ACCELERATION TECHNIQUES FOR LIFE CASH FLOW PROJECTION

Clustering approach

Jan Fojtík, Pavel Zimmermann, Jiří Procházka

18.9.2019

Introduction

Proper valuation of an insurance portfolio is one of the essential actuarial tasks. Common tasks, especially in life insurance, are based on the projection of future expected variables such as claims, premium, expenses, profit or many other for a portfolio under a set of different assumptions. Traditionally, calculations are based on projecting the cash flow per each contract in the portfolio. The advantage of such an approach is that there is no bias and each variable is projected in full scope. Contrary to the advantages, this approach is highly demanding on computation time and space. For example, let's assume an average-sized portfolio consisting of 500 000 contracts. Projecting such a portfolio for 50 years (600 months) for one scenario takes 1 hour. The number of variables to be calculated and stored is about 500. One scenario calculation then represents more than 150 x 10^9 (500 000 x 600 x 500) processes to be stored. Usually, the actuaries need to run multiple scenarios to obtain better information about the portfolio risk profile or from the reasons required by the new legislation. Therefore, running required hundreds or thousands of scenarios would then take a very long time – weeks or even months. Results derived with such a significant delay can be hardly used in practice. The space issue seems to be solved by using the high capacity storages but even with the newest hardware and optimized actuarial software, the long computation time has not been solved entirely. The valuation of multiple scenarios might be an unacceptably long and the derived results may be outdated or not reflecting the actual market situation. A faster approach reducing the computation time would be welcomed by many actuaries because it would allow running more scenarios and stress tests.

Several researchers have already been studying this issue. For example (Freedman, 2008) suggests accelerating the portfolio valuation by running only the selected policies instead of the whole portfolio. The selection of policies is provided by cluster analysis. A recent study (Janeček, 2017) suggests two solutions – proxy function and interpolation between scenarios. In this paper, we primarily discuss in detail the application of cluster analysis as a method to accelerate valuation processes in the life insurance business. This paper is designed as a tutorial presenting the different settings and options of the clustering approach.

The methodology of the clustering approach

The main limitation of the life insurance portfolio valuation is the computation time because a hundred thousand policies are projected individually. The idea of the clustering approach is to reduce the size of the portfolio in terms of a number of policies and preserve the high accuracy of its projection. Projecting only a few selected policies should lead to significantly lower computation time. The crucial aspect is to find such policies which reproduce the whole original portfolio with very high accuracy. The cluster analysis seems to be a suitable method to reduce the portfolio dimension in terms of the number of policies. The original sized portfolio is clustered into the smaller reference portfolio by selecting a few reference policies. The number of clusters defines the size of the reference portfolio.

Application of the clustering approach can be summarized in the following steps:

- selecting the clustering variables;
- defining the number of clusters;
- adjusting the data if necessary and reducing the portfolio size;
- building the reference policies and calculating the system of weights;
- measure the accuracy and reparametrize the clustering to improve the accuracy;
- improving accuracy by reparameterizing the model;
- applying the reference portfolio on different scenarios.

The cluster analysis used in the clustering approach is a general tool containing a variety of algorithms and parametrizations. In this paper, we present the basic application of the clustering approach and provide the analysts with step by step tutorial of the different parametrizations. Namely, we discuss:

- the necessary data transformation;
- the selection of the clustering variables;
- the selection of similarity measures;
- the definition of the suitable number of clusters balancing the speed with the accuracy;
- the selection of the system of weights to adjust the projection of the reference portfolio;

- the measures of the clustering approach success and precision;
- the application of the clustering approach to different scenarios.

