Nonparametric and Model-based Clustering Approaches to Data Compression for Analysing Actuarial Data

Dr. Adrian O'Hagan & Mr. Colm Ferrari, BAFS

January 2015

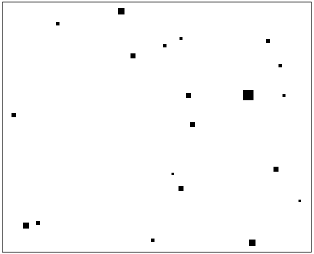
Running order

1. Data compression - non-parametric VS model-based.
2. Model-based clustering - covariance, correlation, PCA.
3. In-sample results - 50 and 250 clusters.
4. Fitting larger numbers of clusters - feedback sampling.
5. In-sample results - 1000, 2500 and 5000 clusters.
6. Out-of-sample results - CTE70, present values of variables.
7. Conclusions and further work - general insurance extensions.

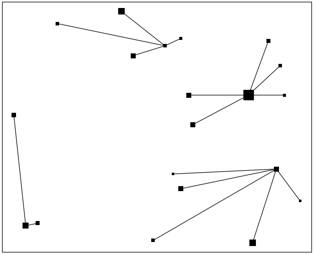
Data Compression by Clustering

- We have a dataset of 110,000 policies with 54 “location” variables and a “size” variable.
- We want to compress the data into clusters that can each be represented by a single, scaled-up policy.
- The aim is for the scaled-up representative policies to replicate the behaviour of the full dataset over a range of stochastic economic scenarios as closely as possible.
- Some compression technique is necessary because it is not feasible to compute a large range of scenarios for the full dataset.

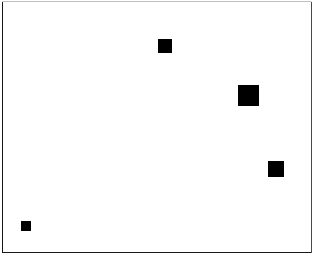
Data Compression by Clustering



(a)



(b)

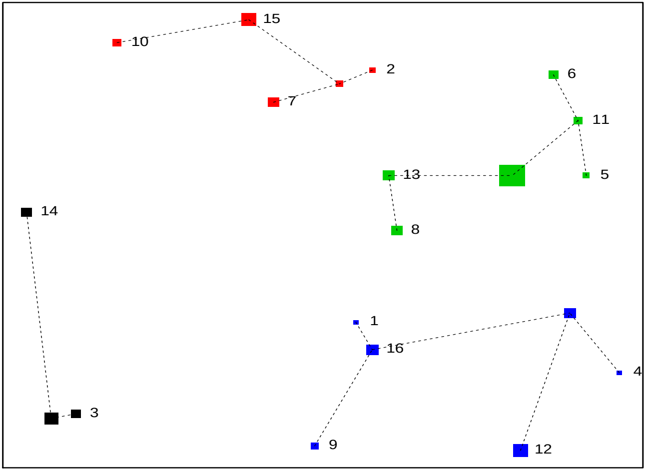


(c)

Data Compression by Clustering

- Current practice (Freedman & Reynolds, 2008) is to use size-weighted hierarchical clustering: iteratively merge the "least important" policy with its nearest neighbour until only the desired number remain.
- A variety of clustering algorithms exist. Can alternative methods result in representative model points that replicate the behaviour of the full data set more accurately over a range of scenarios?
- We test the new clustering methods at various levels of compression - from 110000 policies to 50, 250, 1000, 2500 and 5000 clusters.

Existing Approach - Milliman's Method

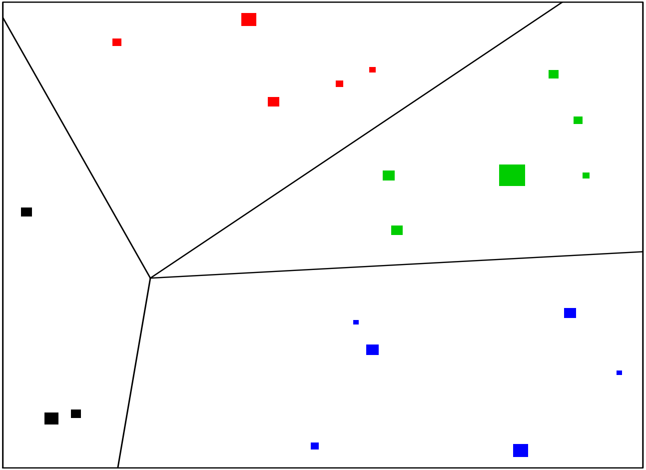


K-medoids Clustering

- Given some initial partition, identify the medoid of each cluster.
- Assign each object to the cluster whose medoid is closest.
- Identify the new cluster medoids.
- Repeat until no more objects are reassigned.

