Enhancing aggregate reserving techniques through clustering-based segmentation

By Stephan Marais

Presented at the Actuarial Society of South Africa's 2023 Convention
Sandton Convention Centre 11–12 October 2023

ABSTRACT

Traditional aggregate reserving techniques assume that each data triangle used is a homogeneous group of claims. However, the practical constraints of management and reporting requirements often hinder the adoption of a statistically optimal approach to segmentation. There is often a reliance on legacy class hierarchies, business processes or expert judgement due to the difficulty of determining an optimal segmentation. This can reduce homogeneity in triangles and affect the performance of aggregate reserving techniques. To address these challenges, this paper proposes a new framework and methodology for automated clustering to allocate individual claims in a reserving class to homogeneous subgroupings suitable for triangle-based projections. The proposed approach involves the following:

— Separation of open claims from closed claims to prevent their unknown development from being a source of bias in the clustering process.
— Defining claim features that describe the development of individual closed claims to optimise the accuracy of reserve estimations produced by triangle reserving methodologies. Features are adjusted for trends, where necessary.
— Implementation of a clustering method on all closed claims for a given reserving class.
— Allocation of open claims to the clusters by finding the most similar closed claims clusters, based on operational time bands which reflect claim characteristics as they vary over time.

— A testing methodology to determine if the clustering method improves the accuracy of reserve estimates.
— Practical considerations for constructing models in a way that allows a dynamic segmentation method and for managing segmentation changes across reporting periods.

This paper includes an illustrative example of closed claims clustering employing the K-Means algorithm. This includes the calculation of claim features, determining the optimal number of clusters, and the application of Principal Component Analysis to reduce the number of variables explaining the distance between data points.

Overall, by providing both a methodology and a practical framework for leveraging clustering techniques as part of a reserving process, this research aims to improve accuracy and reduce the need for intervention of actuaries managing insurers' claims reserves.

KEYWORDS
Loss reserving, reserve segmentation, individual claims clustering

CONTACT DETAILS
Mr Stephan Marais; Email: stephan.marais@dyna-mo.com

1. INTRODUCTION

Aggregate triangle reserving techniques play a crucial role in the short-term insurance industry when estimating and managing claim reserves. These reserves are required by insurers to ensure financial stability, fulfil contractual obligations, and meet regulatory requirements.

Many of the current triangle reserving techniques rely on the assumption that each data triangle is a homogeneous group of claims or, at least, consistently heterogeneous over time. Most of the reserving methods have been developed by actuaries based in stable economies and were generally tested to be fit for purpose with established, stable books of business. However, this assumption does not always hold in practice.

With recent fast-moving global trends in claim experience, lines of business that have generally been stable have been more volatile in recent years. The actuary's ability to identify these trends is limited when working with triangulated data that summarises claims. Summarised claims data prevents the ability to identify underlying changes over time on a claim level.

Reserving actuaries are frequently required to apply judgement in making use of a wide range of established adjustment techniques for methods such as the basic chain ladder to the extent that it is considered standard practice. The Bornhuetter-Ferguson method is a common way actuaries address this problem of unstable claims history, but it can introduce a reliance on a subjective prior assumption of the Initial Expected Loss Ratio

(IELR) and poses the challenge of accurately allocating exposure to different origin year cohorts.

Segmentation aims to address the problem of changes in business mix over time, ensuring reserving is performed on groups of claims with similar development characteristics. There are various popular methods of segmenting claims in reserving processes to obtain more reliable reserve estimates.

Examples of these are:

— claim size,
— policy type,
— geographical location,
— industry, cover type,
— reinsurance structure,
— risks managed together,
— distribution channel, and
— peril.

Adopting such segmentation methods can result in segments of data leading to data-driven assumptions that are not statistically significant. A common example of this is splitting a class by claim size (attritional and large) resulting in a large claims triangle with no data for certain origin years, forcing the actuary to rely on performing manual adjustments to the calculated patterns.

The difficulty in achieving an optimal segmentation arises from various factors, including the complex nature of insurance claims data and the inherent uncertainty in claim development patterns. Identifying and grouping claims with similar characteristics is critical for accurate reserve estimations using methods such as the basic chain ladder. However, the lack of a systematic approach to segmentation can result in suboptimal allocations, potentially leading to bias in reserve estimates.

Additionally, the practical constraints imposed by management and reporting requirements often limit the adoption of statistically optimal approaches to segmentation. Actuaries responsible for estimating claim reserves face challenges in determining the most suitable segmentation method. As a result, there is often a reliance on legacy class hierarchies and previously established business processes. This reliance on traditional methods of segmentation leads to reduced homogeneity within data triangles over time, potentially compromising the reliability and performance of aggregate reserving techniques. This increases the reliance on expert judgment, manual reserving calculations and adjustments even further.

To address these challenges, this paper proposes a new data-driven approach to segmentation for aggregate triangle reserving models leveraging clustering techniques on individual claims.

Clustering methods offer a data-driven approach to identify natural groupings within

data, allowing for the creation of sub-groups that are more homogeneous than the aggregate dataset whilst ensuring sub-groups are of sufficient size so that they remain statistically significant.

This paper describes an approach to feature engineering suitable for grouping claims data in a way that characterises the development pattern of an individual claim.

By automating the clustering process and incorporating relevant claim features, this approach aims to enhance the accuracy of reserve estimations and optimise the management of insurers' claim reserves.

This paper contributes to the field by offering a practical framework that addresses the limitations of existing reserving methods by bridging the gap between traditional aggregate reserving techniques and the statistical advances in clustering algorithms.

The remainder of this paper discusses other related literature and presentations in Section 2. Section 3 presents the new proposed methodology and framework enhancing aggregate reserving models with cluster-based segmentation, followed by a conclusion and discussion of initial findings in implementing this framework in Section 4. An example of implementing this framework is described in Appendix A.