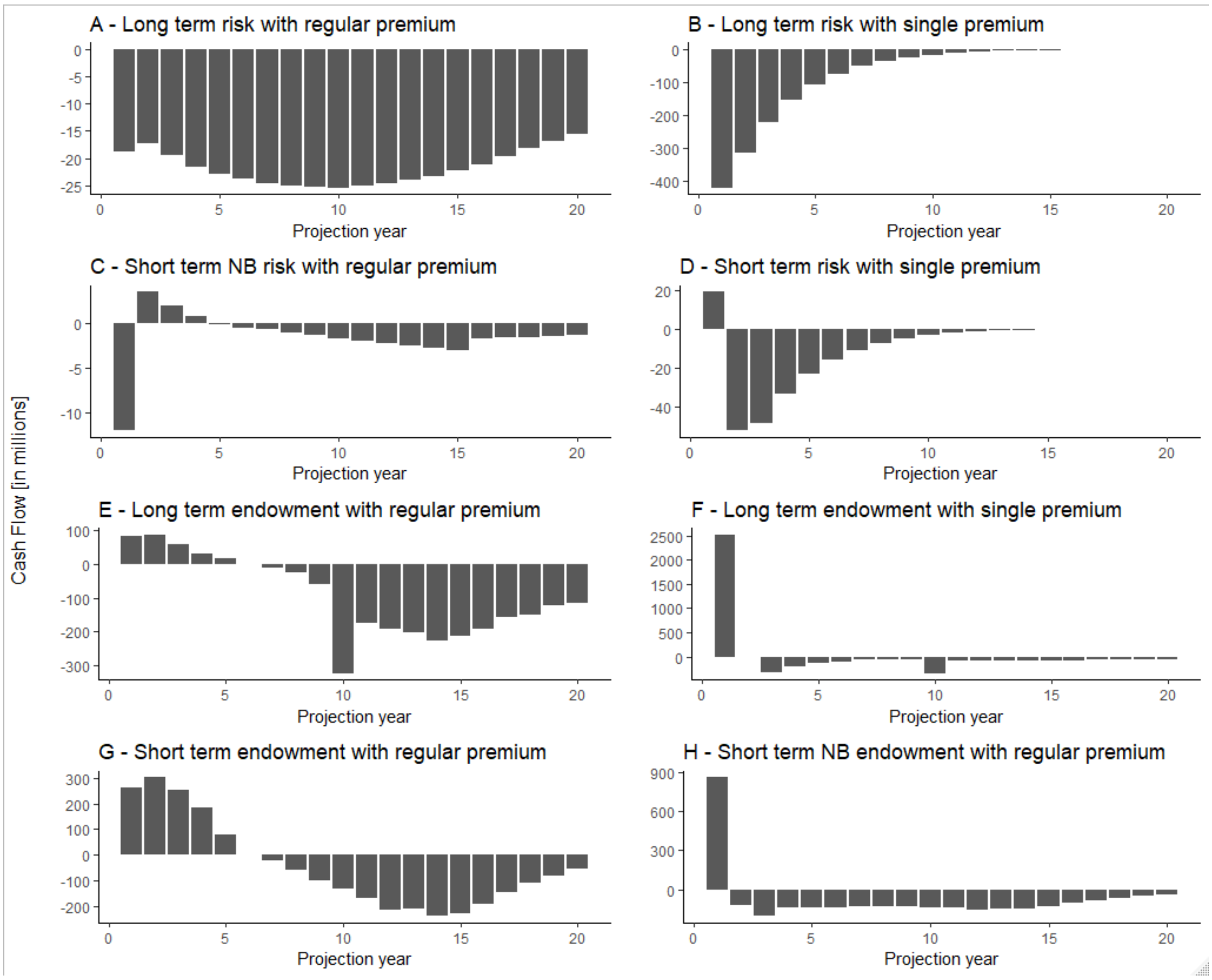
Demo Portfolio

Testing of the clustering approach is performed on two typical life insurance portfolios. The first portfolio consists of flexible Universal Life type insurance products and the other one consists of traditional endowment policies. Both portfolios have one hundred thousand policies. The heterogeneity of each portfolio is ensured by eight different policy products. Usually, we combine the long and short policy duration, regular and single premium, survival and death risk benefit type and availability of new business. The endowment is a product combining the death risk and the survival benefit. All eight products are summarized in table 1.

Table 1 Heterogeneity of the demo portfolio

Product	Premium frequency	Benefit	Duration	New business
A	Regular	Death	Long term	No
B	Single	Death	Long term	No
C	Regular	Death	Short term	Yes
D	Single	Death	Short term	No
E	Regular	Endowment	Long term	No
F	Single	Endowment	Long term	No
G	Regular	Endowment	Short term	No
H	Single	Endowment	Short term	Yes

Figure 1: Cash-Flow development of the heterogeneous portfolio



Both portfolios consist of eight different products equally distributed – 12.5 thousand policies each. In the further text, the demo portfolio will be referred to as the original portfolio because it represents the sized portfolio which we need to reduce. The cash-flow of the traditional (per each policy) approach is presented in Figure 1.

Clustering algorithm

The cluster analysis enables the application of several clustering algorithms to reduce the data dimension. For the purpose of the clustering approach, the nonhierarchical medoid based algorithm CLARA (Clustering LARge Application) (NG, 2002) is employed. The CLARA algorithm is the sampling approach suitable for handling large datasets such as the life insurance portfolio. In this paragraph, we present basic application of the clustering algorithm. The setting and detail parametrization including R code and analysis will be presented in further chapters.

To apply CLARA algorithm in R, the function *clara* from R package *cluster* (Maechler, 2019) is used. The package is available for free in CRAN. After downloading the package *cluster,* the import into R is called by the function *library* as:

```
#Importing the package cluster
library(cluster)
```

Generally, the portfolio reduction is ensured by grouping similar policies into the clusters. Similarity between the two policies is measured by a distance measure. The *clara* function allows using two types of distance measures for quantitative variables – Euclidean and Manhattan measures. The variable can be selected by parameter *metric*. An important aspect when calculating the distance measures is the scale of the variables. Some researchers suggest standardizing the data to make them more comparable (Romesburg, 2004). In this paper, we also prefer standardizing the data before clustering. The standardized score $Z_{i,m}$ of the m^{th} clustering variable $X_{i,m}$ and the i^{th} policy is given as

$$Z_{i,m} = \frac{X_{i,m} - \mu_m}{\sigma_m},$$

where μ_m and σ_m is the mean, respectively standard deviation of the m^{th} clustering variable. The standardizing is implemented into *clara* function under parameter *stand*. The number of predefined clusters is set by parameter *k*. This parameter also defines the size of the reference portfolio in output. The parameter *rngR* ensures that the clustering uses drawing random numbers, therefore each call returns different results. The basic parametrization[1] of *clara* function might be used in R as:

```
#Basic application of clara function
Clustering <- clara(   x = Portfolio[, clustering_variables],
                       k = NbCl,
                       rngR = TRUE,
                       stand = TRUE,
                       correct.d = TRUE,
                       metric = "euclidean",
                       pamLike = TRUE)
```

The output of *clara* function is in list format (specific R format). The position of reference policies in the original portfolio is hidden in *i.med* component. The final reference portfolio can be extracted as:

```
#Extracting the reference portfolio
Refer <- Portfolio[Clustering$i.med, ]
```

Reference portfolio and scaling factors

Projection of only a few reference policies will not replicate the projection of sized original portfolio with high accuracy. Therefore, the system of weights adjusting the projection of the reference portfolio must be introduced. For example, (Freednam, 2008) suggests using the number of policies in each cluster as a weight. This kind of weights ensures that the projected policy count of the reference portfolio matches the size of the original portfolio. The weight w^{Count} of the k^{th} reference policy is calculated as

$$w_k^{Count} = N_k,$$

where N_k is a number of policies assigned to the k^{th} cluster. Another option of weights is the scaling by a certain variable. In this paper, we present the scaling weight w^{Scale} based on the ratio of the best estimate of the liability (BEL). Selection of other scaling variables is also optional. Using this type of weight ensures that the reference portfolio replicates the BEL with no inaccuracy. The scaling weight of the k^{th} reference policy is given as

$$w_k^{Scale} = \frac{BEL_k^{Orig}}{BEL_k^{Refer}},$$

[1] For more details about *clara* function see the documentation of *cluster* package.

where BEL_k^{Orig} is the BEL of all policies assigned to the k^{th} cluster of the original portfolio and the BEL_k^{Refer} is the BEL of the k^{th} reference policy. The scaling weight can be obtained by:

```
#Assigning the cluster to each policy
Portfolio$clusters <- Clustering$clustering

#Calculating the total value of PV_CF per each cluster
Scale.clust <- (xtabs(get("PV_CF") ~ clusters, data = Portfolio))

#The scaling weight for each reference policy
Refer$weight_Scale <- Scale.clust/Refer[, "PV_CF"]

#The weight as the number of policies for each reference policy
Refer$weight_NbPol <- table(Portfolio$clusters)
```

