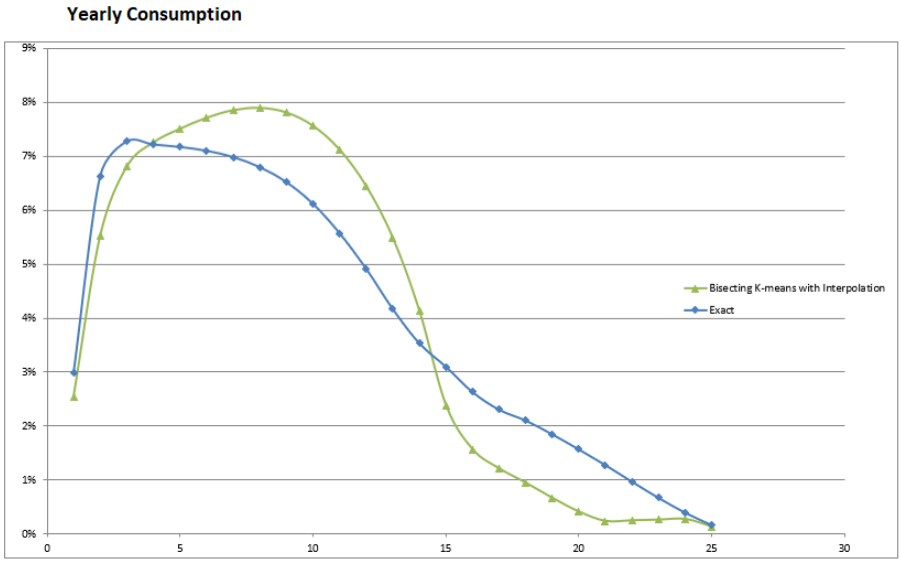

Figure 34: Pattern: Exact and Bisecting K-means with Interpolation[24]

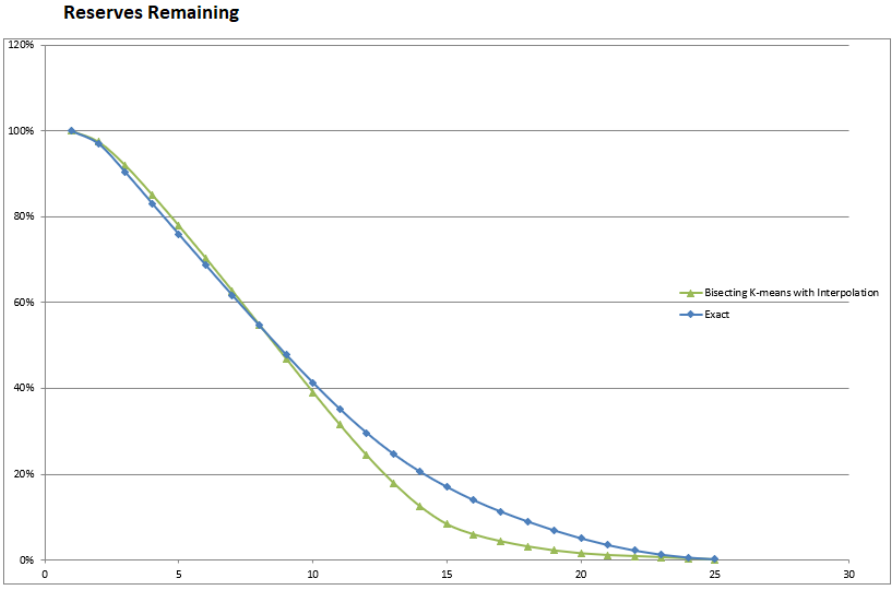

Figure 35: Pattern: Exact and Bisecting K-means with Interpolation[25]

[24] x axis denotes years and y axis denotes yearly consumption in percentage
[25] x axis denotes years and y axis denotes reserves remaining in percentage

The above figures 34 and 35 show the reserving pattern on replicated data. Hence, it is apt to conclude that the outcome of this thesis will be very beneficial towards the more potential researches in the area of reserving.

6 Future work

In this thesis, we have done reserving purely on insureds. Interesting thing would be to do reserving not only on insureds but also on claims data. Then make appropriate proposition for adjustment in reserving and also define profitability during the contract duration. For that we need accounting data and treaty is needed to be monitored throughout the contract duration.

Since at the moment we only have 55,000 insureds and to check profitability we need big treaty. So that it can be monitored throughout the contract duration. Another part of problem is that 55,000 insureds produce only around 30 claims per year, where we would need around 1000 claims to make at least rough calculations.