The k clusters will be linearly separable, similarly sized and approximately spherical.

K-medoids Clustering



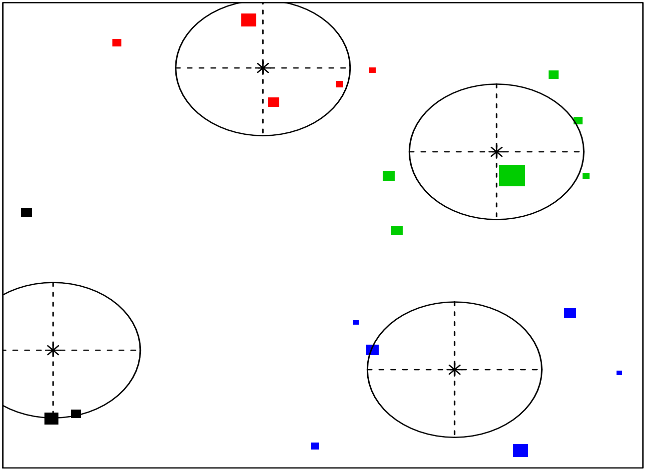
Ward's Minimum Variance Hierarchical Clustering

- Begin by treating each object as an individual cluster.
- Then iteratively merge the pair of clusters that will result in the smallest increase in total within-cluster variance:

$$\sum_{k=1}^{G}\sum_{i=1}^{n_k}\sum_{j=1}^{p}\frac{1}{n_k}(x_{ij}-\bar{x}_{kj})^2 \tag{1}$$

This method produces compact, spherical clusters.

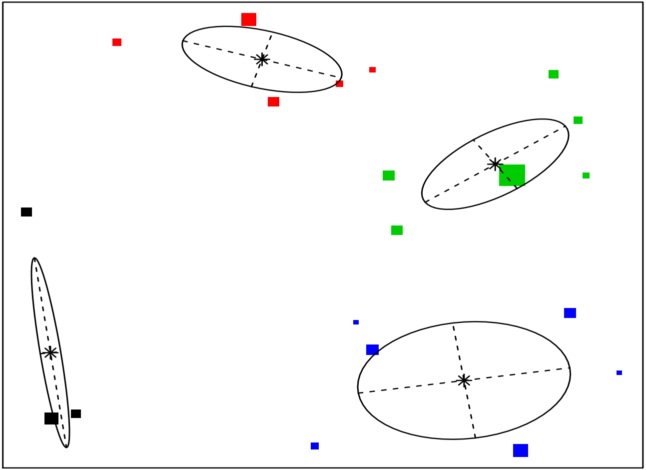
Ward's Minimum Variance Hierarchical Clustering



Model-based Clustering

- Assume that the objects within each cluster follow a multivariate normal distribution.
- Use the EM algorithm to estimate the parameters.