The weighted projection of the m^{th} variable $\tilde{X}_{m,t}$ of the reference portfolio in time t is calculated as

$$\tilde{X}_{m,t} = \sum_{k=1}^{K} w_k X_{m,t,k},$$

where $X_{m,t,k}$ is the non-weighted projection of the k^{th} reference policy in time t and w_k is selected weight. The weighted projection of clustering variables can be obtained by

```
#Weighted projection of the reference portfolio - scaling weights
 Refer.Weighted_Scale <- colSums(Refer[, clustering_variables]*Refer[,
"weight_Scale"], na.rm = TRUE)

#Weighted projection of the reference portfolio - weights by policy count
Refer.Weighted_NbPol <- colSums(Refer[, clustering_variables]*Refer[,
"weight_NbPol"], na.rm = TRUE)
```

Precision

The precision is the measure defining how well the clustering approach replicated the original portfolio. The success of the clustering approach is determined by the percentage error defined as the ratio between the projection of weighted and original portfolio. The error of the m^{th} variable e_m is defined as

$$e_m = \frac{\tilde{X}_m}{X_m} - 1,$$

where $\tilde{X}_m$ is the weighted projection of the reference portfolio and the X_m is the projection of the original portfolio. This type of error is measured for each variable individually. The total clustering error e_{Total} of the whole portfolio is defined as the mean root of the sum of squared as

$$e_{Total} = \sqrt{\frac{\sum_m e_m^2}{M}}.$$

An important aspect to measure the accuracy of the replication consists of variables selection used to calculate the total error. It can be advised to measure the total error using the clustering variables only. The error and total error can be calculated in R as:

```
#Values of the original portfolio in total
Orig.Total <- colSums(Portfolio[, clustering_variables], na.rm = TRUE)

#Precision per each variable
Error.Scale[i, clustering_variables] <- Refer.Weighted_Scale/Orig.Total - 1
Error.NbPol[i, clustering_variables] <-Refer.Weighted_NbPol/Orig.Total - 1

#Total precision
Error.Scale[i, "Total"] <-sqrt(sum((Refer.Weighted_Scale/Orig.Total -
1)^2))/length(Refer.Weighted_Scale)
Error.NbPol[i, "Total"] <-sqrt(sum((Refer.Weighted_NbPol/Orig.Total -
1)^2))/length(Refer.Weighted_NbPol)
```

Clustering variables

The clustering variables represents one of the few factors determining the success of the clustering approach. The proper selection of the clustering variables may increase the precision of the whole clustering process. The importance of clustering variables takes place, especially when grouping similar policies into clusters because the similarity between two policies is based on clustering variables. In the following analysis, we present two types of clustering variables – **basic policy characteristics** and **economic projection**. A basic policy characteristic may be understood as general information about the policy such as age, premium, sum assured or value of reserve. The economic projection represents variables based on cash-flow projection such as individual cash-flows, profits or measures of profitability such as the present value of cash-flow, profit or premium.

The following analysis is applied to flexible Universal life and traditional portfolio separately. The target size of the reference portfolio is set to 500 policies. The similarity between clustering variables is measured by Euclidean distance. The weights represent the number of policies in each cluster. The analysis is performed for the following two types of clustering variables:

a. **Basic policy characteristics** – age at entry, policy period, annual premium, sum assured and reserve of the policy at valuation time.
b. **Economic values** – present values of profit, cash flow and premium. The expected values of cash flow, profit, premium, claims, expenses and commission in the first year and cumulated for ten first years.

Figure 2: Comparisons of different type of the clustering variables – Traditional portfolio

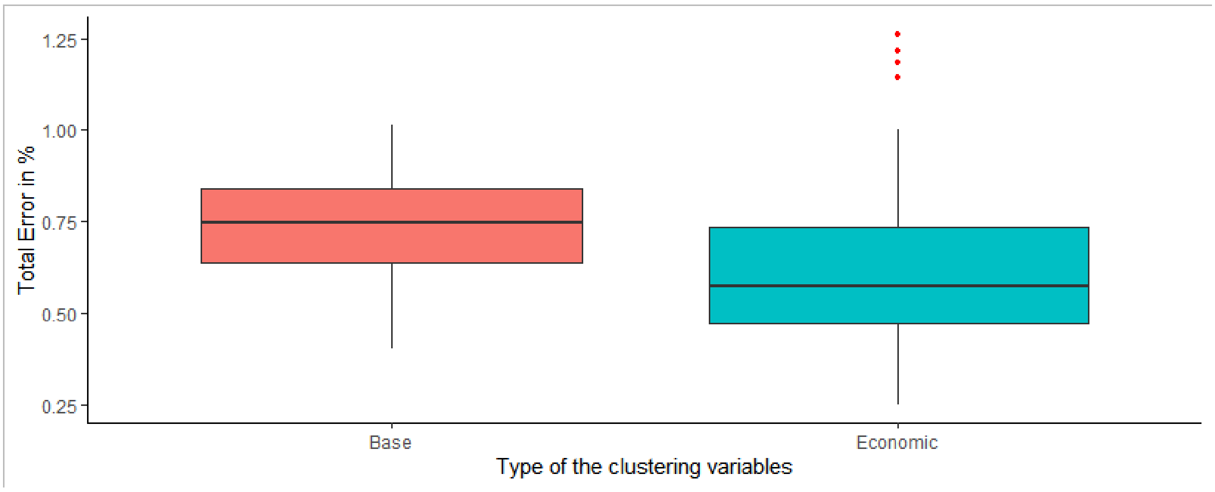
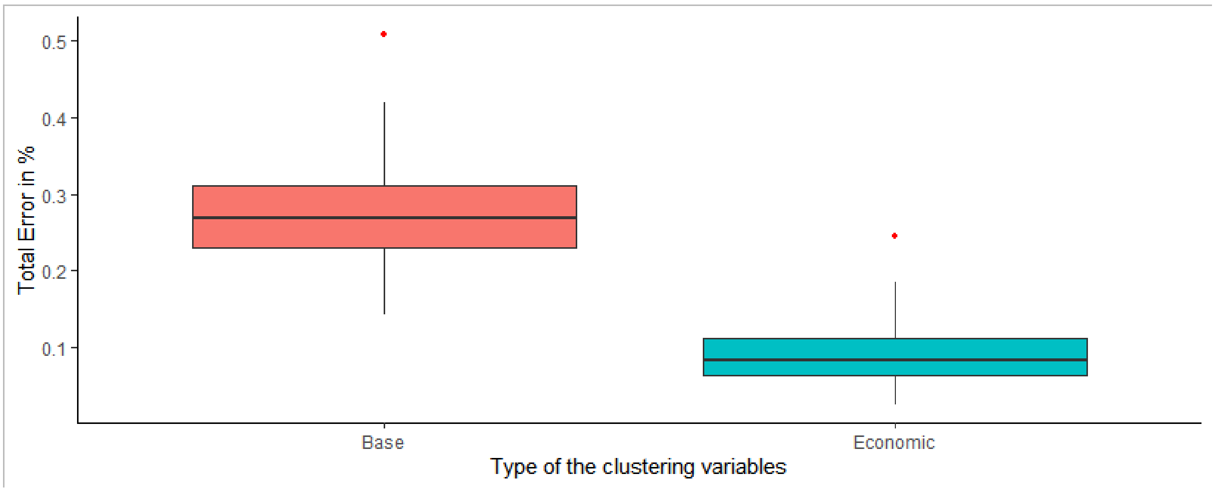