2 LITERATURE REVIEW

2.1 Limitations of homogeneity assumption

Zehnwirth (1989) points out that the chain ladder technique only works well when development factors are homogeneous and raises concerns about the suitability of the homogeneity assumption to real-world scenarios. He also reports that long-tail classes are often described by reserving actuaries as heterogeneous with trends over time being the main cause for this. Consequently, the fact that the assumption of homogeneous development factors is frequently violated in the real world presents a challenge when applying chain ladder techniques.

2.2 Clustering-based segmentation approaches

One way to address the heterogeneity in development factors is to segment claims into groups that are more homogeneous. Avitabile and Cooke (2018) propose clustering as an approach for reserving segmentation that is "data driven, credible and not prone to biases". In this presentation they highlight the need for segmentation and show how it can be achieved through clustering loss development patterns by considering their error distributions. Importantly, they calculate a "stability score" by performing many stochastic simulations and clustering each simulation. This stability score can be used as a diagnostic tool to gauge confidence in selected clusters.

Clark and Jiang (2018) discuss the application of clustering algorithms in loss development analysis. They explore how the K-Means algorithm, Principal Component Analysis (PCA)and Sherman curves can be used to identify patterns and group similar classes by looking at the loss development patterns. They conclude that these methods

enable reserving actuaries to transcend from only considering groupings of data to also identify the variables driving claims development.

Yeo et al. (2001) propose a clustering technique for predicting claim costs in the automobile insurance industry. They compare prediction accuracy when classifying policyholders into risk groups using a clustering algorithm and then predicting the claim costs for each group versus using only heuristic classifications. It is found that the algorithmic clustering approach yields better predictions of claim cost than when using heuristic groups, partially because a clustering approach can consider more variables simultaneously without significantly increasing the number of risk groupings. While the paper focuses on risk rating in the automobile insurance industry, learnings from the clustering technique can be applied to other short-term insurance fields and models.

John (2018) applies clustering methods to help decide on what basis to create segments for a worker's compensation class when performing chain ladder reserving. It was found that using triangles for each group instead of using an aggregate triangle yielded better results for two out of the three segments identified. Hence, it is concluded that more granular assumptions should be used when they bring actual versus expected results closer, otherwise aggregate modelling should be used. Thus, clustering approaches could potentially be integrated and used alongside traditional reserving methods instead of completely replacing them.

While Yeo et al. (2001) showed how clustering could be applied to individual claims and John (2018) showed how clustering could be used to help in choosing which segments to use in a reserving model, this paper proposes a new framework for setting reserving segments through clustering of individual claims.

3. METHODOLOGY

3.1 Framework and methodology overview

This section describes the proposed framework of how clustering can be used to segment a class of business, broken down into five steps. The key steps are stated below and explained further in the following sections.

— Step 1: Separate datasets for open and closed claims (see Section 3.2).
— Step 2: Calculate appropriate claim features on closed claims (see Section 3.3).
— Step 3: Implement a clustering methodology on all closed claims (see Section 3.4).
— Step 4: Allocate open claims to the closed claim clusters by finding the most similar closed claims clusters (see algorithm in Section 3.5).
— Step 5: Build triangles for each cluster and treat them as reserve segments in a standard aggregate triangle reserving model.

3.2 Separation of open and closed claims

Clustering open and closed claims together can cause bias in the open claims cluster allocations if there are features that are not completely known at the report date of the claim.

For example, the total open time of a claim is not known for a claim that is still open. If the variable for an open claim is calculated as the total time this claim has been open to date, clustering it with closed claims can incorrectly group this open claim to a cluster of closed claims which were open for a short period. In this case, it would be better to compare the open claim with other claims variables calculated as if they were open for a similar duration. See Section 3.5.

The definition of a closed claim in this paper is a claim which has been fully paid and is not expected to have any more future development.

It is common to adjust data as a way of dealing with missing or incomplete data in feature engineering, for example by replacing missing values with the mean of the dataset. While this approach is easy to implement, it has obvious flaws for the purpose of a reserving exercise. Reserve and incurred but not reported (IBNR) estimates are largely driven by the open claims present in a triangle. Inserting a value such as the mean or applying other forms of regression to complete open claims data can cause all these open claims to be clustered with a bias to be grouped in a certain cluster which is closest to the entire dataset's mean values for the missing information.

This framework suggests that open claims should not be clustered along with closed claims. This way the assumption can be made that closed claims have a completely known development pattern and hence features that describe the full history of the development pattern can be defined (see Section 3.2).

3.3 Claim feature engineering for clustering

Feature engineering is deemed the key to success in any machine learning or data analytics model. In the context of clustering algorithms, it refers to the process of selecting and transforming the relevant features or variables from the raw data to better represent the data in a way that is desired for the clusters to be formed. These features are then ingested by the algorithm to allow it to cluster the claims in a way that is suitable to achieve better homogeneity in actuarial reserving processes.

For insurance claims data, it involves identifying and creating informative features that capture the underlying patterns and structures within the data based on the characteristics of interest. The structures and characteristics of interest may differ from company to company and between classes of business.

As mentioned in the presentation by John (2018), claim characteristics such as the type or degree of injury can be used as features in a workers' compensation class (also known as employer's liability). The use of such categorical measures of claim characteristics can be constrained by the level of detail available in the data. Separating claims by such characteristics also doesn't necessarily separate heterogeneous trends effectively. For example, both bodily injury and damage to property in a liability class might have similar legal cost trends.

This paper proposes that features should also be calculated from individual claim

transaction data used to build triangles. The benefit of this approach is that it can be applied to any class of business without requiring information other than the individual claim movements which are already available when building triangles. Examples of features that could be calculated on claim transactions directly:

— A measure for loss development factor on an individual claim, such as the final claim size divided by the initial case estimate as per the report date.
— Macaulay Duration calculated on incurred or case estimate movements. This is the total incurred or case estimate movement weighted average term from the origin date. The amounts can be adjusted for inflation.
— The total number of days a claim was active for.