Apart from reserving, there are other things that could be done using analytics and can be very helpful for insurance companies. Just to give a glimpse of that, insurance companies have already started asking questions like:

1. What more can our own data tell us?
2. What else could we learn if we add external data to our models?
3. How can we ensure the power of analytics into day-to-day decision making?
4. What are the technologies required for data analytics?
5. Is big data really the future?

In the life insurance industry, analytics can help a company create a comprehensive road map for managing the entire life-cycle of a customer, from acquisition to lapse or maturity. Analytics also assists an insurer to have an enterprise-wide view of a client which provides an insight and acknowledgment of opportunities across all business lines as we can see below in the Table 15.

Life insurers' top future uses for big data analytics		
53% want to increase market penetration with data & analytics		
Other planned uses	Now	Two years
Transform business model	12%	48%
Expand customer relationships	41%	61%
Enhance customer value proposition	18%	52%
Improve internal performance management	29%	44%

Table 15: Willis Towers Watson's 2015 North American Life Insurance CFO Survey on Big Data Analytics

Nine in ten life insurance companies report using big data analytics to better compete in today′s market, according to a new LIMRA report [26].

According to LIMRA, while most companies are exploring big data analytics programs, less than a third feel they are ahead of their competition in this field. LIMRA believes as companies innovate and find ways to use new technologies to improve business results, there will be a differentiation within the market [10].

6.0.1 Challenges

Challenges faced by insurance industries in implementing big data analytics program are shown below in Table 16.

Top Challenges to Big Data Analytics Programs	
Lack of financial resources	58%
Getting executive buy-in	50%
Accessing data in legacy systems	50%
Don't know enough about it	37%
Building business case/showing value	34%
Lack of human resources or needed skills	34%
Other activities taking higher priority	21%
Confidentiality/privacy concerns	13%
Other	16%

Table 16: Big Data Analytics in Financial Services, LIMRA

Not unexpectedly, funding and executive buy-in are the two most cited hurdles for companies to implement big data analytics programs as shown above in the Table 16, as companies try to decide the importance of these programs to their organizations. But

[26] LIMRA, a worldwide research, learning and development organization headquartered in Connecticut, USA , is the trusted source of industry knowledge for over 850 financial services firms. LIMRA provides its members with the latest insight and analysis on retirement, insurance and distribution, helping them develop effective business strategies that positively impact the bottom line.

legacy systems and staffing are also a challenge. Currently most companies say they have fewer than 10 people dedicated to their big data analytics program and lot of companies report that it is difficult to find people with right skills[10].

Conclusion: Reserving is an important phenomenon in the insurance industry and is generally performed by the long-used conventional mathematical methods. These methods are not capable of handling large datasets. Data is proliferating at a much faster pace than anybody has ever imagined and is one of the main concerns of this digitalized world. Clearly, handling data in a more productive fashion is a need of the hour. This thesis has demonstrated an advanced, efficient, and user friendly approach towards reserving when handling big data using Apache spark. An account of clustering algorithms have been illustrated in the thesis which help in the accurate calculation of reserves.

A rigor of events, led to the understanding that bisecting K-means algorithm has proved to be the most efficient amongst its peers. And the results obtained are comparable to the exact mathematical solution.

References

[1] Ron Adiel. Reinsurance and the management of regulatory ratios and taxes in the property casualty insurance industry. *Journal of Accounting and Economics*, 22(1):207–240, 1996.

[2] Samuel L Baker. Perils of the internal rate of return. *Retrieved January*, 12:2007, 2000.

[3] Thierry Blu, Philippe Thévenaz, and Michael Unser. Linear interpolation revitalized. *IEEE Transactions on Image Processing*, 13(5):710–719, 2004.

[4] Richard A Brealey, Stewart C Myers, Franklin Allen, and Pitabas Mohanty. *Principles of corporate finance*. Tata McGraw-Hill Education, 2012.

[5] Mark Budnitz. The sale of credit life insurance: The bank as fiduciary. *NCL Rev.*, 62:295, 1983.

[6] James H Carr and Lopa Kolluri. Predatory lending: An overview. *Fannie Mae Foundation*, pages 1–17, 2001.

[7] Raymond B Cattell. The scree test for the number of factors. *Multivariate behavioral research*, 1(2):245–276, 1966.

[8] Florence Corpet. Multiple sequence alignment with hierarchical clustering. *Nucleic acids research*, 16(22):10881–10890, 1988.

[9] Aswath Damodaran. Estimating risk free rates. *WP, Stern School of Business, New York*, 1999.

[10] Norah Denley. Life insurers invest in big data analytics. *Retrieved June*, page 6, 2014.

[11] Gary Fagg. *Credit Life and Disability Insurance*. CLICO Management, 1986.