Figure 2 presents the distribution of the total error for both types of clustering variables in the traditionally based portfolio. The distribution is obtained by 100 re-clustering of the original portfolio. Figure 3 presents the same results for the traditional portfolio. Both types of clustering variables achieve a high accuracy but using economic variables seems to be more accurate. In general, it can be advised to select as clustering variables the variable which should be replicated with the highest accuracy.

Figure 3: Comparisons of different type of the clustering variables – Universal life portfolio



Similarity measures

Another factor determining the accuracy is the similarity measure. The function *clara* from package *cluster* allows measuring the similarity by two types of measures – **Manhattan** and **Euclidean** distance measure. The similarity can be understood as a process defining the distance between two policies based on comparing the variables.

The Manhattan distance measure $Manhattan_{i,j}$ is defined as the sum of absolute differences between *M* clustering variables of the *i*th and the *j*th policy as

$$Manhattan_{i,j} = \sum_{m=1}^{M} |Z_{m,i} - Z_{m,j}|,$$

where $Z_{m,i}$, respectively $Z_{m,j}$ is the *m*th standardized clustering variable. The Euclidean distance $Euklid_{i,j}$ is defined as the sum of squared differences between *M* clustering variables of the *i*th and the *j*th policy as

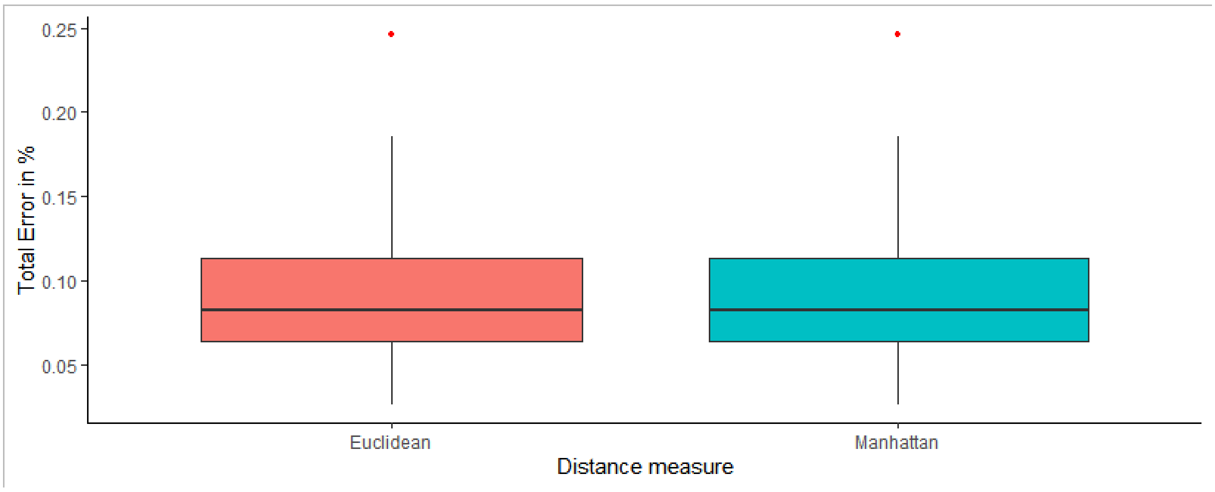
$$Euklid_{i,j} = \sqrt{\sum_{m=1}^{M} (Z_{m,i} - Z_{m,j})^2},$$

where $Z_{m,i}$, respectively $Z_{m,j}$ is the m^{th} standardized clustering variable.

The following analysis compares the distribution of the total error for both types of distance measure. The distribution is obtained from 100 re-clustering of the original portfolio. The target size of the reference portfolio is set to 500 policies and the weights represent the number of policies in each cluster.

Figure 5 presents the boxplot of the total error of both types of distance measure for Universal life portfolio. Analogically, figure 4 presents the total error distribution of the traditional portfolio. In both cases no significant effect between similarity measures was found. Also, there is no obvious reason to assume that one distance measure should lead to better results than the other one. For the purposes of the other analysis, the Euclidean distance measure will be employed.

Figure 5: the accuracy of the Euclidean and Manhattan distance measure – Universal life portfolio



Computation time and its acceleration

The main goal of the clustering approach is to speed up life insurance valuation with an acceptable level of inaccuracy. The faster valuation takes effect especially when testing a high number of scenarios or stress tests. The total computation time consists of two components – **clustering** and **projection** time. The projection time is required for calculating the portfolio projections and the clustering time is required for reducing the original portfolio into the smaller reference portfolio.

