Approximate Valuation of Life Insurance Portfolio with the Cluster Analysis: Trade-Off Between Computation Time and Precision

Jan Fojtík[1] | *Prague University of Economics and Business, Prague, Czech Republic*
Jiří Procházka[2] | *Prague University of Economics and Business, Prague, Czech Republic*
Pavel Zimmermann[3] | *Prague University of Economics and Business, Prague, Czech Republic*

Abstract

Valuation of the insurance portfolio is one of the essential actuarial tasks. Life insurance valuation is usually based on a projection of cash flows for each policy which is demanding computation time. Furthermore, modern financial management requires multiple valuations under different scenarios or input parameters. A method to reduce computation time while preserving as much accuracy as possible based on cluster analysis is presented. The basic idea of the method is to replace the original portfolio by a smaller representative portfolio based on clusters with some weights that would ensure the similarity of the valuation results to the original portfolio. Valuation is then significantly faster but requires initial time for clustering and the results are only approximate – different from the original results. The difference is studied for a different number of clusters and the trade-off between the approximation error and calculation time is evaluated.

Keywords	*DOI*	*JEL code*
Life insurance, portfolio valuation, cash flow projection, cluster analyses	*https://doi.org/10.54694/stat.2021.23*	*C38, C63, G22*

INTRODUCTION

The proper valuation of the life insurance portfolio is one of the essential actuarial tasks. In general, there are several metrics describing the portfolio value such as the liability value, profit or loss or distributive

[1] Faculty of Informatics and Statistics, Prague University of Economics and Business, Prague, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic. E-mail: xfojj00@vse.cz.
[2] Faculty of Informatics and Statistics, Prague University of Economics and Business, Prague, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic. E-mail: xproj16@vse.cz.
[3] Faculty of Informatics and Statistics, Prague University of Economics and Business, Prague, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic. E-mail: zimmerp@vse.cz.

earnings. These metrics will be referred to as the economic metrics. Calculation of these *economic metrics* will be referred to as the *portfolio valuation*. Proper calculation of these metrics is important for reporting or accounting purposes, especially with new legislation of IFRS 17 or Solvency II. Actual and accurate valuation of life insurance portfolio has also a great importance in management decision or creating future business plans.

Ordinary methods for valuating life insurance products are usually not complicated or mathematically sophisticated. The valuation is based on per-policy projection – projecting the expected development and economic metrics of each policy (Selimovic, 2010) which is computationally demanding. Furthermore, actuaries usually need to valuate the portfolio multiple times under different assumptions of expected reality – e.g. valuation on different interest, mortality or lapse rates. For this purpose, the portfolio needs to be valuated under a wide range of different scenarios (Giamouridis et al., 2016; or Kaucic and Daris, 2015). Even with the newest technologies such as optimized actuarial software and powerful hardware, the results are derived with significant delay. For example, for a portfolio of a smaller insurance company with 300 000 policies, projecting 1 000 scenarios for 50 years (600 months), valuation may, depending on the performance of hardware and software, last even a month. The valuation time increases with the portfolio size or the number of required scenarios. The results derived with such a delay may be outdated or not satisfying actual market assumptions. Therefore, new opportunities of faster valuation present an active area of research.

Several researches suggest that data-mining methods can be used to solve this task (see Mohammed et al., 2016; or Devale and Kulkarni, 2012). A recent study (Janecek, 2017) presents two solutions for faster liability modeling based on mathematical proxy-functions and interpolations among different interest rate scenarios. Authors Freedman and Reynold (2008), and Fojtik et al. (2017) suggest cluster analysis as a good alternative way to accelerate valuation of life insurance portfolio. Presently, the application of cluster analysis has been in life insurance used mainly for the client segmentation purposes (Jandaghi et al., 2015), or experience analysis and assumption-setting process (Purushotham, 2016). Another approach for faster life insurance portfolio valuation is based on least squares Monte Carlo techniques. These techniques are studied in Turnbull (2014), Nteukam et al. (2014), and Krah and Nikolič (2018).

The approach researched in this paper is based on cluster analysis and will be referred to as *clustering approach*. This approach can be described in the following steps:

1. Calculate the economic metrics for each policy of the original portfolio based on a small number of initial runs – e.g. the basic (best estimate) scenario and few stress scenarios for the most important inputs.
2. Perform cluster analysis on economic metrics and potentially even on the policy parameters for a fixed, reasonably chosen number of clusters. Find a suitable representative (one policy) for each cluster. The portfolio of representatives will be referred to as the *reference portfolio*.
3. Calculate the projections of the economic metrics for all further scenarios only for the representatives of the clusters.
4. Find suitable weights of the representatives ensuring that the projection of economic metrics of the weighted reference portfolio will be as close as possible to the projection of the economic metrics of the original portfolio.

The more clusters (and hence representatives) are used, the slower the calculation and more precise the approximation. This constitutes a trade-off between computation time and accuracy. Analysis of this trade-off is the main contribution of this article.

In the first chapter the most typical method for life-insurance portfolio valuation is introduced including detail description of each component and calculation. The second chapter presents the alternative approach for faster valuation of life-insurance portfolio based on clustering analysis including basic parametrization, advantages and settings. The third chapter presents the main analysis of this paper.

In this section the alternative valuation approach is applied on artificial portfolio and its time and accuracy performance is discussed.

1 PROJECTING CASH FLOW PER EACH POLICY

The most common methods for life-insurance portfolio valuation are based on the cash flow projection. Typically, these models are a projection of future cash in-flows and out-flows for each policy or an aggregate of policies. Modelling all policies one by one will be referred to as *per-policy projection*. A detailed description of per-policy projection can be found in Dickson and Hardy (2013).

1.1 Components of cash flow model

In this section, we present a cash flow model that can be considered as a typical valuation approach for life insurance policies. Specific companies of course have specific variations of these models. However, it is not essential (nor possible) for this analysis to cover all particular deviations of valuation models. The typical components of the cash flow model are:

In-flows:

- Premium.

