ALTERNATIVE APPROACH FOR FAST ESTIMATION OF LIFE INSURANCE LIABILITIES

JAN FOJTÍK, JIŘÍ PROCHÁZKA, PAVEL ZIMMERMANN, SIMONA MACKOVA, MARKÉTA ŠVEHLÁKOVÁ
University of Economics, Prague, Faculty of informatics and statistics,
Department of statistics and probability calculus, W. Churchill Square. 4, Prague 3, Czech Republic
email: xfojj00@vse.cz, xproj16@vse.cz, zimmerp@vse.cz, xsvem35@vse.cz,
simona.mackova@vse.cz

Abstract

Estimation of future liabilities is one of the essential actuarial tasks. With the huge client portfolios nowadays, not only the accuracy of liability estimates is of great importance but also the time within which the results are calculated. Especially in the case of estimates of life insurance liability, the computing time can be very high, because the estimation is based on a projection of future cash flow of each contract separately. Therefore, methods to reduce computation time while as do not significantly decrease accuracy are welcomed by many actuaries. Cluster analysis can be applied for this purpose. Basic idea is to split contracts into clusters and represent all contracts within the cluster by a specific contract, so called model point. Projection is only calculated for these model points and weights are assigned to reflect the number of contracts of the cluster. The main contribution of this paper consists in the analysis of clustering variables in the case of approximate life insurance liability model.

Key words: *life insurance, estimation of BEL, cluster analysis*

JEL Codes: *C38, C63, G22*

DOI: 10.15611/amse.2017.20.11

1. Introduction

Ordinary methods for valuing life insurance products and testing different scenarios are usually not complicated or mathematically sophisticated and specialized software designed for valuing life insurance products can be used see (Bacher 2010). However, despite the simplicity of the calculations, the high number of computing steps and the large sizes of the portfolios make the calculation very time-consuming and even with the newest technologies, the results may be derived with significant delay. Especially when testing a high number of investment strategies, see (Giamouridis, 2016) or (Kaucic, 2015). For example, let's assume a portfolio consisting of 100 000 different contracts. Estimating the value of liability and economic profit of such portfolio for one scenario can take approximately 5 minutes (depending on the complexity of cash flow model, software and hardware). Usually, insurance companies calculate hundreds or thousands of scenarios and obtaining the solution of 1 000 scenarios take several days or weeks. It is obvious that testing different scenarios by traditional approach is very time-demanding and the results derived with a significant delay may be outdated and may not follow current market information. The actuaries have usually only a few possibilities to reduce the computational time, such as purchasing faster hardware or better software, using more machines or testing fewer scenarios. These options may bring extra expenses or a loss of important information. Moreover, those solutions are only temporary and no single one solves the entire problem. Several researchers of statistics and

data-mining field have already been studying this issue see (Mohammed, et al. 2016). Freedman (2008) suggests that cluster analysis seems to be an alternative way for life product valuation.

In this paper, the use of cluster analysis for liability estimation and scenario testing of life insurance products are discussed. The main principle of clustering approach is to reduce the size of the original portfolio and create a smaller portfolio of reasonably selected representatives ('representative portfolio'), where each representative is assigned a weight reflecting the number of contracts that it represents. The liability estimate of the representative portfolio should approximate the result of the original large portfolio. Using smaller portfolio reduce computational time. Since the cluster analysis is only an approximate approach of real liability estimate, the reduction of precision of such estimate must be taken into consideration. An important task is the selection of clustering variables which are used to create the smaller representative portfolio. There are two main groups of clustering variables that can be considered. An obvious first choice of clustering variables is basic characteristics describing the contracts, such as the age of the client or the premium. It is presented in this paper that this selection of clustering variables does not need to lead to the best results. A better solution is obtained when the clustering variables are metrics of economic profits rather than basic client´s characteristics. The comparisons of liability estimates and time efficiency of both selection of clustering variables are presented in the results of this paper.