Let's assume that valuating one scenario of one policy by classical per-policy cash-flow model lasts in average time T_{avg}. The total computation time of $N_{scenarios}$ without any clustering of the whole portfolio of size $N_{policies}$ then lasts

$$T_{avg} N_{policies} N_{scenarios}.$$

In this case, the total computation time is equal to the projection time. But in case of the clustering approach, the total computation time of the same number of scenarios with the reference portfolio of size $N_{reference}$ then lasts

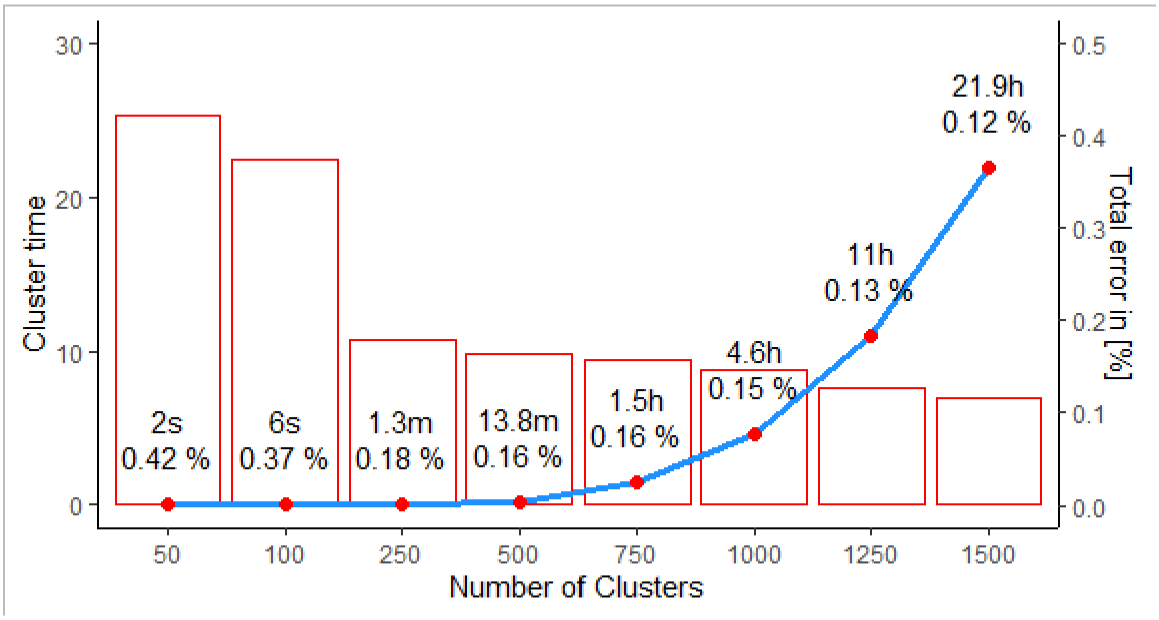
$$T_{clustering} + T_{avg} N_{reference} N_{scenarios}.$$

The total computation time consists of both clustering time $T_{clustering}$ and the projection time derived by modeling smaller reference portfolio. For simplicity, let's assume that the average projection time of one policy will remain approximately the same after the reduction. The acceleration of the clustering approach is given by

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

The acceleration of the clustering approach is determined by two aspects. The first one is clustering time $T_{clustering}$ and the second one is given by the size of the reference portfolio $N_{reference}$ (number of policies to be projected). Figure 6 presents the relation between the clustering time (blue line) and the total clustering error (red boxes) for the different number of clusters.

Figure 6: Precision and calculation time of the clustering approach with respect to the number of clusters

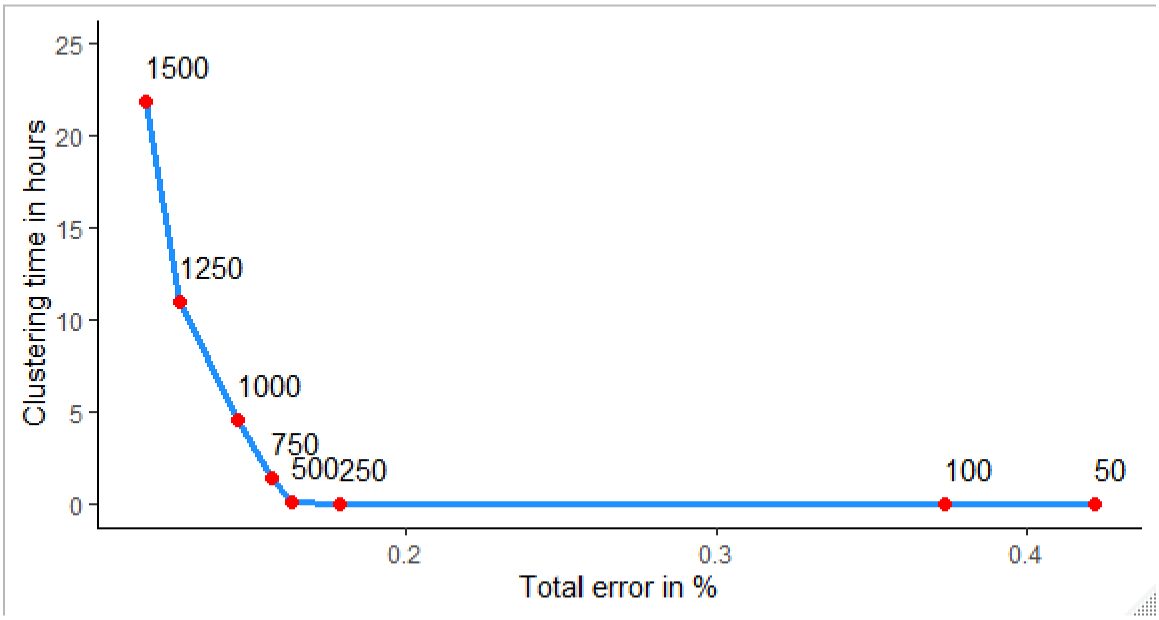


It can be concluded that the high number of clusters leads in slightly better precision but with higher calculation time. The calculation time also increased faster with the number of clusters thus some results may not be derived in a reasonable time. The number of clusters should be set in order to ensure fast computation time.

The number of clusters

The method defining the suitable number of clusters is based on a tradeoff between the clustering time, precision and computation time. For the purpose of this work, we used only 500 clusters. The reason can be seen in figure 7 which presents the tradeoff between accuracy and clustering time. 500 clusters represent a breaking point where the additional increase in calculation time is not compensated by any higher accuracy. The acceleration of 500 clusters is 200 (100 000 original policies divided by 500 reference policies). This means that 500 reference policies can be used to test 200 time more scenarios than by modeling the original sized portfolio with a very high level of accuracy.