[12] Nuno Fernandes. *Finance for Executives: A practical guide for managers*. NPVPublishing, 2014.

[13] Ronald Aylmer Fisher et al. 012: A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error. 1920.

[14] Aristides Gionis, Heikki Mannila, and Panayiotis Tsaparas. Clustering aggregation. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):4, 2007.

[15] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[16] Edwin C Hustead. Trends in retirement income plan administrative expenses. *Living with defined contribution pensions*, pages 166–177, 1998.

[17] Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence*, 24(7):881–892, 2002.

[18] John Case Karen Berman, Joe Knight. *Financial Intelligence, Revised Edition: A Manager's Guide to Knowing What the Numbers Really Mean*. HARVARD BUSINESS REVIEW PRESS, 2013.

[19] Sidney Katz, Laurence G Branch, Michael H Branson, Joseph A Papsidero, John C Beck, and David S Greer. Active life expectancy. *New England journal of medicine*, 309(20):1218–1224, 1983.

[20] Donald E Knuth. Computer programming as an art. In *ACM Turing award lectures*, page 1974. ACM, 2007.

[21] Chad McCracken and Eric Wertheimer. Underwriting: The poetics of insurance in america, 1722-1872, 2008.

[22] Xiangrui Meng, Joseph Bradley, Burak Yavuz, Evan Sparks, Shivaram Venkataraman, Davies Liu, Jeremy Freeman, DB Tsai, Manish Amde, Sean Owen, et al. Mllib: Machine learning in apache spark. *The Journal of Machine Learning Research*, 17(1):1235–1241, 2016.

[23] Don Harper Mills. Medical insurance feasibility study: a technical summary. *Western Journal of Medicine*, 128(4):360, 1978.

[24] Christopher M Murtaugh, Brenda C Spillman, and Mark J Warshawsky. In sickness and in health: An annuity approach to financing long-term care and retirement income. *Journal of Risk and Insurance*, pages 225–253, 2001.

[25] Sev V Nagalingam. *CIM justification and optimisation*. CRC Press, 1999.

[26] Louis Perridon, Manfred Steiner, and Andreas W Rathgeber. Financial management of the company, page 178-412.

[27] Alexis Pozen and David M Cutler. Medical spending differences in the united states and canada: the role of prices, procedures, and administrative expenses. *INQUIRY: The Journal of Health Care Organization, Provision, and Financing*, 47(2):124–134, 2010.

[28] Samuel Preston, Patrick Heuveline, and Michel Guillot. Demography: measuring and modeling population processes. 2000.

[29] Michael Rothschild and Joseph Stiglitz. Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *The quarterly journal of economics*, pages 629–649, 1976.

[30] Mark A Rothstein. *Genetics and life insurance: Medical underwriting and social policy*. MIT Press, 2004.

[31] Florian Schewe. Reserve estimation and analysis with generalized additive models for location, scale and shape, 2012.

[32] D Sai Srinivas and Stuart Land . Credit life insurance. *8th Global Conference of Actuaries, Mumbai*, 2006.

[33] Michael Steinbach, George Karypis, Vipin Kumar, et al. A comparison of document clustering techniques. In *KDD workshop on text mining*, volume 400, pages 525–526. Boston, 2000.

[34] Erwin Straub and Dawson Grubbs. The faculty and institute of actuaries claims reserving manual. volume 1 and 2. *ASTIN Bulletin*, 28(02):287–289, 1998.

[35] R Studio. Rstudio: integrated development environment for r. *RStudio Inc, Boston, Massachusetts*, 2012.

[36] R Core Team. R language definition. *Vienna, Austria: R foundation for statistical computing*, 2000.

[37] Tim Verdonck, Martine Van Wouwe, and Jan Dhaene. A robustification of the chain-ladder method. *North American Actuarial Journal*, 13(2):280–298, 2009.

[38] Richard Verrall. Claims reserving and generalised additive models. *Insurance: Mathematics and Economics*, 19(1):31–43, 1996.

[39] Hendrik Wagenaar. *Meaning in action: Interpretation and dialogue in policy analysis*. Routledge, 2014.

[40] John Walkenbach. *Excel 2010 power programming with VBA*, volume 6. John Wiley & Sons, 2010.

[41] Xiaojie Wang et al. *Introduction to statutory reserves in life insurance companies*. PhD thesis, 2011.

[42] Menahem E Yaari. Uncertain lifetime, life insurance, and the theory of the consumer. *The Review of Economic Studies*, 32(2):137–150, 1965.

[43] Anna Karin Spångberg Zepezauer. Handelsgesetzbuch–handelsbilanz. In *Steuerlehre und Bilanzierung für das Bachelor-Studium*, pages 140–170. Springer, 2017.