The aim of this proposed framework is to segment a class through clustering to produce a more reliable reserve estimate using aggregate triangle reserving techniques. These methods, such as the chain ladder method, produce more reliable projections when the underlying claims in the triangle develop in a similar pattern. This means that **features used in clustering should aim to describe the development pattern of claims as closely as possible**. Claim features that are drivers for claims development can also be included in the method but should not be considered the main variables used for clustering. A feature which could be a driver of a claim's development pattern is its initial reported case estimate size. For example, large claims are often handled differently by claim handlers or are covered as part of a reinsurance treaty which means they follow different admin processes that can result in different payments or development patterns compared to attritional claims.

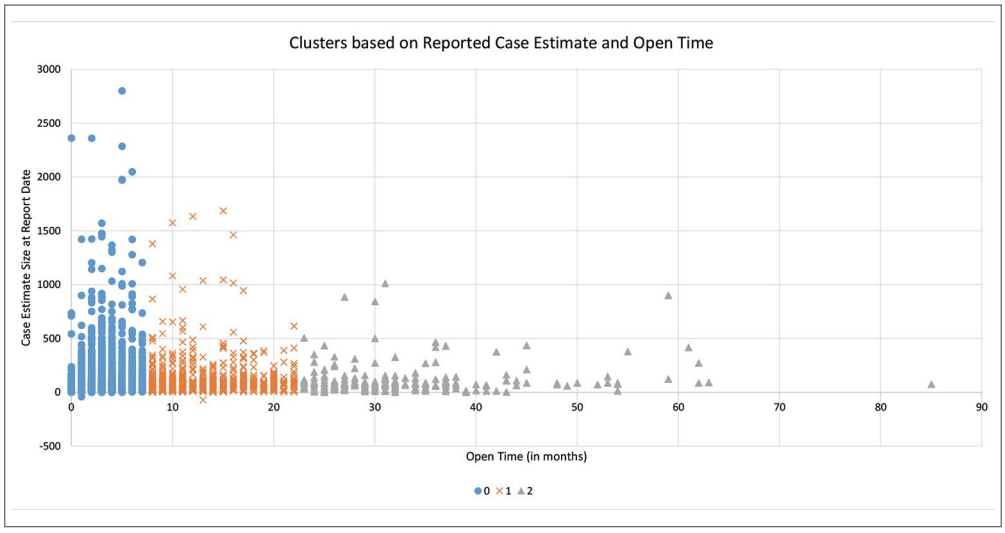


FIGURE 1 This graph illustrates how claims can be clustered by two features into three groups, active time in number of months and reported case estimate size.

Features are only used in the clustering process rather than in the reserving process itself. This means that amounts can be adjusted or weighted without affecting the integrity of the reserve estimate. An example would be to adjust claim amounts for inflation, or to manipulate claim transaction data during the period of a pandemic, such as Covid19, as it is used in feature calculations. This is common in experience rating processes for pricing where the historical experience data are adjusted to be comparable to the current valuation period.

3.4 Implementation of clustering method on closed claims

The framework that this paper proposes is not prescriptive on which clustering technique should be used on the closed claims. There is a wide range of techniques available such as K-Means, Mean-shift, Spectral Clustering and K-Nearest Neighbour (KNN) which all have different strengths and limitations.

The choice of suitable features and clustering method used depends on the nature of the claims, quality and level of information available in the data. This is an exercise which is familiar in the field of data science. The clustering of closed claims can be done by employing a wide range of available techniques. Further research can be done to establish optimal ways of clustering these closed claims for various types of classes.

It is important to note that separating closed from open claims does not imply that the method assumes that closed claims cannot be reopened. The aim is to simply separate claims which have sufficiently developed such that features that assume a claim's full transaction history is available can be calculated.

See Appendix A for an example of a closed claim clustering implementation.

3.5 Allocation of open claims

This paper suggests an algorithm for allocating open claims to existing closed claim clusters in operational time bands. Each open claim is allocated to a closed claim cluster by finding the most similar characteristics to the existing closed claim clusters. It is like the K-Means clustering algorithm in the sense that it chooses the best cluster by minimising the sum squared error (SSE) of features calculated in operational time.

The term "operational time" is used in the context of individual claims as defined in Taylor et al. (2008). The concept of operational time was introduced into the actuarial literature by Buhlmann (1970), and first applied to loss reserving by Reid (1978).

Using operational time bands (OT bands) means that movements are grouped according to the proportion of claims settled at each point in development as opposed to the number of financial periods since origin as with triangles.

3.5.1 THE METHODOLOGY

First, choose suitable operational time bands, e.g., the 10th, 20th, . . . , up to 90th percentiles. These percentiles will represent the operational time bands used in the allocation method.

For each operational time band, apply the following steps:

1. **Determine the number of days representing the boundaries of the current operational time band.**
   This can be done by ordering closed claims by the number of days they were open and finding the claims at the given claims band's percentile.
   For example, 20 and 42 active days might represent the 10th and 20th percentile in the closed claim dataset. This range of open time represents the 10%–20% operational time band.

2. **Find all open claims which fall within the lower and upper bounds of the band.**
   For example, there might be 150 open claims that have been open between 20 and 42 days and are now allocated to the 10%–20% operational time band.

3. **Find relevant closed claims to compare with open claims.**
   This is done by searching for all closed claims in the dataset which were open for the same duration of time relative to the operational time band. All movement data beyond the upper bound needs to be removed. This way it is effectively assumed that the closed claims are observed as if they were open for the same duration as the open claims in the current OT band have been open for.
   For example, there might be 300 closed claims that were open for at least 20 days and remove all information beyond 42 days.

4. **Calculate appropriate features for the claims in operational time.**
   Features calculated on open claims (Step 2) and the closed claims in (Step 3) as per the age of the given operational time band are now comparable since they are the same age.
   Because the claims at this point in operational time are still considered open, appropriate features to compare them are likely to be different to the features used in the closed claims clustering process (as per Section 3.4) since some pieces of information are now missing or not relevant anymore. For example, the total number of open days cannot be calculated because, at this point in the OT band, it is not known when claims will be closed since we removed this information in Step 3. Considering how long a claim has been open to date as a feature is not useful because all claims relating to the same operational time band have been open for the same duration. An example of a new useful feature to include is the ratio of 'paid to date' to 'incurred to date'. This ratio for a closed claim is always expected to be 1.