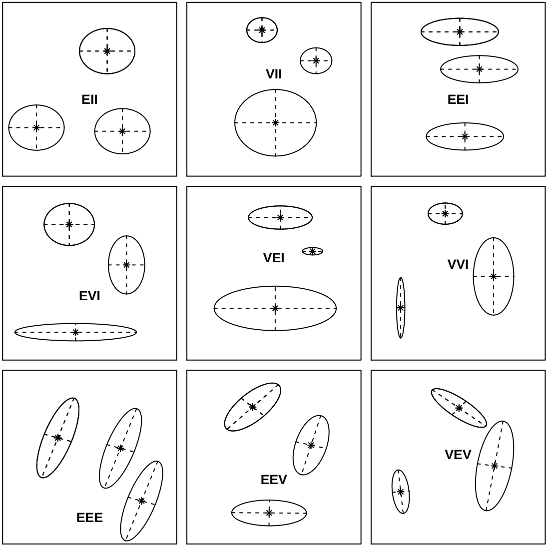
Model-based Clustering



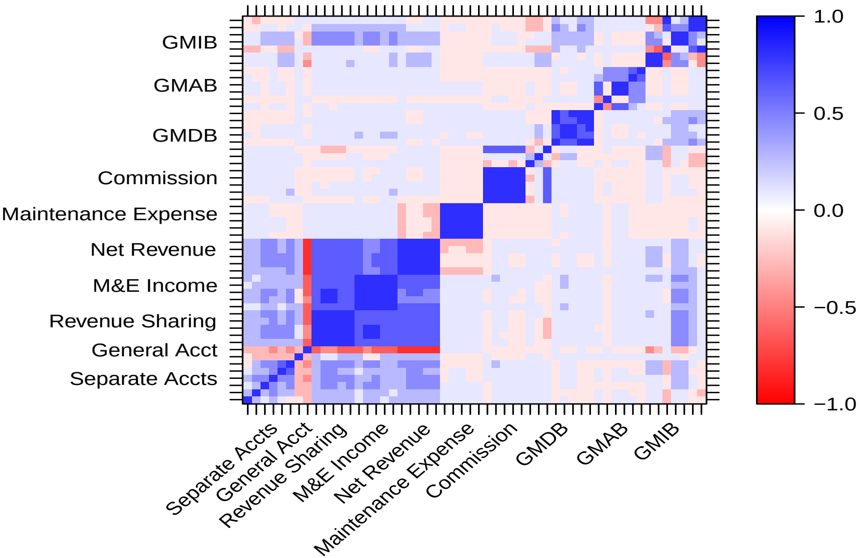
Implementing the model-based approach

- Fit the Gaussian mixture model using **mclust** in **R** with the **me.weighted** step to account for policy size.
- 50 and 250 clusters: exact **mclust** solution available via laptop processing.
- 1000 and 2500 clusters: exact **mclust** solution available via cluster processing.
- 5000 clusters: exact solution not available via **mclust**.

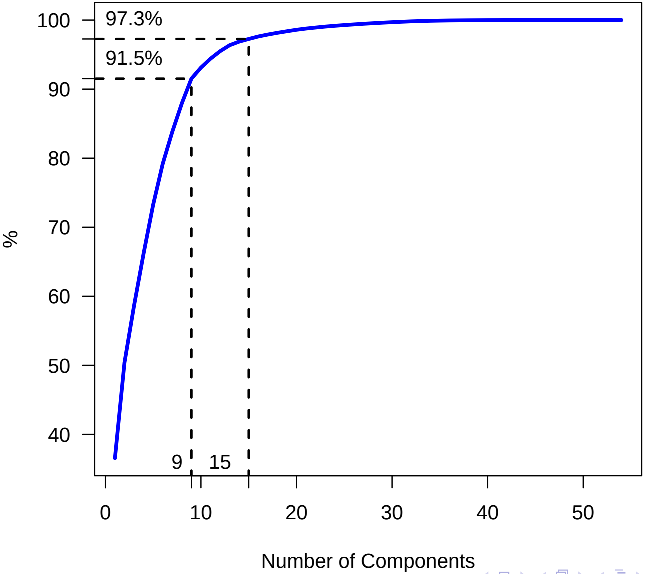
Available Covariance Structures



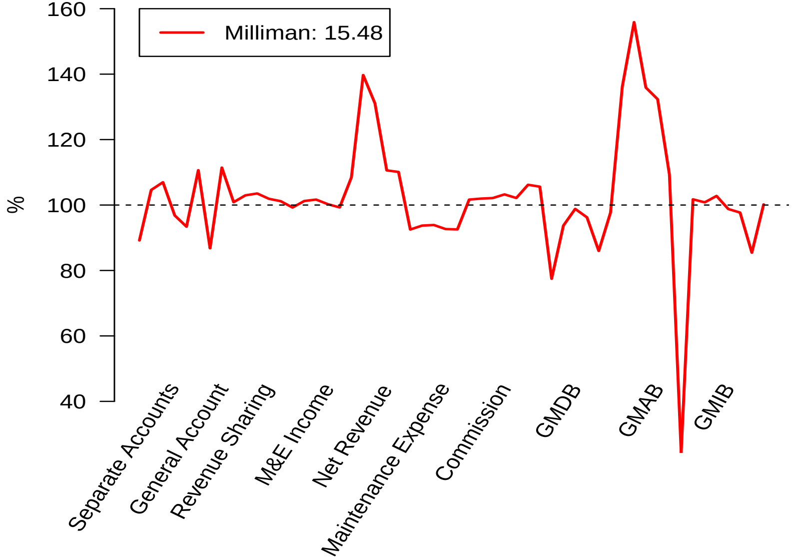
Weighted Correlation of Location Variables