The paper is divided into three parts. The first section introduces the clustering methodology. It describes the main principles of clustering approach, selection of variables and precision measures. In the second section, the input data, cash-flow model and individual scenarios are presented. The result of analysis and model comparisons is summarized in the result section. The cash-flow model, as well as the whole study, is created in R software. The package used for cluster analysis is *Cluster see* (Meacher, 2017).

2. Methodology of clustering approach

In this section, the main idea of application of cluster analysis on life insurance portfolio will be explained. Furthermore, the variables selection for cluster analysis, the method of clustering, selection of clustering variables and the evaluation of results will be discussed.

2.1. Cluster analysis

The main principle of the clustering approach is to create several clusters, each cluster grouping contracts with similarities. Each cluster includes representative model point describing a group of model points within the cluster. Each of these representative model points is assigned the weight equal to the number of model points within the relevant cluster. The liability estimation based on the smaller portfolio (including only the representative model points) should reduce the computational time and lead to a high precision of the estimate.

The procedure of correct clustering approach can be divided into following steps:

- collection of model points and calculation of economic profit metrics;
- selection of the clustering algorithm;
- identification of clustering variables;
- selection of proper number of clusters;
- creating clusters – finding representative model points;
- creating representative portfolio and testing the precision;
- calculation of scenarios.

2.2. Clustering variables

Clustering variables are used to create suitable reference model points representing the original portfolio. *K* clustering variables are selected. $X_{i,k}$ is *k*-th clustering variable of *i*-th model point. The first and obvious choice is to cluster using basic model point characteristics such as age, premium or sum assured. The second choice is to use economic profit as PVFC or individual cash flows. Two approaches are discussed when selecting proper clustering variables:

- basic model point characteristic;
- metrics of economic profit of each model point.

The character of both variable selections is very different. In the first case, the variables bring only a descriptive character (position) of each contract. On the other hand, the variables of economic profit describe more the dynamics and development (trajectory) of each contract. Both approaches are demonstrated and compared in the results section. The clustering variables in both approaches are:

Basic model point characteristics	**Metrics of economic profit**
Age	PVFC
Annual premium	PVPL
Policy term in month	NPVPL
Duration in force	PVDE
Sum assured	PVPrem
Fund value	CF

The potential problem of clustering using economic profit as clustering variables consists in the wide dispersion of nominal values. Let's imagine two model points with different nominal values of cash flows but similar development. Using CLARA algorithm these model points may be treated as not-similar because of the wide range of their nominal values. In fact, this two model points should be similar because in the case of economic profit the development is the main aspect. The solution is to transfer nominal clustering variables of economic profit to the relative values. In order to provide such a transformation, individual clustering variables $X_{i,k}$ are adjusted by reference variable PVFC. The relative value $R_{i,k}$ of *k*-th variable of *i*-th model point is computed as

$$R_{i,k} = \frac{100 \cdot \left(X_{i,k} - PVFC_i\right)}{PVFC_i}. \tag{3}$$

Note that modified values of PVFC after using formula 3 are equal to zero. This variable is then omitted from clustering procedure.

2.3. Clustering algorithm

The clustering algorithm used in this paper is one of the centroid model algorithms. Center of each cluster is represented by medoids (i.e. by one specific model point). Most of the standard PAM like algorithms are inappropriate for clustering very large data sets, which is why Clustering Large Applications algorithm or CLARA is used as clustering algorithm (NG, 2002). Dissimilarities between each model points are measured by Euclidean distance. Euclidean distance between *i*-th model point MP_i and *j*-th model point MP_j is defined as

$$d\left(MP_i, MP_j\right) = \sqrt{\sum_{k=1}^{K}\left(X_{k,i} - X_{k,j}\right)^2}\ , \tag{4}$$

where $X_{k,i}$ and $X_{k,j}$ for $k = 1, 2, \ldots, K$ are values of clustered variables for *i*-th model point MP_i and *j*-th model point MP_j.