Out-flows:

- Claims,
- Surrenders,
- Maturities,
- Commissions,
- Expenses.

The cash flow is calculated as the difference between the expected in-flows and the expected out-flows. The cash flow $CF_{i,t}$ formula for the i^{th} policy at the time t is:

$$CF_{i,t} = EPrem_{i,t} - EClaims_{i,t} - ESurr_{i,t} - EMat_{i,t} - EComms_{i,t} - EExpens_{i,t}\,, \tag{1}$$

where the $EPrem_{i,t}$ is the expected value of premium paid at the beginning of the period t. The $EClaims_{i,t}$ is the expected value of coverage paid for loss or policy event that occurs during period t. The $ESurr_{i,t}$ is the expected value of surrenders paid for early contract cancellation at the end of the period t before the end of the policy period. The $EMat_{i,t}$ is the expected value of maturities paid at the end of the policy. The ECommsi,t is the value of commissions and $EExpens_{i,t}$ is the value of expenses expected to be paid at the beginning of the time period t. Cash flows and its components are projected in discrete time intervals indexed by t (months or years). The expected values of each component are the nominal values adjusted by the probability that the component will be paid.

The probability of death and the probability of surrender (pre-mature end of contract) are involved in the model. The notation q_x refers to the probability of death for a person in age x. The s_t denotes the probability of surrender in policy period t.

1.1.1 Expected premium

The premium $Prem_{i,t}$ is paid at the begin of the period t only if the insured person is alive and the contract has not been cancelled before. The expected value of premium for the i^{th} contract and period t is:

$$EPrem_{i,t} = \begin{cases} Prem_{i,t} & t = 1 \\ Prem_{i,t} \prod_{j=1}^{t-1}(1 - q_{x+j-1})(1 - s_j) & t > 1 \end{cases}, \tag{2}$$

where x is the age of insured person at the beginning of the valuation. The premium is paid at the beginning of the projected period therefore, the first payment in time $t = 0$ is not adjusted by any probability.

1.1.2 Expected claim

The claim $Claim_{i,t}$ is paid only if the accident happened in period t, the insured person was alive before the accident occurred and the contract has not been cancelled before the accident. The claim expected value for the i^{th} contract in time t is:

$$EClaim_{i,t} = \begin{cases} Claim_{i,t}(1 - s_t) & t = 1 \\ Claim_{i,t}\, q_{x+t-1}(1 - s_t) \prod_{j=1}^{t-1}(1 - q_{x+j-1})(1 - s_j) & t > 1 \end{cases}, \tag{3}$$

where x is the age of the insured person at the beginning of the valuation. The claim is assumed to be paid at the end of projected period t.

1.1.3 Expected surrender

The surrender value $Surr_{i,t}$ is paid only if the contract is cancelled before the end of the policy period and the insured person is still alive. The expected surrender value for i^{th} contract in time t is:

$$ESurr_{i,t} = \begin{cases} Surr_{i,t}(1 - q_x)s_t & t = 1 \\ Surr_{i,t}\, q_{x+t-1}(1 - q_{x+t-1})s_t \prod_{j=1}^{t-1}(1 - q_{x+j-1})(1 - s_j) & t > 1 \end{cases}, \tag{4}$$

where x is the age of the insured person at the beginning of the contract. The surrender is assumed to be paid at the end of projected period t.

1.1.4 Expected maturity

The maturity value $Mat_{i,t}$ is paid only if the contract has not cancelled and the insured person is alive at the end of the contract in time T. As the maturity is paid only at the end of the contract the expected maturity value before the end of contract is equal to zero. For the i^{th} contract and time equal to end policy period T the expected maturity is:

$$EMat_{i,t} = \begin{cases} 0 & t < T \\ Mat_{i,t} \prod_{j=1}^{t}(1 - q_{x+j-1})(1 - s_j) & t = T \end{cases}, \tag{5}$$

1.1.5 Expected commissions

The commission $Comms_{i,t}$ is paid at the beginning of the period t only if the insured person is alive and the contract has not been cancelled before. The expected commission value for the i^{th} contract in time t is:

$$EComms_{i,t} = \begin{cases} Comms_{i,t} & t = 1 \\ Comms_{i,t} \prod_{j=1}^{t-1}(1 - q_{x+j-1})(1 - s_j) & t > 1 \end{cases}, \tag{6}$$

where x is the age of the insured person at the beginning of the contract. The commission is assumed to be paid at the beginning of the projected period therefore, the first commissions in time $t = 0$ is not adjusted by probabilities.

1.1.6 Expected expenses

The expenses $Expens_{i,t}$ are paid in period t only if the insured person is alive and the contract has not been cancelled before. The expected value of expenses for the i^{th} contract in time t is:

$$EExpens_{i,t} = \begin{cases} Expens_{i,t} & t = 1 \\ Expens_{i,t} \prod_{j=1}^{t-1}(1 - q_{x+j-1})(1 - s_j) & t > 1 \end{cases}, \tag{7}$$

where x is the age of the insured person at the beginning of the contract. The expenses are assumed to be paid at the beginning of the projected period therefore, the first expenses in time $t = 0$ are not adjusted by probabilities.

1.2 Economic metrics

In this section the calculation of basic economic metrics is presented.

1.2.1 Present value of liability

The portfolio value is usually calculated as the present value of future cash flows *PVCF*. The present value of future cash flows for i^{th} policy is calculated as discounted sum of cash flows:

$$PVCF_i = \sum_t CF_{i,t}\, v^t\,, \tag{8}$$

where v^t is the flat discount factor at projection time t. For simplicity PVFC will be referred to the value of liability. The total value of the liability for the whole portfolio is sum of *PVCF* for each policy.