PCA - Proportion of variance explained

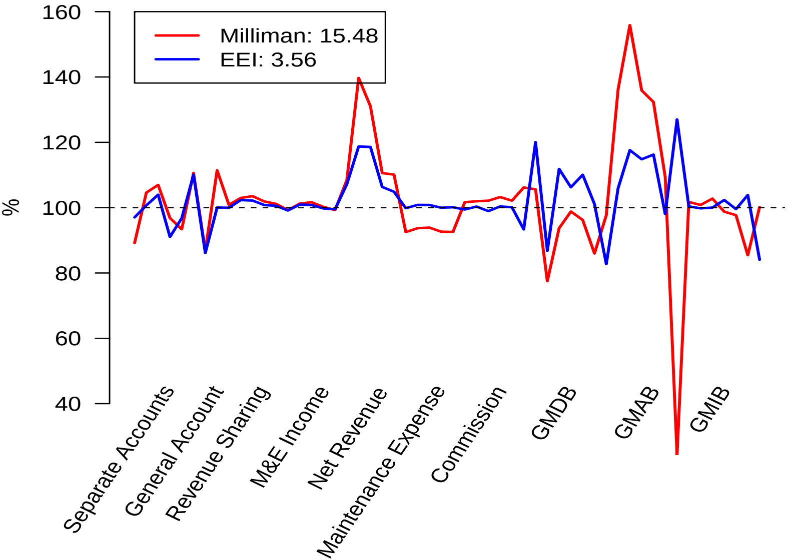


50 Clusters



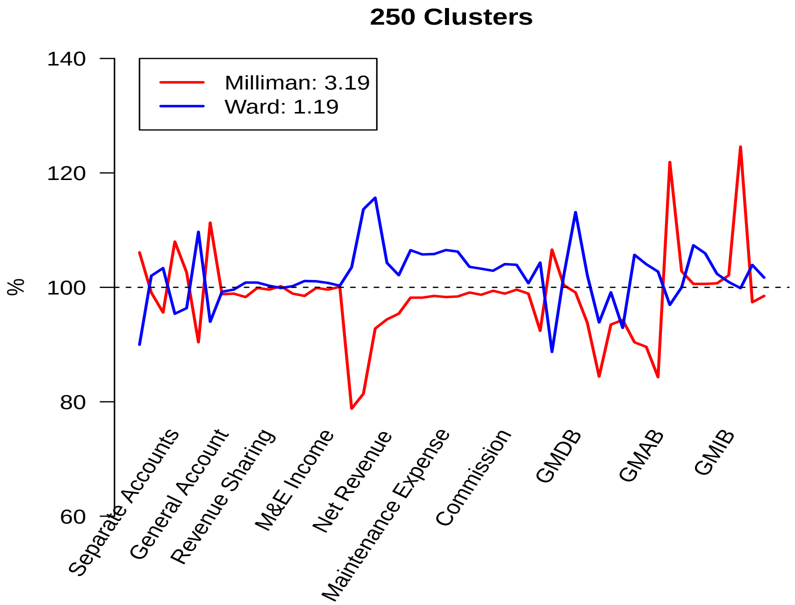
Milliman: 15.48
%
160
140
120
100
80
60
40
Separate Accounts
General Account
Revenue Sharing
M&E Income
Net Revenue
Maintenance Expense
Commission
GMDB
GMAB
GMIB

50 Clusters



Milliman: 15.48
EEI: 3.56
160
140
120
100
80
60
40
%
Separate Accounts
General Account
Revenue Sharing
M&E Income
Net Revenue
Maintenance Expense
Commission
GMDB
GMAB
GMIB