2.4. Precision measure

It is necessary to prepare error measures to quantify the precision of clustering approach. These measures compare the results of economic profit among both approaches and they are tested for all scenarios. The liability estimates calculated by traditional cash-flow approach are exact representation with no inaccuracy. On the other hand, the liability estimates of clustering approach are not exact representation and may include difference. The smaller the difference between results is the better is the approximation of liability by clustering approach. The error measure of *k*-th variable $error_k$ is calculated as a relative difference of results of both approached. The formula is:

$$error_k = 100 \cdot \left(\frac{approx_k}{real_k} - 1 \right), \tag{5}$$

where $approx_k$ is the result of *k*-th variable of clustering approach and $real_k$ is the results of *k*-th variable of traditional cash-flow approach. To compare multiple clustering solutions, the error is compared using the maximum, average and median value of an absolute value of an error.

3. Input model points

In this section, the input data with all supporting calculation as cash-flow model and metrics of economic profit are introduced. Also, several scenarios tested in this paper are presented in the last sub-section.

3.1. Input data

Input data includes model points representing a real insured portfolio of the life insurance company. The size of the portfolio is 106 524 model points, where each model point represents one policy contract with personal information about the client. Due to this fact, the adjustment to secure confidential information about the clients was applied on the whole portfolio. A sample of the insured portfolio is presented in Table 1.

There are four different types of product in the portfolio, A, B, C and D. The difference between the products consists of different settings of charges, surrender fees or surrender period. The second column gives information about the age of the insured person at the beginning of the contract. The Annual premium is the amount of money the policyholder is issued to pay for the contract. If the insured person is male, the value for sex is 0, for women this value is set as 1. Policy term in month is the length of the policy period in months. Duration in force shows how long the contract is in force. Sum assured is the amount of money paid to the client in the case of death of the insured person. The remains of premium after the application of charges and other fees are usually saved into clients account. This account can be seen in last column fund value.

Table 1: Sample of insured portfolio.

Product type	Entry age	Annual premium	Sex	Policy term in month	Duration in force	Sum assured	Fund value

C	22	2 321	1	588	252	93 745	53 927
A	30	2 629	1	444	288	117 024	61 930
C	25	2 335	0	552	192	138 408	46 170
A	41	7 955	1	384	204	233 706	128 731
B	43	3 610	1	300	252	32 681	0
C	19	3 631	0	684	300	207 905	102 791
C	39	4 122	1	264	204	101 248	40 148

Source: the author's work

3.2. Cash-flow model

Input data include only basic characteristics about clients but they do not say anything about the economic profit of contracts. There are plenty of metrics that can be considered as economic profit but almost all of them are based on cash-flow projection. We present several variables widely used in actuary practice:

- present value of future cash flows (PVFC);
- present value of profit and loss (PVPL);
- present value of premium (PVPrem);
- present value of distributive earnings (PVDE).

One of the very common metrics describing economic profit is the present value of future cash flow. To calculate this value the cash-flow model must be built first. Basic introduction into cash-flow model construction is described in (Cipra, 2014). Since the cash-flow model is applied on each model point, this procedure may be time-consuming in case the of many model points. The cash-flow model used in this analysis has the following form:

$$\mathrm{CF}_t = E\mathrm{Prem}_{t-1} - E\mathrm{Surrender}_t - E\mathrm{Death}_t - E\mathrm{Maturity}_t - E\mathrm{Commisions}_t - E\mathrm{Expenses}_t . \quad (1)$$

$E\mathrm{Death}_t$ is the expected value of benefit paid in case of death at time t adjusted by the probability of dying at time t. $E\mathrm{Prem}_{t-1}$ stands for the expected premium at the beginning of the period t, $E\mathrm{Surrender}_t$ stands for the expected surrender value at the end of period t, $E\mathrm{Maturity}_t$ stands for the expected value at maturity, $E\mathrm{Commission}_t$ represents the expected commission and $E\mathrm{Expenses}_t$ stands for the expected expenses. For simplicity, the model is built on an annual basis. Therefore, index t stands for the year of the projection.