1.2.2 Present value of profit and loss

The profit of the i^{th} policy in projection time t is calculated as:

$$Profit_{i,t} = CF_{i,t} + \Delta Reserve_{i,t} + InvIncome_{i,t}\,, \tag{9}$$

where the $\Delta Reserve_{i,t}$ is the change of reserve (change of fund value between times $t - 1$ and t) and $InvIncome_{i,t}$ is the investment income realized between times $t - 1$ and t. The present value of profit for i^{th} policy is calculated as discounted sum of individual profits as:

$$PVprof_i = \sum_t Profit_{i,t}\, v^t, \tag{10}$$

where v^t is the flat discount factor at projection time t. The portfolio present value of profit is calculated as sum of present values of profits for all policies.

1.2.3 Present value of premium

The present value of premium for i^{th} policy in time t is calculated as discounted sum of individual expected premium payments:

$$PVPrem_i = \sum_t Eprem_{i,t}\, v^t, \tag{11}$$

where v^t is the flat discount factor at projection time t. The portfolio present value of premium is calculated as sum of present values of premium for all policies.

2 CLUSTER ANALYSIS AND PARAMETER SETTING

As stated above, some acceleration techniques focus on replacing the original portfolio by representatives of clusters. The cash flow projection is then performed for these representatives since there is potentially a major reduction of computation time. This is however connected with certain inaccuracy. Inaccuracy in this context means the difference between the projection results of the original portfolio and the weighted projection results of cluster representatives for a given scenario.

There are two essential decisions within the clustering approach to make. The number of clusters required and the set of clustering variables to be used. The number of clusters as well as the number of clustering variables increases computational time on one hand but may increase the accuracy on the other hand. A reasonably selected set of clustering variables may also increase the accuracy. Therefore, both should be optimized at least to some extent. The number of clusters represents the size of the reference portfolio which should be manually pre-selected. In the following analysis, the different number of clusters will be analysed to present its accuracy and speed trade-off.

2.1 Clustering variables

There are two distinct types of clustering variables available in this task. The first type is the basic policy characteristics such as age, gender, premium, sum assured or policy reserve. The second type is the economic metrics such as profits, premiums, claims or even cash flows in individual projected periods or intervals and its sensitivities to certain stress scenarios. The character of both types of clustering variables is very different. The economic metrics describe more the dynamics of the policy rather than position as the basic characteristics.

It has been confirmed in Fojtik (2017), that for the purposes of scenario valuation using the economic metrics as clustering variables leads to significantly better results than using the policy characteristics. Clustering variables assumed in this study are:

- the present values of future cash flows,
- the present value of profit or loss,
- the present value of premium,
- the sum of cash flows in the first five years,
- the sum of profits in the first five years,
- the sum of claims in the first five years,
- the sum of expenses and commissions in the first five years,
- the sum of premium in the first five years.

The selected clustering variables represent the common metrics describing the insurance portfolio from economical perspective. When using clustering variables, the problem of the different scale of nominal values may arise. To avoid this problem, it is advised to standardize the data before clustering. The standardization process consists of subtracting the mean and dividing it by the standard deviation.

2.2 Clustering algorithm

For the purposes of this paper, the non-hierarchical medoid based algorithm CLARA (Clustering LARge Application) was applied (Ng and Han, 2002). The CLARA algorithm is suitable for handling large datasets such as a life insurance portfolio (Hebak, 2013). In general, clusters are created by grouping similar observations. In the case of life insurance portfolio, clusters of policies with similar economic variables selected as clustering variables are created. The dissimilarity between the two policies is defined by the distance measure. In this paper, the dissimilarities are measured by the Euclidean distance between policies. The Euclidean distance is defined as the sum of squared differences between the i^{th} and the j^{th} policy:

$$d_{i,j} = \sqrt{\sum_{m=1}^{M} (Z_{m,i} - Z_{m,i})^2}\,, \tag{12}$$

where $Z_{m,i}$ and $Z_{m,j}$ are the standardized values of the m^{th} clustering variable of the i^{th} respectively j^{th} policy. The setting of R function is as follows: clara (x = Clustering_data, k = K, samples = 200, rngR = TRUE, stand = TRUE, correct.d = TRUE, metric = "Euclidean", pamLike = TRUE). The data object Clustering_data is insurance portfolio with clustering variables only and K stands for the number of clusters (size of reference portfolio). For more information about additional parameters see the documentation of the clustering function in the package *cluster* (Maechler et al., 2021).

2.3 System of weights

A system of weights must be assigned to the reference portfolio in order to replicate the projection of the original portfolio. Authors Freedman and Reynold (2008) suggest using the number of policies in each cluster as weight. This ensures that the number of reference policies matches the size of the original portfolio. Another option of the weighting system is scaling by some financial variable. In this paper, we present a weight based on the ratio of the present values of cash flows between the original portfolio and the reference portfolio. The weights are calculated for each cluster individually on the basic (best estimate) scenario. The weight of the kth cluster is given by:

$$w_k = \frac{PVCF_k^{Orig}}{PVCF_k}, \tag{13}$$

where the $PVCF_k^{Orig}$ represents the total present value of cash flows of policies from the original portfolio belonging to the k^{th} cluster and the $PVCF_k$ is the present value of cash flows of the k^{th} representative. This ensures that the reference portfolio will, in the basic scenario, replicate the present value of cash flow of the original portfolio exactly.

The approximate total value of the m_{th} projected variable $\widetilde{X}$ is for each scenario calculated as a weighted sum of the m_{th} projected variable $X_{m,k}$ of the representatives from the reference portfolio:

$$\widetilde{X}_m = \sum_{k=1}^{K} w_k X_{m,k}. \tag{14}$$

The symbol K stands for the number of clusters given by the size of the reference portfolio which is set manually before the clustering. Note that the weights are built on the best estimate scenario but can be used for projecting other stress scenarios.

2.4 Error measure

As we are trying to replicate the results of the per-policy projection of the original portfolio, the error in this context is the relative difference between the approximate value calculated by Formula (14) of the m_{th} projected variable and the corresponding variable of the original portfolio.