5. **Standardise all the feature data.**
   This can be done by subtracting the mean from each feature data point calculated in Step 4 and dividing by the standard deviation. Open and closed claims will have to be combined in determining the mean and standard deviation.

6. **Calculate the new centroid of each closed claims cluster.**
   Group the closed claims feature data by cluster numbers allocated through the closed claim clustering process (Section 3.4). Calculate the new centroid for each cluster number based on the new standardised features in operational time for that cluster. The centroid is simply calculated as the mean of each feature.

7. **Allocate open claims to the nearest closed claims clusters centroids.**
   For every open claim, calculate the distance to each closed claim centroid. The distance between the open claim and the centroid can be calculated as the sum squared error (SSE) of the standardised feature data in operational time. Choose the cluster number which minimises SSE.

All the open claims should have an allocated cluster number after the algorithm has been repeated for all the operational time bands.

3.6 Testing methodology

The aim of the testing should be to see whether adding clustering-based segmentation addresses the limitations of aggregate reserving techniques and improves the accuracy of the reserve estimates when no manual adjustments have been made to the method.

The suggested approach to measuring accuracy in reserve estimates is to perform back-testing on an algorithmic reserving model based on the aggregate triangle reserving method of choice that requires no manual intervention by the actuary. For example, using the basic chain ladder method and applying some basic algorithmic smoothing methods to the patterns calculated based on the entire triangle data, such as curve fitting or automatic exclusions of outliers.

Since the projection estimates the best estimate reserve, the projected lower triangle resulting from segmented classes can be summed up and compared with the reserve estimate produced by the aggregate class triangle and back-tested on the data.

Back-testing involves creating a smaller subset of the total claims by removing transaction data from the most recent calendar periods and fitting the reserving model to the reduced triangle. This fitted model is then used to project into the future and these projected values (expected) are then compared to the removed (actual) data in recent calendar periods. As described by Balona and Richman (2020), this process can be done iteratively, considering single calendar periods at a time.

It is of interest to identify which model and parameter selection minimises the root-mean-square error (RMSE) of the predicted versus the actual values.

If employing a clustering-based segmentation approach significantly improves the performance of the reserving model, it is likely that there are some interesting trends present in the individual claims data.

The clustering process can be extended to identify obvious trends in the data by setting

up visual charts of various pairs of features plotted together and grouped with the same colour by cluster number.

The actuary should be able to see patterns according to how clusters were formed. For example, a specific cluster might be grouped such that claims have a longer reporting delay than the other clusters.

Charts of the incurred, paid and reported number triangles per segment (formed from the cluster) can then also be analysed to see if the basis for forming the cluster has trends throughout calendar periods. For example, it might be that a segmented triangle (formed by a specific cluster) has a larger number of reported claims in recent accident periods. This could be an indication of a changing trend over time, i.e., that insured risks have an increasing reporting delay.

3.7 Considerations for model construction

As per Section 3.6, the software used to build the model should be able to allow for algorithmic, rule-based reserves calculations and be able to generate predictions for performance metrics.

The feature selection process can be time-consuming because it must be done iteratively.

The actuary will want to have dashboards and performance metrics preconfigured in an end-to-end model so that the impact of every parameterisation change can be seen quickly with the entire reserving process running automatically. The ability to perform scenario analysis, and select features used in the clustering process iteratively, will be useful.

The reserving model should ideally be able to adapt dynamically to a differing number of segments or sub-classes to make the process of iteratively comparing reserve performance based on a differing number of clusters easier. The number of clusters is likely to be a model parameter. Many clustering methods require the number of clusters created as an input parameter and do not automatically determine an optimal number of clusters.

4. CONCLUSION AND DISCUSSION

This paper presents a framework that can be used for employing clustering techniques as a basis for segmentation in triangle reserving models. It described a methodology for including clustering features that are based on the transaction history.

The framework aims to address the limitations of the homogeneity assumption in triangle reserving methods. A key benefit of employing this framework is that it can easily be adopted in practice since it can be used within the context of regulatory reporting or legacy class hierarchy constraints as it can be used to create sub-class reserve segments. Another important benefit of employing this framework in a reserving process is to get meaningful insights into claim trends through cluster analysis.

This framework was initially tested on a small dataset from a short-tail class of business that is traditionally stable and relatively homogeneous. The framework was found to be able to slightly improve the reserving accuracy of both the paid and incurred chain

ladder using the testing methodology described (See Appendix A). However, the results of this test are not convincing since the results were sensitive to the model parameters used.

To better measure the benefit of employing this framework, performance should be tested on a dataset of a heterogeneous class of business that is more volatile over time and has claims with varying development patterns. Ideally, this test should be done over more than one valuation period to see if its implementation can help the actuary by better monitoring trends in claims experience.

Further work can be conducted to find a way to build a model which is able to find the claim features of a dataset which minimises the RMSE test statistic by running several different permutations of including different claim features for clustering automatically. Identifying the optimal features can help the reserving actuary find trends and understand what drives the heterogeneity in claims for a class of business.

Acknowledgements

I'm extremely grateful to the Dynamo Analytics team who supported the writing of this paper. Special thanks to Matthew Webster who contributed to the ideas in this paper, Ronald Richman for acting as a soundboard and helping with the sourcing of test data. Jack Manning and Simon Duncan for proofreading, Petroné Moolman and Bronwyn Muir for helping to build the testing model and James Ellis for sharing his machine-learning experience. Lastly, I would like to extend my sincerest thanks to Lisa Pines for reviewing this paper and contributing by sharing her practical experience.

REFERENCES

Avitabile, J & Cooke, J (2018). Clustering for reserving segmentation. Presentation. https://www.casact.org/sites/default/files/presentation/clrs_2018_presentations_st-1_avitabile.pdf.

Balona, C & Richman, R (2020). The actuary and IBNR techniques: A machine learning approach. Available at SSRN 3697256 .

Buhlmann, H (1970). Mathematical methods in risk theory. Springer-Verlag, Berlin, Heidelberg, New York.