Individual cash flows from formula 1 are used to calculate metrics of economic profit. For example, the present value of future cash flows for each model point is calculated as the sum of discounted expected cash flows over the policy period i.e:

$$PVFC = \sum_{t=1}^{n} (CF_t \cdot \prod_{k=1}^{t} \frac{1}{1+i_t}) , \quad (2)$$

where n is the number of policy years to maturity and i_t is the expected investment return at time t. A similar analogy is used for other metrics.

3.3. Sensitivity calculation

The sensitivity testing of liabilities and economic profit is an important aspect of an actuary work. Liability estimates are usually sensitive to the unexpected changes in various assumptions, e.g. different development of lapse rates, mortality rates, expenses or interest rates. Calculating different scenarios is supposed to explain the impacts of the assumption

changes on liability estimate. Therefore, the clustering model needs to provide a good approximation of liability estimate on all scenarios not only the best estimate. The results of clustering model are compared on seven different scenarios with the result based on the original cash-flow model to demonstrate the precision of the clustering approach. Each scenario represents shock on rates influencing the development of model points cash flows. The first scenario is set as the best estimate and the other six provide shocks on each assumption rate separately. Table 2 summarizes all scenarios. The portfolio of shocks was inspired by Solvency II directive (EIOPA, 2009), as a mandatory guide of risk analysis for insurance companies.

Table 2: List of scenarios with shocks.

Scenario	Shock
1 - Best estimate	No shock applied
2 - Mortality rate up	Mortality rate increased by 15%
3 - Mortality rate down	Mortality rate decreased by 20%
4 – Lapse rate up	Lapse rate increased by 50%
5 – Lapse rate down	Lapse rate decreased by 50%
6 – Interest rate up	Interest rate increased by 2pb
7 – Interest rate down	Interest rate decreased by 2pb

Source: the author's work

4. Results

In this section, results of clustering approach for two different sets of clustering variables, basic model point characteristics and metrics of economic profit, are compared. The results are first presented on the first scenario – best estimate and then on the other 6 scenarios. Table 3 summarizes errors realized for both the settings of clustering variables. The clustering is provided on 500 clusters.

Table 3: Results of best estimate scenario.

Statistic of percentage error	Basic model point characteristics	Metrics of economic profit
Maximum of absolute error	200.278	0.061
Average of absolute error	43.016	0.01
Median of absolute error	25.45	0.001

Source: the author's work

In the case of best estimate scenario, the results for clustering using metrics of economic profit are significantly better for all three error measures. A similar statement was confirmed by the results of the other six scenarios presented in Table 4 for basic model point characteristics and in Table 5 for economic profit. To simplify the output table, the results are presented on four variables, PVFC and individual cash flows for first three years.

Table 4: Precision of estimates based on basic model points characteristics.

Statistic of percentage error	Percentage errors			
	PVFC	CF_1	CF_2	CF_3
Scenario 2	-6.355	-4.253	103.332	-51.277
Scenario 3	-6.395	-4.984	108.718	-53.688
Scenario 4	-7.166	-6.083	69.577	-40.978
Scenario 5	-5.601	-0.477	195.192	-79.173
Scenario 6	-9.826	-4.763	103.677	-51.093
Scenario 7	-5.480	-4.558	105.582	-52.285

Source: the author's work

Table 5: Precision of estimation based on economic profit metrics.

Statistic of percentage error	Percentage errors			
	PVFC	CF_1	CF_2	CF_3
Scenario 2	-0.051	-0.208	-0.215	-0.362
Scenario 3	0.070	0.386	0.305	0.423
Scenario 4	0.112	0.092	0.129	0.159
Scenario 5	-0.066	-0.100	-0.314	-0.493
Scenario 6	0.013	0.038	-0.017	-0.081
Scenario 7	0.010	0.039	0.002	-0.034

Source: the author's work

Table 6: Number of clusters and precision of estimates based on economic profit.