The error measure of the m_{th} variable is given by the following:

$$e_m = \frac{\widetilde{X}_m}{X_m} - 1, \tag{15}$$

where the $\widetilde{X}$ is obtained by Formula (14) and X_m is the total value of the m_{th} projected variable of the original portfolio. The total error of the reference portfolio is then defined as the average square root sum of squares over all selected variables as:

$$e = \frac{\sqrt{\sum_m e_m^2}}{M}. \tag{16}$$

It is advisable to measure the error only for the important variables in terms of actuarial modelling. In this paper, the error is measured on the clustering variable from section Clustering variables.

2.5 Computation time

The main goal of the clustering approach is to reduce the number of projected policies in order to speed up life insurance portfolio valuation with an acceptable level of inaccuracy. The total *computation time* consists of two components – *clustering time* and *valuation time*. The valuation time is required for projecting the policies and the clustering time is required for reducing the size of the original portfolio and building the reference portfolio.

Let's assume that the one scenario valuation of one policy by classical per-policy cash flow model takes in average time T_{avg}. The valuation time of $N_{scenarios}$ on the whole original portfolio of size $N_{policies}$ then lasts approximately:

$$T_{avg}\, N_{policies}\, N_{scenarios}\,. \tag{17}$$

In the case of the original portfolio, the total computation time is equal to the valuation time because no clustering is performed. But in the case of the reference portfolio, the total computation time is given by the sum of clustering and valuation time of the reference portfolio of size $N_{reference}$ as:

$$T_{clustering} + T_{avg} N_{reference}\, N_{scenarios}\,. \tag{18}$$

where the first component $T_{clustering}$ is the time required for clustering. For simplicity, let's assume that the average valuation time of one policy will remain approximately the same after the reduction. The acceleration by clustering approach is then:

$$\frac{T_{avg} N_{policies} N_{scenarios}}{T_{clustering} + T_{avg} N_{reference} N_{scenarios}}\,. \tag{19}$$

The section Analysis confirms that the size of the reference portfolio (number of clusters) increases both – the clustering time as well as the valuation time. The significant time saving is evident especially when testing more scenarios.

3 ANALYSIS

The goal of this analysis is to present an approach how to select the suitable number of clusters for the specific portfolio and the number of scenarios that preserves the high accuracy and significantly speeds up the portfolio valuation.

There are three essential aspects that need to be considered before selecting the number of clusters, namely:

- accuracy of the clustering approach,
- clustering time,
- total acceleration.

In this part, we present the relation between accuracy and clustering time with respect to the different number of clusters and the total acceleration for the different number of scenarios.

3.1 Experimental artificial portfolio

The analysis of the clustering approach is performed on an artificial life insurance portfolio that consists of universal-life insurance policies. The portfolio includes 100 000 policies. The 8 different policy products are ensuring a reasonable level of heterogeneity that may be observed in real portfolios. Each product has 12 500 policies. The products differ in the premium frequency, length of policy period or the system of benefit payments. The basic parameters of the portfolio are presented in Table 1. The artificial portfolio includes the basic policy characteristics and the metrics of economic profit based on best estimate assumptions.

In the first step of making the artificial portfolio, the basic policy characteristics were generated for each product individually.

Table 1 Basic overview of the artificial portfolio

Product	**A**	**B**	**C**	**D**
Average age	25	25	25	25
Average policy period	30	30	5	5
Max age	80	80	50	50
Term coefficient	1	1	0.25	0.25
Min policy period	10	10	5	5
Policy duration	10	10	1	1
Sum assured	500 000	500 000	500 000	500 000
Premium frequency	Regular	Single	Regular	Single
Benefit type	SA	SA	SA	SA
Product	**E**	**F**	**G**	**D**
Average age	30	30	30	30
Average policy period	30	30	5	5
Max age	80	80	50	50
Term coefficient	1	1	0,25	0,25
Min policy period	10	10	5	5
Policy duration	10	10	1	1
Sum assured	1 000 000	1 000 000	1 000 000	1 000 000
Premium frequency	Regular	Single	Regular	Single
Benefit type	SA+CV	SA+CV	SA+CV	SA+CV

Source: Own construction

The policy characteristics were generated as follows:

- The age of the client at the start of the valuation is generated from Poisson distribution with the specific mean for each product presented in Table 1.
- The policy period was calculated as follows:
  1. Firstly, the maximal length of the policy period h is calculated from the age of the client obtained from the previous step to a maximum possible age considered. The maximal possible age is presented in Table 1.

2. Secondly, the maximal period *h* is multiplied by a random variable generated from uniform distribution with minimum set to 0.1 and maximum to the term coefficient presented in Table 1.
3. In the last step, the minimal length of the policy period for each product is ensured by parameter *Min policy period* from Table 1.

- The policy duration is given by the policy period obtained from the previous step multiplied by a specific coefficient l generated from uniform distribution with minimum set to 0 and maximum 1. The l coefficient ensures that the policy duration is lower than the policy period for each contract.
- Sum assured was generated from normal distribution with the same mean and standard deviation parameter (Sum assured). This parameter can be seen in Table 1. To eliminate negative or very low values, the lower bound of the sum assured was set as 10 000.
- The premium was calculated by the deterministic pricing formulas for premium (see Cipra, 2014), for two specific types of benefit payments:
  – SA: the benefit of sum assured is paid only in case of death. The premium is calculated as for the term insurance products.