Clark, D & Jiang, V (2018). Cluster analysis in loss development. Presentation. https://www.casact.org/sites/default/files/presentation/rpm_2019_presentations_r-4_clark.pdf

Ding, C & He, X (2004). K-means clustering via principal component analysis, in Proceedings of the twenty-first international conference on Machine learning, p. 29.

John, A (2018). Granular reserving dialogistic in machine learning. Presentation. https://www.actuaries.org.uk/system/files/field/document/Reserving%20ML%20Presentation%2020Jun18%20v1.2_2.pdf.

Kaloyanova, E (2021). How to combine PCA and K-means clustering in python? https://365datascience.com/tutorials/python-tutorials/pca-k-means/

Reid, D (1978). Claim reserves in general insurance, *Journal of the Institute of Actuaries.* **105**(3): 211–315.

Taylor, G, McGuire, G & Sullivan, J (2008). Individual claim loss reserving conditioned by case estimates, *Annals of Actuarial Science,* **3**(1–2), 215–256.

Yeo, A, Smith-Miles, K, Willis, R & Brooks, M (2001). Clustering technique for risk classification and prediction of claim costs in the automobile insurance industry, *Int. Syst. in Accounting, Finance and Management.* 10, 39–50.

Zehnwirth, B (1989). The chain ladder technique – a stochastic model, *Claims Reserving Manual.* 2, 2–9.

APPENDIX A: ILLUSTRATIVE EXAMPLE

A.1 Data description

10 000 claims were randomly selected from a large class of business that is usually considered to be a short-tail class.

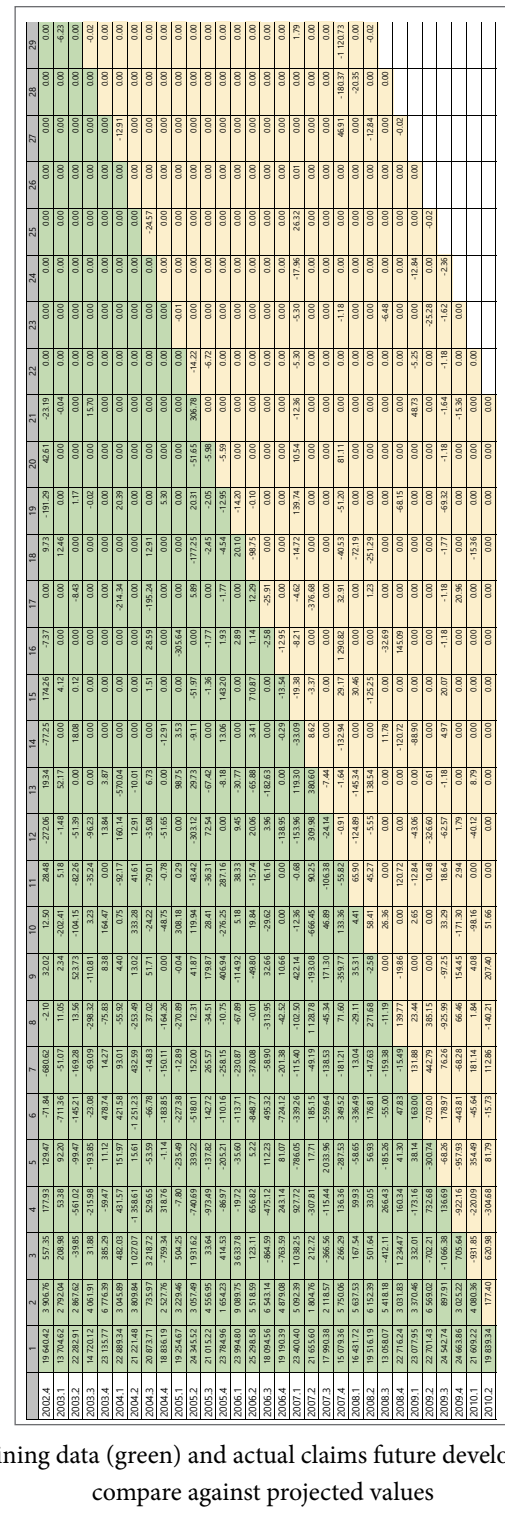
The claims' origin dates start from 1 October 2002. Claims younger than 30 June 2010 are cut off from the testing sample, leaving 8710 claims in the training dataset.

◢ **ClaimID: 120205407**					
01/10/2002	01/10/2002	01/10/2002	01/01/2003	3.533	102.076
01/10/2002	01/11/2002	01/10/2002	01/01/2003	0.000	0.000
01/10/2002	01/12/2002	01/10/2002	01/01/2003	0.000	0.000
01/10/2002	01/01/2003	01/10/2002	01/01/2003	44.351	-54.192
◢ **ClaimID: 120208063**					
01/10/2002	01/10/2002	01/10/2002	01/12/2002	3.533	29.356
01/10/2002	01/11/2002	01/10/2002	01/12/2002	0.000	0.000
▸ **ClaimID: 120208269**					
▸ **ClaimID: 120208570**					

FIGURE 2 Snapshot of the data structure of the claim transactions table

The full history of claims development data is available up to 30 June 2015. The actual development data between 1 July 2010 and 30 June 2015 can be used to compare against the projected amounts.