Figure 7: Relation between accuracy and time of clustering approach



Replication of Universal-live and traditional life insurance

Generally, the clustering approach should cope with all types of insurance products. In the previous chapters, we presented the results of Universal life and traditional portfolio. The precision of the Universal life portfolio was in all cases higher than in traditional portfolio. Figure 8 compares the distribution of total error for both types of portfolio. The error distribution of each portfolio is obtained from 100 re-clustering into smaller reference portfolio of 500 policies. The weights are set as the number of policies in each cluster. The similarity is measured by the Euclidean distance between economic variables.

Figure 8: Comparisons of total error between both types of portfolio

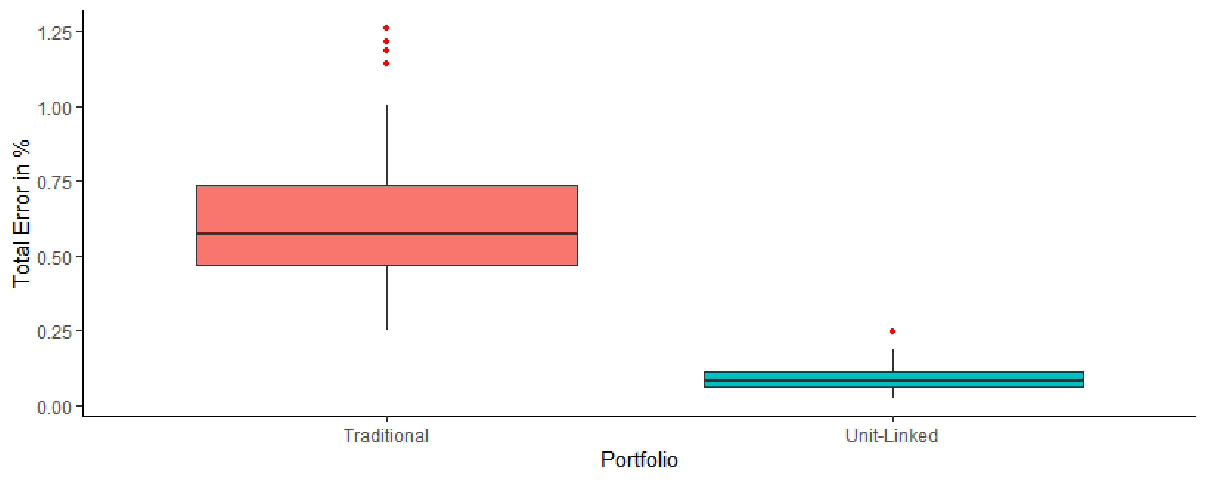


Figure 9 and 10 present precision of clustering variables in both portfolios. The precision of each variable can be considered stable and do not suffer from extreme values. It can be concluded that the clustering approach can be also used to approximate also certain variables.

Figure 9: Precision of the clustering variables in the traditional portfolio

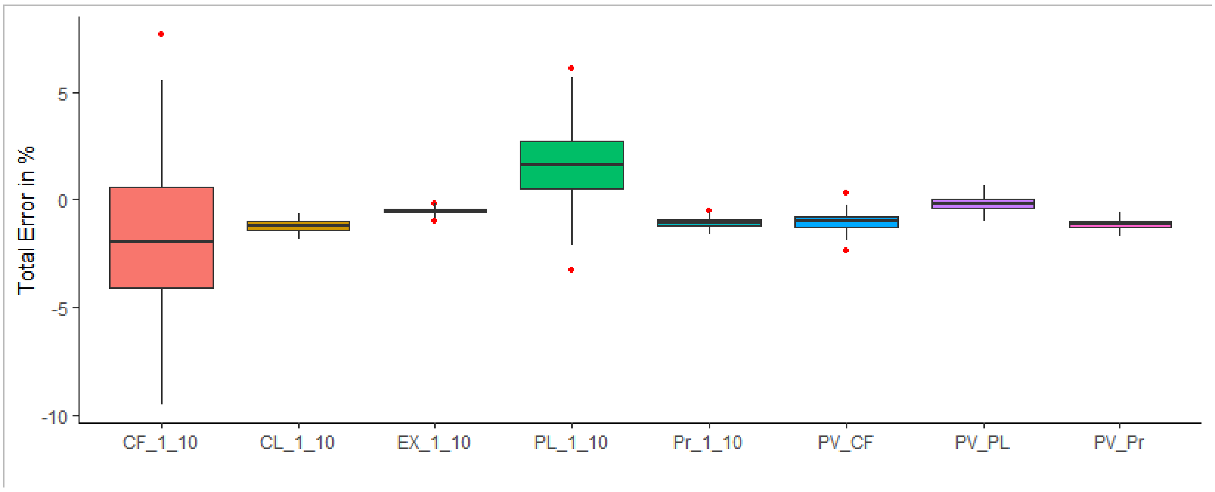
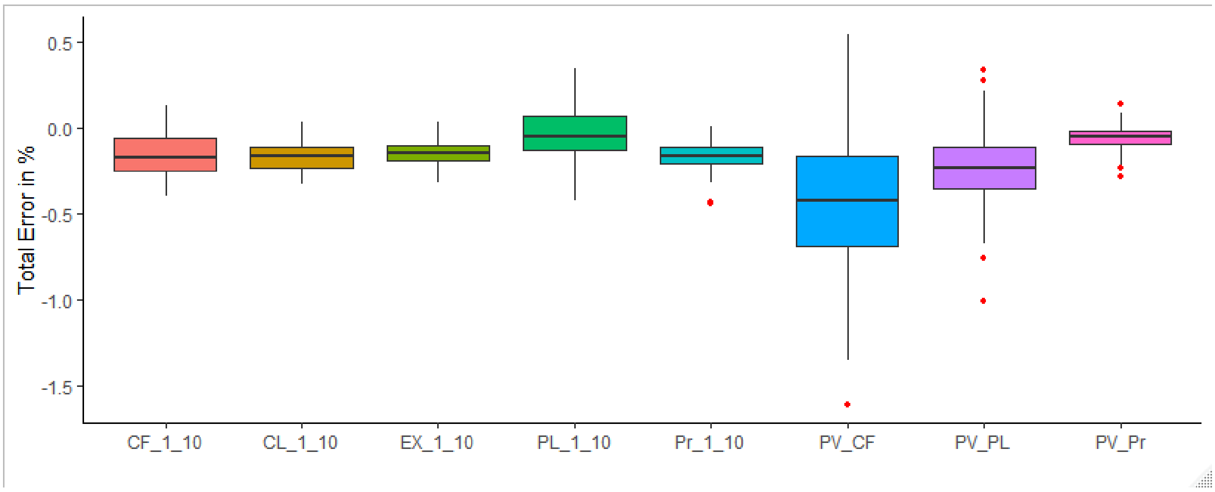