Figure 1 Examples of Cash-Flow projection in artificial portfolio

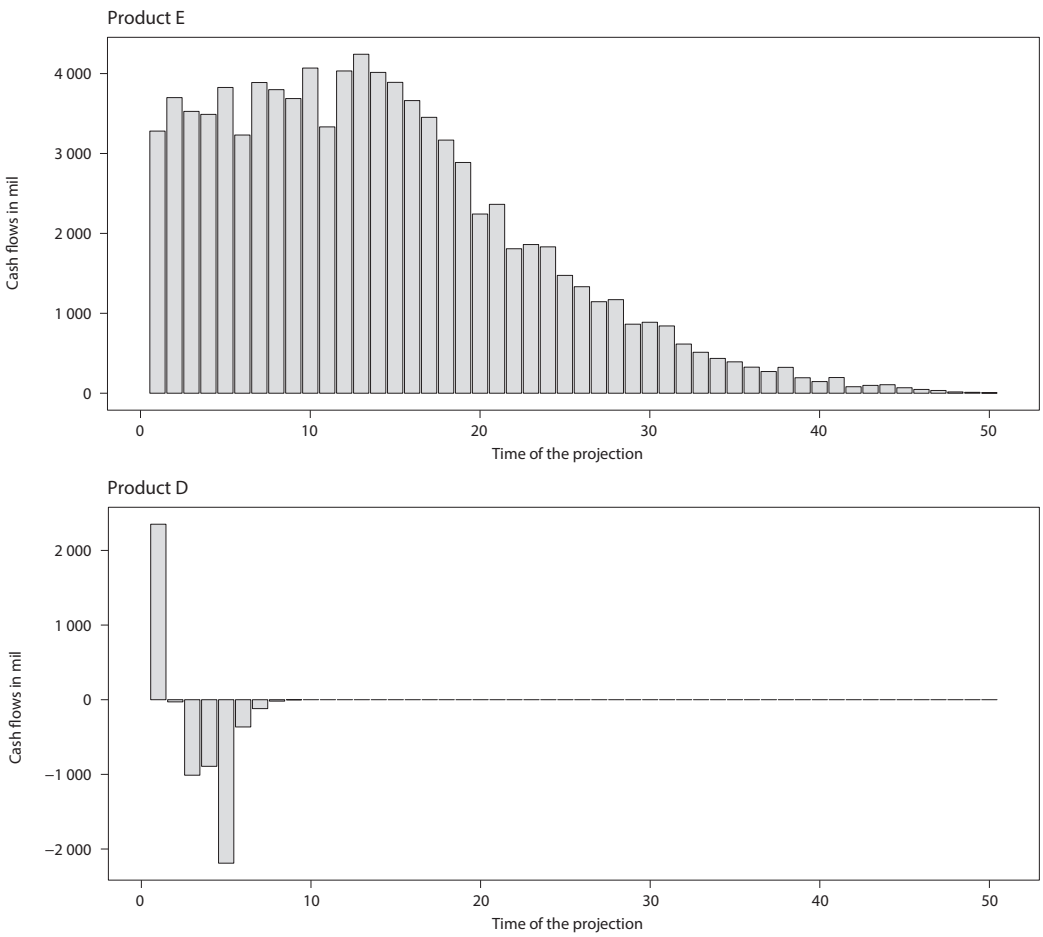
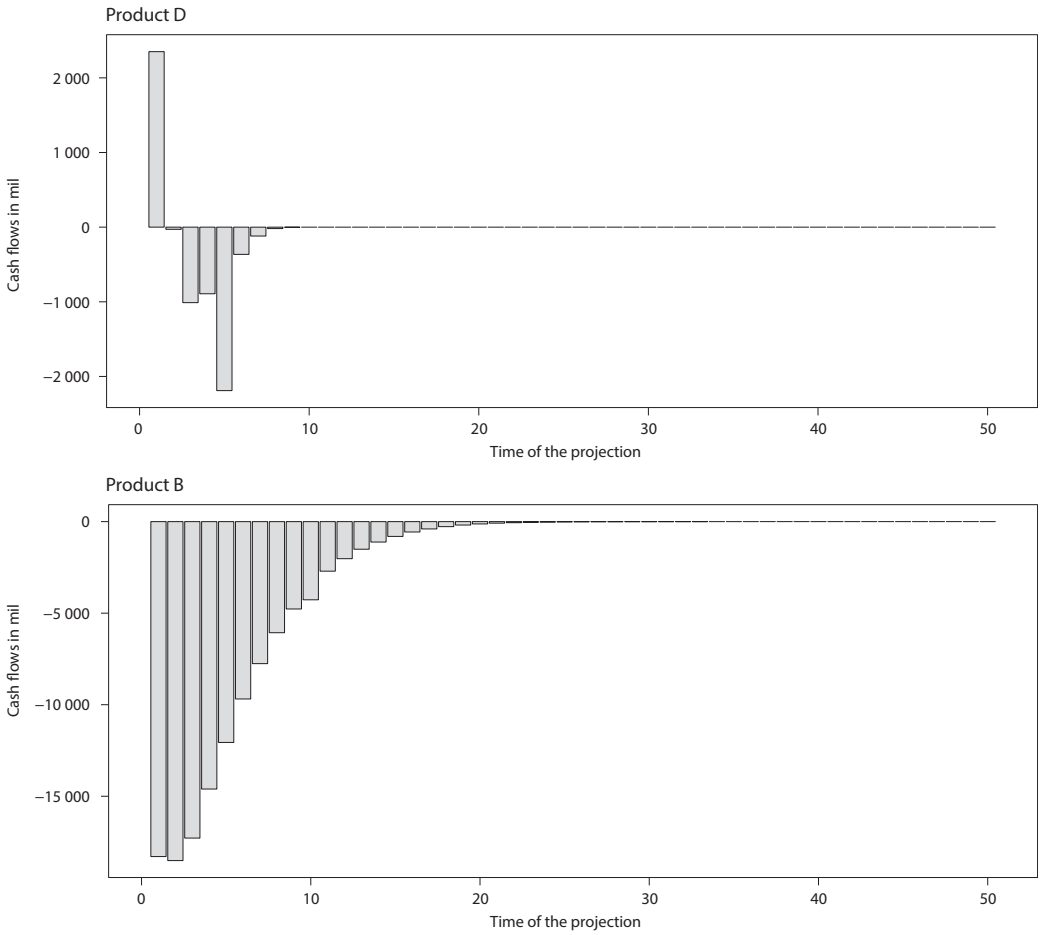


Figure 1 (continuation)



Source: Own construction

  - SA + CV: the benefit is paid in two cases – death and maturity (surviving to the end of the policy period). The premium is calculated as for the endowment insurance product where the death benefit is a sum assured and the survival benefit is the value of the fund at the end of the policy period.
- The fund value was calculated as a difference between premium paid over the policy duration with interest minus the expenses paid over the policy duration.