Number of clusters	Statistic of percentage error	Percentage error	Calculation time of clustering procedure
10 clusters	Maximum of absolute error	195.134	0.12 hours
	Average of absolute error	19.525	
	Median of absolute error	4.618	
100 clusters	Maximum of absolute error	0.705	0.52 hours
	Average of absolute error	0.092	
	Median of absolute error	0.003	
200 clusters	Maximum of absolute error	0.013	0.73 hours
	Average of absolute error	0.004	
	Median of absolute error	0.099	
500 clusters	Maximum of absolute error	0.061	2.25 hours
	Average of absolute error	0.01	
	Median of absolute error	0.001	

Source: the author's work

The main reason for using cluster analysis is the reduction of calculation time. The liability estimate of the presented portfolio takes about 2.3 hours. Table 6 presents the results of calculation time with respect to the number of selected clusters and precision of the estimate. The computational time grows exponentially with the number of selected and the precision of the estimate is also better with a higher number of selected clusters. Therefore, there may need to be set up a good tradeoff between sufficient estimation error and computational time.

5. Conclusion

The paper presents cluster analysis as a useful tool for decreasing computation time of valuation of life insurance portfolios. Reduction of computational time may be particularly high when testing a large number of different scenarios. Insurance companies usually need to test hundreds or thousands of scenarios and the results may be derived after several days or even weeks. The results derived with such delay may be outdated and of low informational benefit.

One of the main aspects of clustering approach is to select proper clustering variables. We suggest using metrics of economic profit rather than basic contract characteristics. We show that in case of economic profit, the precision of clustering approach is very high. The error of model with an adequate selection of the number of clusters is lower than 0.1 percent. The more clusters are created the better the results of approximate approach fit the original portfolio. Due to the character of cluster analysis, the increase of the number of clusters leads to the increase of calculation time exponentially. Therefore, a good tradeoff between accuracy of estimates and computational time needs to be considered.

Cluster analysis is a method with a variety of algorithms. In this paper, only the CLARA algorithm is used. One of the topics for further research is whether other algorithms may lead to a higher quality of estimates with lower computational time.

Acknowledgements

The support of the grant scheme METHODS FOR FAST ESTIMATION OF LIFE INSURANCE LIABILITIES WITH RESPECT TO DIFFERENT INVESTMENT STRATEGIES IG410017 is gladly acknowledged.

References

[1] Bacher, U., Barkovic, D., & Dernoscheg, K. 2010. Actuarial Estimation of Technical Provisions' Adequacy in Life Insurance Companies. In Interdisciplinary Management Research-Interdisziplinare Managementforschung 6, pp. 523-533.

[2] Cipra, T. 2014. Financial and insurance formulas. Place of publication not identified: Physica Springer.

[3] EIOPA 2009: Solvency II Directive: DIRECTIVE 2009/138/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL, https://eiopa.europa.eu/.

[4] Freedman, A., Reynold, C., W. 2008. Cluster analysis: A spatial approach to actuarial modeling. http://www.milliman.com/uploadedFiles/insight/research/life-rr/cluster-analysis-a-spatial-rr08-01-08.pdf.

[5] Giamouridis, D., Sakkas, A., & Tessaromatis, N. 2016. Dynamic Asset Allocation with Liabilities. European Financial Management, 23(2), pp. 254-291. doi:10.1111/eufm.12097¨

[6] Kaucic, M., & Daris, R. 2015. Multi-Objective Stochastic Optimization Programs for a Non-Life Insurance Company under Solvency Constraints. Risks, 3(3), pp. 390-419. doi:10.3390/risks3030390

[7] Maechler, M., Rousseeuw, P., Et al. 2017. Finding Groups in Data: Cluster Analysis Extended Rousseeuw et al. Vienna: R Foundation for Statistical Computing, http://www.r-project.org/.

[8] 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). doi:10.14569/ijacsa.2016.071031

[9] 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. pp. 1003-1016.