Figure 10: Precision of the clustering variables in Universal life portfolio



Replication of homogenous and heterogeneous portfolio

In the previous analyses, we presented results of the clustering approach only on the heterogeneous portfolio – portfolio with different combinations of products. It may be interesting to observe if the homogenous portfolio will be replicated with higher accuracy. We present eight homogenous portfolios where each portfolio consists of 100 000 policies of the same type presented in part Demo Portfolio.

On Table 1figure 11 and 12 we present the distribution of total error for each portfolio for traditional based and Universal life products. The error distribution of each portfolio is obtained from 100 re-clustering. The size of the reference portfolio is 500 policies. The weights are set as the number of policies in each cluster. The similarity is measured by the Euclidean distance between economic variables.

In both cases, the results are characterized by high accuracy and stability. The stability is an important feature of the clustering approach which concludes its application on a variety of different products without any deeper knowledge of the products or portfolio.

Figure 11: Eight different portfolios of traditional based products – total error comparisons

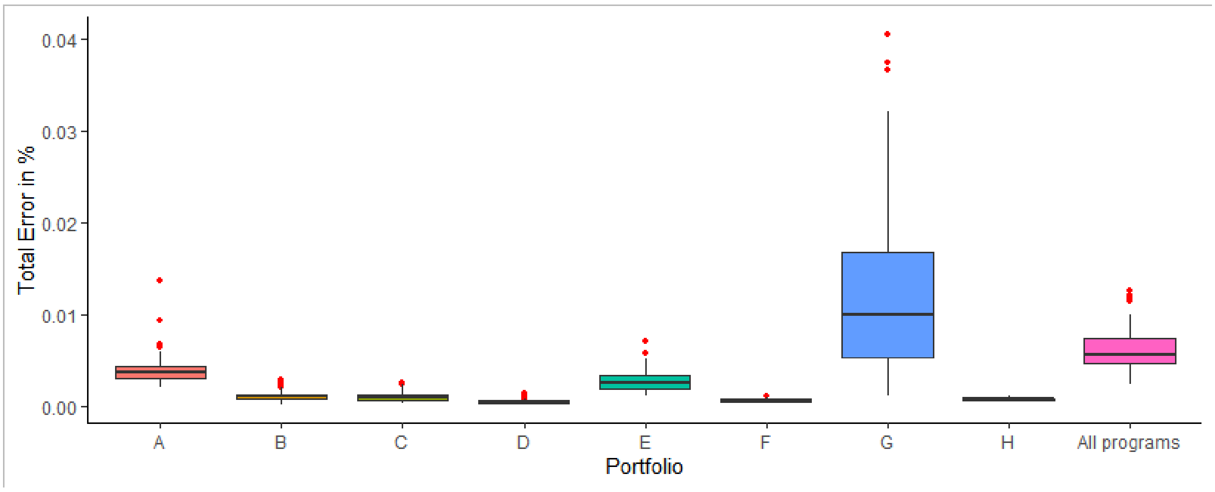
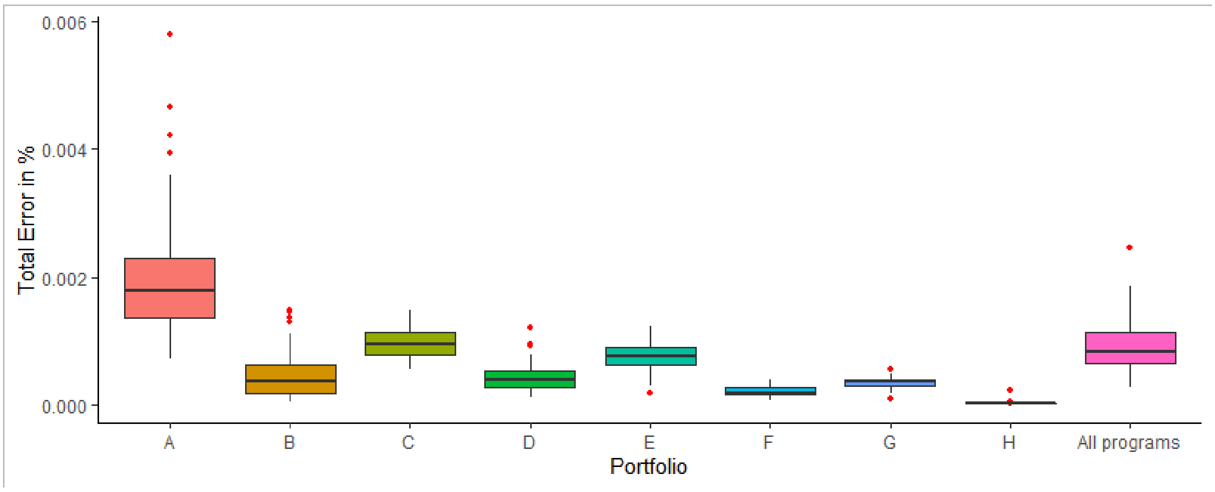


Figure 12: Eight different portfolios of Universal life products – total error comparisons



Application on different scenarios

One of the most typical actuarial tasks consists in testing the insurance portfolio under a different set of assumptions or investment scenarios. For instance, stressing mortality or lapse shocks are common tests for the purposes of Solvency II. Valuation under different investment scenarios finds its importance for managing the assets or liabilities. In this part, we present the usage of the clustering approach in terms of stress testing. We present the following stress tests:

- Lapse shock 15% up;
- Lapse shock 15% down;
- Mortality shocks 15% up;
- Mortality shocks 15% down;
- Invest rates 0% flat;
- Invest rates 2% flat;
- Invest rates generated from uniform distribution between 2% to 10% for each year of projection;
- Invest rates generated from uniform distribution between 4% to 10% for each year of projection;
- BEL – best estimate assumptions with invest rate 4% flat.