After generating the policy characteristics, the economic metrics are calculated by per-policy projection described in the section Components of cash flow model.

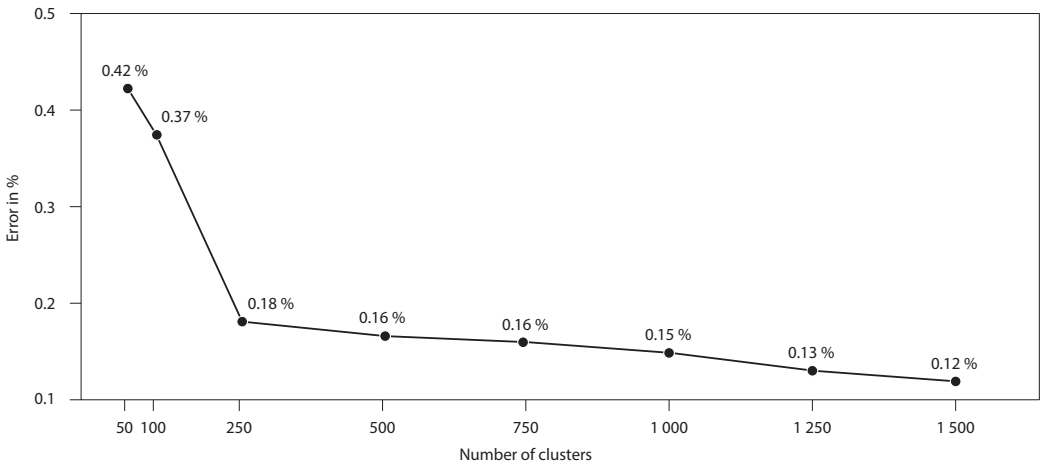
The examples of cash flows can be seen in Figure 1 on three selected products.

3.2 Number of clusters and the accuracy

Figure 2 presents the accuracy of the approximation for a different number of clusters used. The analysis is provided on the original portfolio designed in section Artificial portfolio. As stated previously, accuracy

increases (error decreases) with the number of clusters. At first, the error decreases very fast. Somewhere around 250 clusters, the decrease of the error slows down significantly and continues steadily.

Figure 2 Relation between the accuracy and the number of clusters

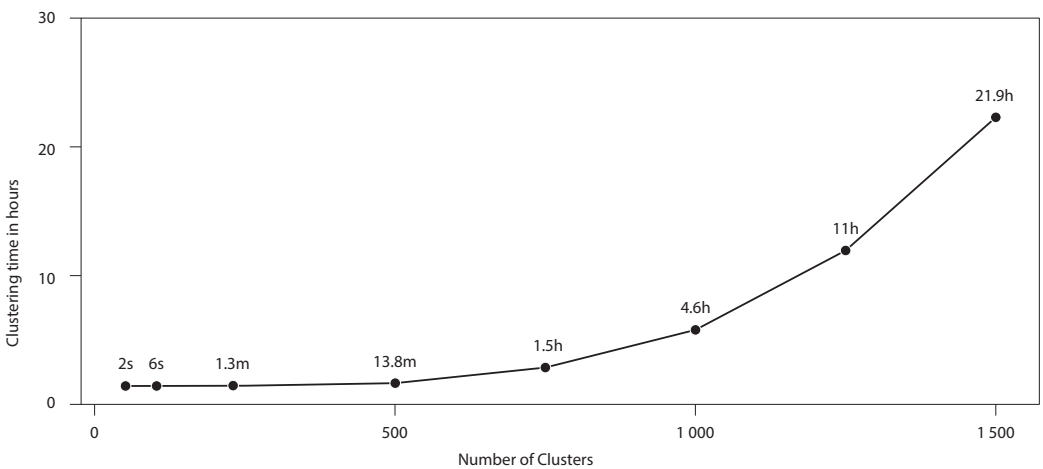


Source: Own construction

3.3 Number of clusters and the clustering time

Figure 3 presents the clustering time for the different number of clusters. The clustering time increases with the number of clusters. The increase is not linear but significantly faster. Therefore, the results for a high number of clusters (more than 10 000) may not be achieved in real time.

Figure 3 Relation between the clustering time and the number of clusters

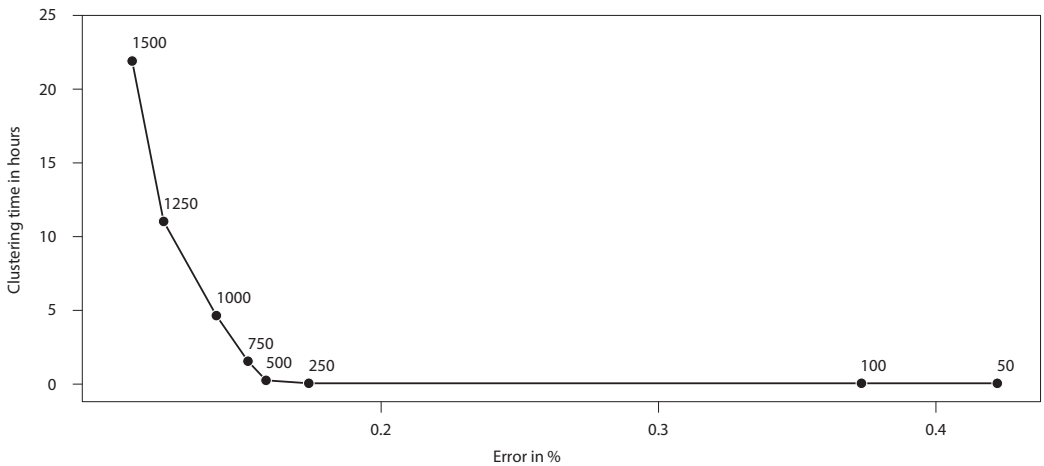


Source: Own construction

3.4 Accuracy and clustering time trade-off

Figure 4 puts the previous analysis together and presents the relation between the clustering time and accuracy achieved for the different number of clusters. The label of the line represents the number of clusters. The trade-off between the accuracy and clustering time, in this case, suggests that the reasonable number of clusters is somewhere between 250 and 500 where the additional increase in clustering time is not compensated by the significant increase in the corresponding accuracy.