Figure 3 illustrates the sample training data in green and the actual data to test against the projected amounts in yellow.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
2002.4	19 640 422	3 906 776	557 351	177 393	129 497	-71 944	-680 652	-2 310	32 022	12 501	28 448	-272 096	19 344	-77 255	174 256	-2 377	0.00	9 731	191 295	40 611	-23 119	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2003.1	13 704 622	2 792 094	208 980	92 330	53 318	-711 366	-5 107	11 085	2 344	-202 401	5 118	-1 483	52 177	0.00	0.00	4 122	0.00	0.00	12 466	0.00	-4 044	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-4 233
2003.2	22 282 911	2 867 652	-39 885	-564 022	-99 457	-145 231	169 281	13 566	553 273	-104 115	-82 256	-51 299	0.00	18 088	0.12	0.00	0.00	8 458	0.00	1 177	0.00	15 705	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2003.3	14 720 124	4 496 191	31 088	-215 958	-193 885	-23 088	-69 699	-298 332	110 881	-3 233	35 244	-96 223	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.02
2003.4	23 135 777	6 776 839	385 290	-594 477	11 112	478 744	14 277	-75 883	8 398	164 477	0.00	13 984	3 307	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2004.1	22 889 344	3 945 889	462 030	-431 577	151 697	421 558	93 011	-55 582	4 440	0.75	-98 177	160 114	-570 044	0.00	0.00	0.00	-214 344	0.00	20 389	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-12 951	0.00	0.00
2004.2	21 221 148	3 809 984	1 027 077	-1 358 611	13 651	-1 251 223	-452 599	-255 489	13 012	333 286	41 611	12 951	-10 011	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2004.3	20 873 711	735 917	3 218 72	529 655	-53 599	-66 781	-14 833	37 052	5 171	-24 422	-79 011	-35 088	6 731	0.00	1 511	28 559	-195 254	12 931	0.00	0.00	0.00	0.00	0.00	-24 577	0.00	0.00	0.00	0.00	0.00
2004.4	18 836 619	2 527 76	-799 34	318 76	-1 14	-188 385	-150 111	-164 266	0.00	-48 75	-0 78	-51 65	0.00	-12 911	0.00	0.00	0.00	0.00	5 301	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2005.1	19 254 467	3 229 46	504 25	-7 80	-235 49	-227 38	-12 869	-270 89	-0 04	308 18	0.29	0.00	98 75	3 53	0.00	-305 64	0.00	0.00	0.00	0.00	0.00	0.00	-0 01	0.00	0.00	0.00	0.00	0.00	0.00
2005.2	24 345 52	3 057 49	1 931 62	-740 69	339 22	-518 01	15 200	12 31	41 87	119 94	43 42	-303 12	29 73	-9 11	-51 97	0.00	5 891	-177 25	20 31	-51 65	306 78	-14 22	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2005.3	21 015 22	4 556 95	33 64	-973 49	-137 82	142 72	265 57	-34 51	179 87	28 41	-36 31	72 54	-67 42	0.00	-1 36	-1 77	0.00	-2 45	-2 05	-5 98	0.00	-6 72	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2005.4	23 784 96	1 654 23	414 53	-86 97	-205 21	-110 16	-258 15	-10 75	406 94	-276 25	287 16	0.00	-8 18	13 06	143 20	1 93	-1 77	-4 54	-12 95	-5 59	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2006.1	23 994 80	9 089 75	3 633 78	-19 72	-35 60	-113 71	-230 87	-67 89	-11 492	5 18	38 33	9 45	-30 77	0.00	0.00	2 89	0.00	20 10	-14 20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2006.2	25 296 58	5 118 99	123 11	656 82	5 22	-848 77	-374 08	-0 01	-49 80	19 84	-15 74	20 06	-65 88	3 41	710 87	1 14	12 29	98 75	-0 10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2006.3	18 094 56	6 543 14	-864 39	-475 12	112 23	495 32	-58 90	-313 95	32 66	-256 2	16 16	3 96	-182 63	0.00	0.00	-2 58	-25 91	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2006.4	19 190 39	4 879 08	-763 99	243 14	81 07	-724 12	-201 38	-42 52	10 66	0.00	0.00	-138 95	0.00	-0 29	-13 54	-12 95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2007.1	23 460 40	5 092 39	1 038 25	927 72	-786 05	-339 26	-115 40	-102 90	422 14	-12 36	-0 68	-153 96	119 30	-33 09	-19 38	-8 21	-4 62	-14 72	139 74	10 54	-12 36	-5 30	-5 30	-17 96	-26 32	0.01	0.00	0.00	1 79
2007.2	21 655 60	1 804 76	212 72	-307 81	17 71	185 15	-49 19	1 128 78	-193 08	-666 45	90 25	309 98	380 60	8 62	-3 37	0.00	-376 68	0.00	0.00	0.00	0.00	0.00	48 73	-5 25	0.00	-12 84	0.00	0.00	0.00
2007.3	17 960 38	2 118 57	-366 56	-115 44	2 033 96	-559 64	138 53	-45 34	171 30	-46 89	-106 38	-24 14	-7 44	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-25 28	0.00	-0 62	0.00	0.00	0.00
2007.4	15 013 16	5 750 06	246 29	136 16	-287 53	340 52	187 21	71 60	-359 77	133 36	-55 82	-0 91	-1 64	-132 94	29 17	1 290 82	32 91	-40 53	-51 20	81 11	0.00	0.00	-1 18	0.00	0.00	46 91	-180 37	-1 120 73	
2008.1	19 481 72	6 612 39	601 64	99 93	35 05	54 93	176 81	147 65	27 68	-2 58	34 31	4 41	65 10	-124 89	-145 54	0.00	0.00	-72 19	0.00	0.00	0.00	0.00	-1 77	-1 64	-15 36	0.00	0.00	-20 35	0.00
2008.2	19 516 19	6 112 39	601 64	54 93	176 81	147 65	27 68	-2 58	34 31	4 41	65 10	-124 89	-145 54	0.00	-125 25	0.00	1 23	-251 29	0.00	0.00	0.00	0.00	-69 32	-1 18	-15 36	0.00	0.00	-12 84	0.00
2008.3	13 056 07	5 418 18	-412 11	266 43	-165 36	-55 00	-159 38	-11 19	0.00	26 36	0.00	0.00	11 78	0.00	-32 69	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-6 48	0.00	0.00	0.00	0.00	-0 02
2008.4	22 716 24	3 001 83	12 944 7	160 34	41 30	-97 83	-15 49	139 77	-19 66	0.00	120 72	0.00	0.00	-120 72	0.00	141 09	0.00	0.00	-68 15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0 02		
2009.1	23 077 95	3 370 46	32 01	-173 16	38 14	163 00	131 88	23 44	0.00	2 65	-12 84	-43 06	0.00	-88 90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	48 73	-5 25	0.00	-12 84	0.00	0.00		
2009.2	22 701 43	6 569 02	-702 21	732 68	-300 79	-703 00	-442 79	385 15	0.00	0.00	10 48	-326 60	0.61	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-25 28	0.00	-0 62				
2009.3	24 542 74	897 91	-1 066 38	136 69	-68 26	178 97	76 26	-925 99	-97 25	33 29	18 64	-62 57	-1 18	4 97	20 07	-1 18	-1 18	-1 77	-69 32	-1 18	-1 64	-1 18	-1 62	-2 36					
2009.4	24 663 86	3 025 22	705 64	-922 16	-957 93	-443 81	-68 28	66 46	154 45	-171 30	2 94	1 79	0.00	0.00	0.00	0.00	20 96	0.00	0.00	0.00	-15 36	0.00	0.00						
2010.1	21 609 22	4 080 36	-931 85	-220 09	354 49	-45 64	181 14	1 84	4 08	-98 16	0.00	-40 12	8 79	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00							
2010.2	19 839 34	177 40	620 38	-304 68	81 79	-15 73	112 86	-140 21	207 40	51 66	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00									