On figure 13 and 14 we present the distribution of total error for each stress test for traditional based and Universal life heterogeneous portfolio. The error distribution of each portfolio is obtained from 100 re-clustering. The size of the reference portfolio is 500 policies. The weights are set as the number of policies in each cluster. The similarity is measured by the Euclidean distance between economic variables obtained using best estimate assumption.

In this case, both results are characterized by high accuracy and stability. This confirms that the clustering approach may be also used for testing different scenarios of assumption or investment changes.

Figure 13: Traditional approach – total error comparisons

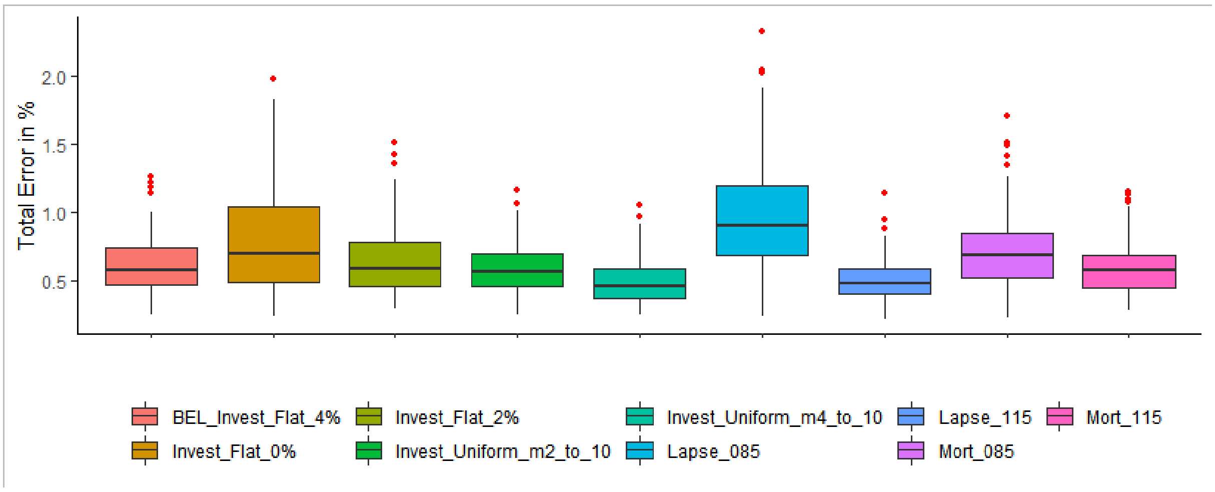
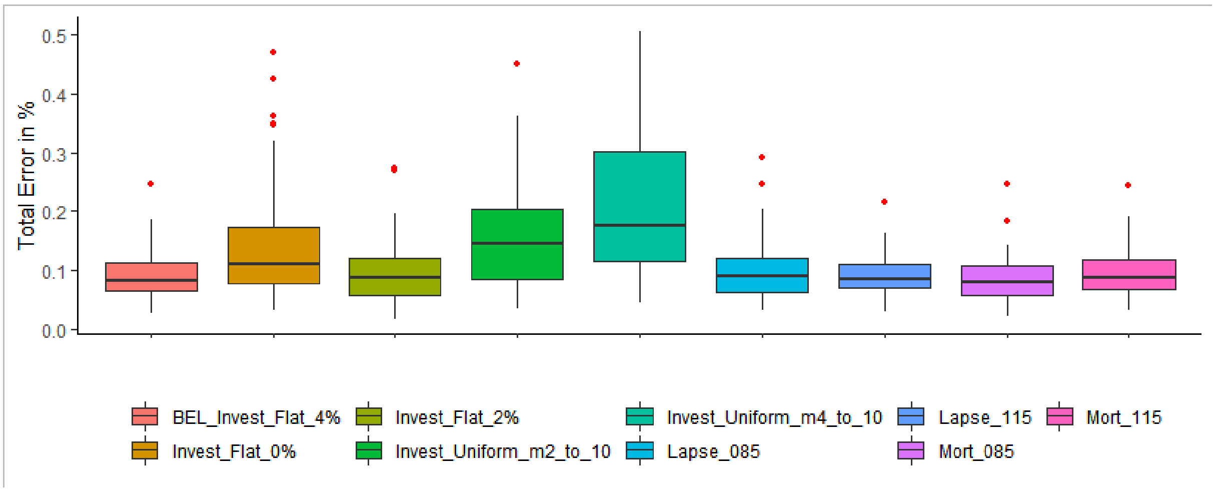


Figure 14: Unit link approach – total error comparisons



Summary

Cluster analysis is one of the tools that can be applied to accelerate multiple scenario valuation of the 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 the selection of clustering variables, the suitable number of clusters (size of the reference portfolio) or system of weight to adjust the reference projections.

The experiment was performed on artificial portfolio inspired by common life insurance products. Application on other portfolios may lead to different results. From our experiment, we may conclude:

Pros

- Stability and high accuracy over changes in the assumption or portfolio structure.
- Faster computation time.
- The clustering success depends on the proper parametrization such as clustering variables and the number of clusters.
- Clustering approach is designed to be easily automated on:

    - different type of life insurance policies;
    - heterogeneous or homogenous portfolio in terms of policy characteristics.
- Clustering approach can be used to valuate the portfolio under different Solvency scenarios or for the purposes of the asset-liability management.

Cons

- Any changes in the clustering process may be demanding on technical skills and statistical knowledge of the analysts.
- For the high number of clusters, the results may not be finished in a reasonable time – time grows faster with the number of clusters.

References

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

Janeček, M. (2017) Acceleration Techniques for Life Cash Flow Projection Based on Many Interest Scenarios – Cash Flow Proxy Functions, Czech actuarial society, Actuaria.org

Maechler, M., Rousseeuw, P. (2019). cluster: Cluster Analysis Basics and Extensions. R package version 2.1.0

Ng, R., & Han, J. (2002). CLARANS: a method for clustering objects for spatial data mining. IEEE Transactions on Knowledge and Data Engineering, 14(5), 1003-1016. doi:10.1109/tkde.2002.1033770

Romesburg, C. (2004) Cluster Analysis for Researchers. Lulu Press.