Figure 4 Trade-off between the accuracy and clustering time



Source: Own construction

3.5 Acceleration of clustering approach

Table 2 present the acceleration calculated by formula 19 for the different number of clusters and one scenario. Acceleration naturally decreases towards 0 with the increasing number of clusters. For example, using the reference portfolio of 500 policies defined by the clustering approach seems to be beneficial because the whole calculation is 21 times faster already for one scenario. But using the reference portfolio of 1 500 policies is 4 times slower for one scenario as the valuation time and especially the clustering increase materially.

Table 3 presents the acceleration of the clustering approach for the different number of scenarios and the different number of clusters. The acceleration may differ for the different number of scenarios

Table 2 Reference portfolio acceleration for one scenario

Number of clusters	50	100	250	500	750	1 000	1 250	1 500
Valuation time	0.17	0.33	0.83	1.67	2.5	3.33	4.17	5
Clustering time	0.03	0.1	1.29	13.83	90.24	274	662	1 314
Calculation time	0.2	0.44	2.12	15.5	92.74	277	667	1 319
Acceleration	1 709	762	156	21	3.59	1.2	0.5	0.25

Source: Own construction

but usually raises with the number of scenarios. For the high number of scenarios, clustering time is negligible, and the acceleration is proportional to the number of clusters.

The results from Table 3 suggest that using 1 500 clusters for modelling only one scenario does not save any time but modelling 10 scenarios would be 2.44 times faster. The boundary $N^{+}_{Scenario}$ defines the minimal number of scenarios, where acceleration is higher than 1 (the clustering approach is beneficial). This boundary has the following:

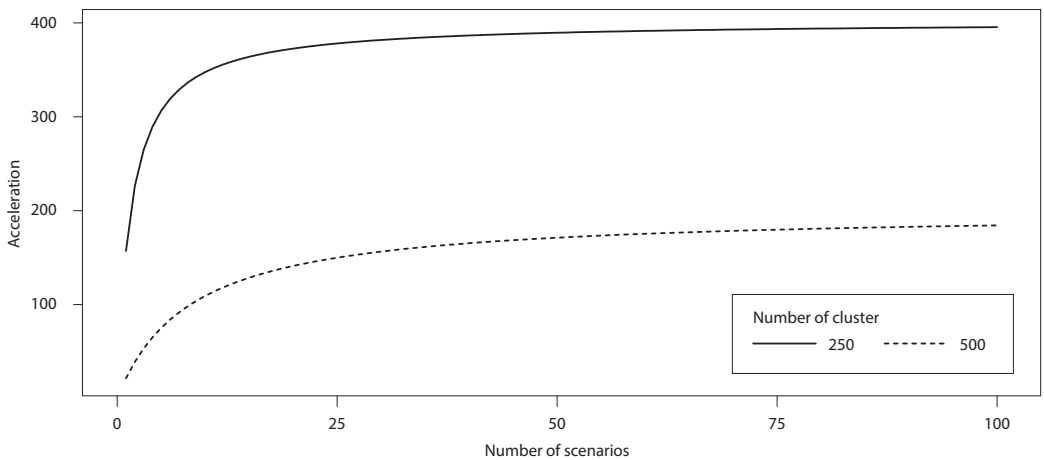
$$N^{+}_{Scenario} > \left\lceil \frac{T_{Clustering}}{T_{avg}\,(N_{Orig} - N_{Refer})} \right\rceil, \tag{20}$$

Table 3 Reference portfolio acceleration for one scenario

Number of clusters	Number of scenarios				
	1	**10**	**50**	**100**	**1000**
50	1 709.04	1 966.52	1 993.21	1 996.6	1 999.66
100	762.77	969.84	993.82	996.9	999.69
250	156.96	346.37	387.99	393.9	399.38
500	21.51	109.29	171.53	184.67	198.35
750	3.59	28.92	77.43	97.97	128.69
1 000	1.20	10.83	37.78	54.84	92.39
1 250	0.50	4.73	19.13	30.87	69.02
1 500	0.25	2.44	10.65	18.37	52.79

Source: Own construction

Figure 5 Comparision of acceleration for two clustering settings



Source: Own construction

For 1 500 clusters the clustering approach seems to be faster for at least 5 scenarios.

In section Accuracy and clustering time trade-off, it has been mentioned that a reasonable number of clusters based on error criteria should be between 250 or 500 clusters. Using the 250 clusters to replicate the original portfolio on a high number of scenarios (100 and more) is almost 400 times faster than modelling the original portfolio. Using the 500 clusters on the same number of scenarios is almost 200 times faster. This means that using the reference portfolio of 250 respectively 500 policies the analysts may test 400 respectively 200 times more scenarios in the same amount of time with a very high level of accuracy.

Figure 5 compares the acceleration between 250 and 500 clusters. The acceleration for 250 clusters dramatically increases when modelling a low number of scenarios and the growth slowly stabilizes after 20 scenarios. Using 500 clusters the acceleration grows slowly and doesa not stabilize so fast as using a lower number of clusters.

This task may be posted as an optimization task, where we search for maximum accuracy given the computation time available or minimum computation time for a given acceptable accuracy.

CONCLUSION

The proper valuation of the life insurance portfolio is one of the essential actuarial tasks. Traditionally used valuation techniques are based on modelling all policies of the portfolio which is time demanding. This takes effect, especially when valuating a high number of scenarios. Reducing the portfolio size in terms of the number of policies seems to be a good approach to speed up the computation time of the valuation.