FIGURE 3 Showing training data (green) and actual claims future developments (yellow) used to compare against projected values

A.2 Closed claims clustering using PCA and the K-Means algorithm

Using the data as described in Appendix A.1 with valuation period 30 June 2010, separating closed from open claims, resulted in 8411 closed claims and 299 open claims.

This means that 8411 closed claims will be clustered together and 299 open claims will get allocated to the closed claims clusters.

The model was built to allow the following nine claim features on closed claims:

— **ReportedCaseEstimate** – The initial estimated total claim size.
— **OpenTime** – Total number of days between the close and opening dates of the claim.
— **ReportingDelay** – The number of days between the occurrence of the insured event and the report date.
— **IncurredMovementCount** and **PaidMovementCount** – The total number of nonzero incurred/paid values.
— **IncurredTimeWeightedAverage** and **PaidTimeWeightedAverage** – The incremental incurred/paid movement amounts, weighted by the distance in days of the transaction date and the report date, divided by the total open time in days.
— **IncurredDuration** and **PaidDuration** – Macaulay Duration, calculated on the incremental incurred/paid movements. The feature in this example is calculated with a discount rate of 0%, but including a discount rate would be a useful addition.

These features are calculated for every closed claim and standardised by subtracting the mean of each feature and dividing by the standard deviation with the code in Figure 4.

```
from sklearn.preprocessing import StandardScaler
######## Standardized Features ######## scaler = StandardScaler()
feature_std = scaler.fit_transform(feature_data)
```

Ding and He (2004) showed that transforming the data with PCA can reduce the dimensions and noise of the data, and ultimately improve the performance of K-Means clustering. This method is used in this closed claim clustering example.

The feature data is transformed through PCA:

```
from sklearn.decomposition import PCA
######## Perform PCA with 2 Components ######## pca =
PCA(n_components=2) pca.fit(feature_std)
scores_pca = pca.transform(feature_std)
```

Clustering is done through K-Means:

```
from sklearn.cluster import KMeans
######## Perform KMeans with 2 Components ######## kmeans_pca = KMeans(n_clusters=2,
nit='k-means++', n_init=10, random_state=42) kmeans_pca.fit(scores_pca)
```

PCA and K-Means were implemented based on the tutorial by Kaloyanova (2021).

ID	ReportedCaseEstimate	OpenTime	ReportingDelay	IncurredMovementCount	IncurredTimeWeighedAverage	IncurredDuration	PaidMovementCount	PaidTimeWeighedAverage	PaidDuration
121091375	51.65	29	0	5	151.08	75.85	4	204.51	102.67
121872878	23.24	1	0	2	-536.88	-132.00	1	113.88	28.00
120746473	224.30	9	0	4	687.50	25.74	4	1324.49	49.59
122363954	107.58	1	0	2	1432.77	9.03	1	4445.07	28.00
120726498	29.18	1	0	2	-308.79	-16.07	1	595.82	31.00
122114894	12.91	1	0	2	-164.54	-21.64	1	235.73	31.00
121475212	81.59	2	0	2	5.23	0.13	2	2406.84	58.87
120833543	51.34	3	1	2	186.95	9.76	2	1644.10	85.87
121729434	2.94	0	0	1	0.00	0.00	1	0.00	0.00
121908635	179.99	2	0	2	116.73	1.27	1	5606.45	61.00
120780299	18.85	2	1	2	-32.23	-3.62	1	552.16	62.00

Figure 4 Features calculated for closed claims

A.3 Open claims cluster allocation features

The algorithm was implemented as per Section 3.5, using 10 equally spaced operational time bands.

The model was built to allow the following feature calculations in each operational time band as in Step 4:

— ReportedCaseEstimate – The initial estimated total claim size.
— **ReportingDelay** – The number of days between the occurrence of the insured event and the report date.
— I**ncurredMovementFrequencyRate** and **PaidMovementFrequencyRate**– The total number of nonzero incurred/paid values, divided by the total number of days a claim has been open.
— **PaidToIncurredLatest** – The sum of all paid amounts to date divided by the total incurred amount to date.
— **IncurredGrowth** – The factor by which the incurred amount has grown to date. i.e., the total incurred amount to date divided by the initial case estimate at the report date.
— **IncurredDuration** and **PaidDuration** – Macaulay Duration, calculated on the incremental incurred/paid movements, calculated with a discount rate of 0%.

A.4 Results and discussion

A.4.1 TESTING METHODOLOGY

The model was run with various iterations to test the clustering method and performance, with each iteration changing:

— the valuation date, back-testing at different points in time in the data's history.
— various features included in the closed and open claims clustering methods.
— the number of clusters created by the closed claims clustering method.

The RMSE was calculated as a testing measure, comparing the actual claims development amounts and the sum of each cluster segment's incremental projected lower triangle produced by chain ladder projections on incurred and paid data respectively. The same test statistic was calculated using the same chain ladder method assumptions on the aggregate triangulated dataset (in other words, without segmentation).