Cluster analysis is one of the tools that can be applied to accelerate multiple scenario valuation of life insurance portfolio by reducing the size of the original portfolio into smaller reference portfolio. Results are on one hand obtained much faster as the per-policy projection is performed only for the reference portfolio. On the other hand, certain inaccuracy occurs as there is a difference between the projection results of the reference and the original portfolio.

The proper application of clustering approach requires the setting of several parameters such as selection of clustering variables and the suitable size of the reference portfolio determined by the number of clusters. The selection of clustering variables may increase the precision of the clustering approach. It can be advised to select clustering variables as the variables that the model should reproduce with the highest accuracy. The higher number of clusters may increase accuracy but also increase the computation time. When comparing the computation time, one has to include also the clustering time. The accuracy of the approximation is driven by the number of clusters used. An increasing number of clusters, on the other hand, increases both the clustering time as well as the valuation time of the reference portfolio. From our experiment, we may conclude:

1. The general level of error is relatively low.
2. An error of the approximation decreases with the increasing number of clusters at first relatively fast. At some point, the decrease slows down and continues steadily further at a slower rate.
3. For the high number of scenarios, the clustering time tends to be negligible, and the acceleration is proportional to the ratio of the size of the reference portfolio to the size of the original portfolio.
4. This means that the reasonable number of clusters is for our experiment somewhere between 250 and 500 as for the higher number of clusters, error rate decreases only slowly while as computation time increases rather fast and for the lower number of clusters the situation is opposite.

ACKNOWLEDGEMENT

The support of grant scheme Methods for fast estimation of life insurance liabilities with respect to different investment strategies IG410017 is gladly acknowledged.

References

CIPRA, T. (2014). *Financial and insurance formulas.* Physica Springer.

DEVALE, A. B., KULKARNI, R. V. (2012). Applications of data mining techniques in life insurance. *International Journal of Data Mining and Knowledge Management Process* (IJDKP), 2(4). <https://doi.org/10.5121/ijdkp.2012.2404>.

DICKSON, D. C., HARDY, M. R. (2013). *Actuarial Mathematics for life contingent risks.* 2nd Ed. Cambridge University Press, 616 p.

FOJTÍK, J. PROCHÁZKA, J. ZIMMERMANN, P. MACKOVÁ, S. ŠVEHLAKOVÁ, M. (2019). Alternative approach for fast estimation of life insurance liabilities. *20th Application of Mathematics and Statistics in Economics*, Sklarska Poreba, Poland. <https://doi.org/10.15611/amse.2017.20.11>.

FREEDMAN, A. REYNOLD, C., W. (2008). *Cluster analysis: A spatial approach to actuarial modeling* [online]. Milliman research report. [cit. 13.9.2021]. <https://bit.ly/2Uq48Kr>.

GIAMOURIDIS, D. SAKKAS, A., TESSAROMATIS, N. (2016). Dynamic Asset Allocation with Liabilities. *European Financial Management*, 23(2): 254–291. <https://doi.org/10.1111/eufm.12097>.

HEBAK, P. (2013). *Statistické myšlení a nástroje analýzy dat.* Informatorium.

JANDAGHI, G., MOAZZEZ, H., MORADPOUR, Z. (2015). *Life Insurance Customers segmentation using fuzzy clustering* [online]. Tehran: Faculty of Management and Accounting, Farabi College, University of Tehran. [cit. 13.9.2021]. <https://bit.ly/2Vyrga2>.

JANEČEK, M. (2017). Acceleration Techniques for Life Cash Flow Projection Based on Many Interest Scenarios – Cash Flow Proxy Functions [online]. *Czech Actuarial Society*. <https://bit.ly/2MBPtcb>.

KAUCIC, M., DARIS, R. (2015). Multi-Objective Stochastic Optimization Programs for a Non-Life Insurance Company under Solvency Constraints. *Risks*, 3(3): 390–419. <https://doi.org/10.3390/risks3030390>.

KRAH, A. S., NIKOLIČ, Z., KORN, R. (2018). A Least-Squares Monte Carlo Framework in Proxy Modeling of Life Insurance Companies. *Risks*, 6(2): 62. <https://doi.org/10.3390/risks6020062>.

MAECHLER, M., ROUSSEEUW, P., STRUYF, A., HUBERT, M., HORNIK, K. (2021). *Cluster: Cluster Analysis Basics and Extensions* [online]. R package version 2.1.2. <https://CRAN.R-project.org/package=cluster>.

MOHAMMED, M., YOUSSEF, B., TAOUFIQ, G. (2016). Time-Saving Approach for Optimal Mining of Association Rules. *International Journal of Advanced Computer Science and Applications*, 7(10). <https://doi.org/10.14569/IJACSA.2016.071031>.

NG, R. T., HAN, J. (2002). CLARANS: A method for clustering objects for spatial data mining. *IEEE transactions on knowledge and data engineering*, 14(5): 1003–1016. <https://doi.org/10.1109/TKDE.2002.1033770>.

NTEUKAM, O., JIAEN, R., PLANCHET, F. (2014). Internal model in life insurance: Application of least squares Monte Carlo in risk assessment. *Ekonomia. Rynek, Gospodarka, Społeczeństwo*, 41: 81–93. <http://dx.doi.org/10.17451/eko/41/2015/93>.

PURUSHOTHAM, M. (2016). Cluster analysis: Applications in experience analysis and assumption-setting [online]. *The Actuary.* <https://theactuarymagazine.org/cluster-analysis>.

SELIMOVIC, J. (2010). Actuarial Estimation of Technical Provisions' Adequacy in Life Insurance Companies. *Interdisciplinary Management Research*, 6: 523–533.

TURNBULL, C. (2014). Implementation of Least-Squares Monte Carlo (LSMC) in a Life Insurance Context – a Case Study [online]. *Moody's analytics risk perspectives, managing insurance risk*, 3. [cit. 13.9.2021]. <https://www.moodysanalytics.com/risk-perspectives-magazine/managing-insurance-risk/principles-and-practices/implementation-of-least-squares-monte-carlo-in-a-life-insurance-context>.