Testing the model iteratively produced mixed results. In some tests, the clustered method produced better results than the aggregate projection and vice versa. Different feature selections also optimised either the paid chain ladder or the incurred chain ladder. Interestingly, in this dataset, it was most often the paid projection which was improved by employing the clustered segmentation.

A.4.2 EXAMPLE RESULT OF TEST SCENARIO

Valuation period:

The valuation period of the test scenario was set to 30 June 2010, using all the available claims in the data and hence not back-testing on a reduced training dataset as the model is capable of doing dynamically as mentioned in Appendix A.4.1.

Number of clusters:

Given that the size of the sample dataset is relatively small, the model was set to only create two clusters to ensure sufficient volumes of claims in the segmented triangles.

Feature selection:

To focus on optimising the performance of the incurred chain ladderproject with clustering, features were selected that best describe the development pattern of the incurred data. The following features were selected:

Closed claims clustering features:

— OpenTime
— ReportingDelay
— IncurredTimeWeightedAverage
— IncurredDuration

Open claims allocation features:

— ReportingDelay
— IncurredToPaidLatest
— IncurredGrowth
— IncurredDuration

Most of these claim features describe the incurred development of a claim. This resulted in the following cluster allocations:

Table 1 Number of claims allocated to each cluster

Cluster Number	Closed Claims	Open Claims
0	7735	98
1	676	201

The chain ladder methods produced the following pattern calculations for the different clustered segments compared to the aggregate triangles (Figure 5).

This parameterisation produced the following RMSE (Table 2).

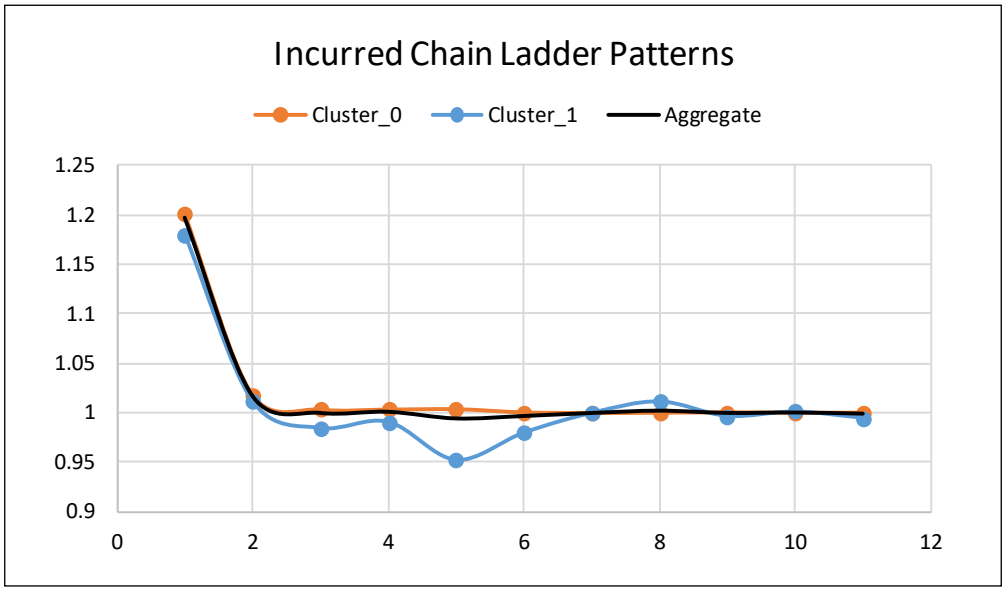


FIGURE 5 Showing chain ladder link ratio patterns by development period based on the aggregate versus segmented incurred triangle training data

TABLE 2 RMSE by method

Object Name	Clustered	Aggregate
IncurredProjected	229.58	234.34
PaidProjected	268.75	296.89

It can be seen here that employing clustering-based segmentation is slightly able to outperform the aggregate chain ladder using this specific feature selection for the dataset.

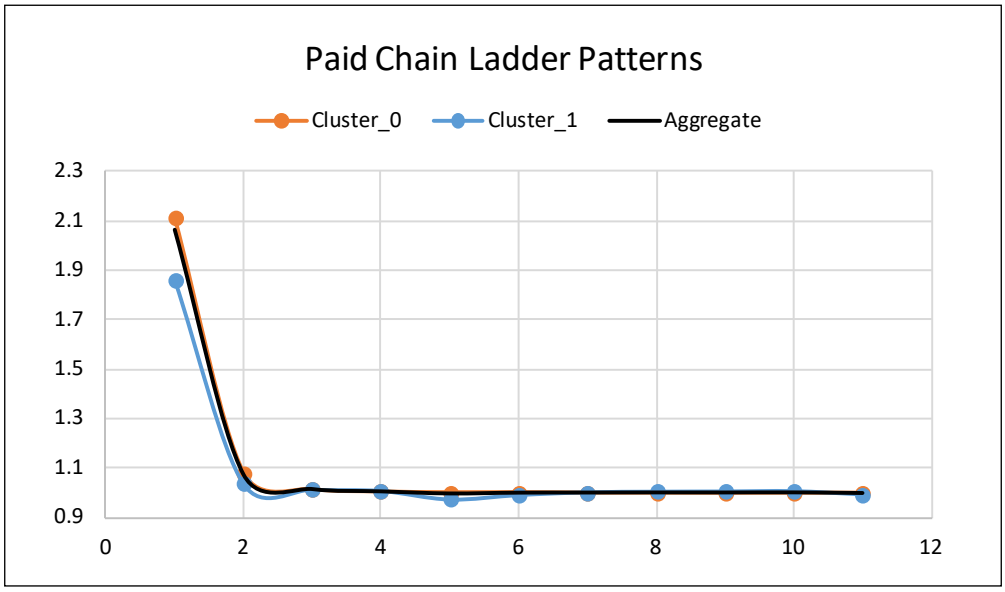


FIGURE 6 Showing chain ladder link ratio patterns by development period based on the aggregate versus segmented paid triangle training data