250 Clusters
Milliman: 3.19
140
120
100
80
60
%
Separate Accounts
General Account
Revenue Sharing
M&E Income
Net Revenue
Maintenance Expense
Commission
GMDB
GMAB
GMIB


250 Clusters
Milliman: 3.19
Ward: 1.19
140
120
100
80
60
%
Separate Accounts
General Account
Revenue Sharing
M&E Income
Net Revenue
Maintenance Expense
Commission
GMDB
GMAB
GMIB

Fitting Larger Numbers of Clusters

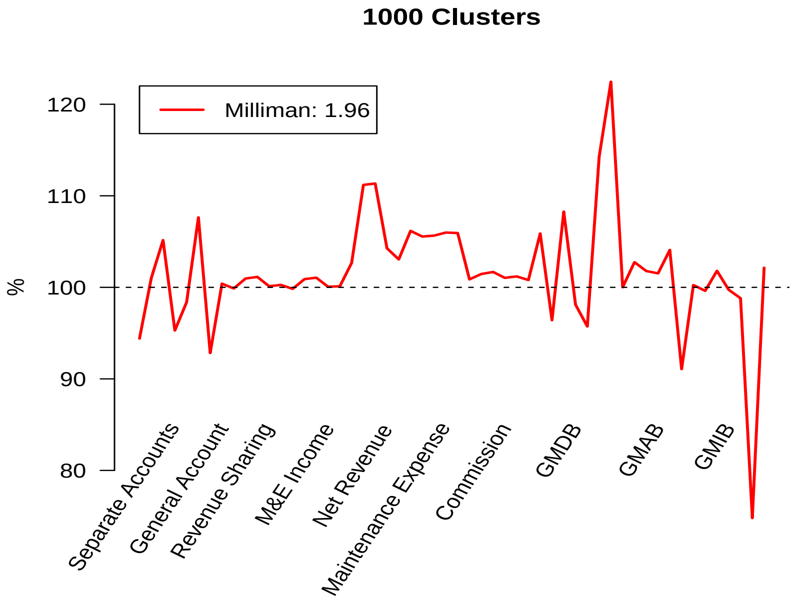
- Direct application of model-based clustering to large datasets with large numbers of clusters can be prohibitively expensive in terms of computer time and memory.
- e.g. a VVV model with 5000 clusters and 15 location variables would require the estimation of hundreds of thousands parameters.
- Feedback sampling is an approach we have developed that takes advantage of the size-weighted nature of the data to partition the data into large numbers of clusters.

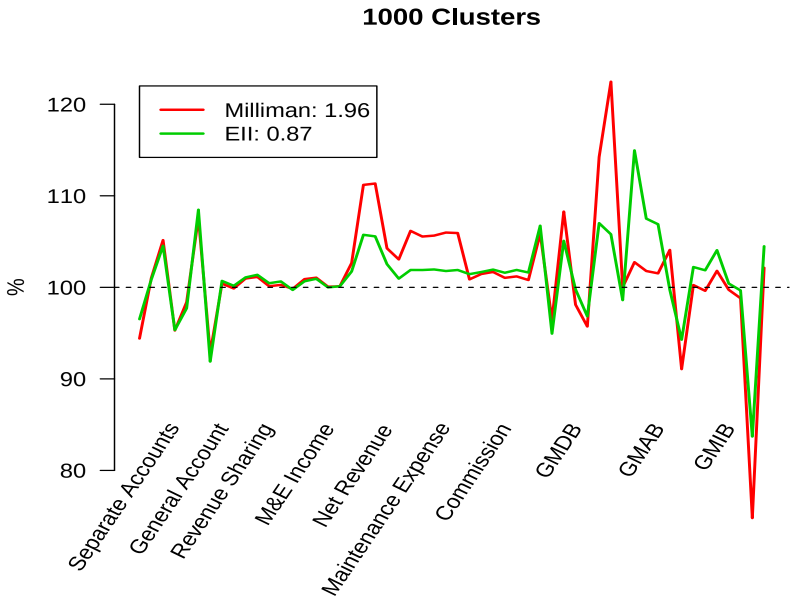
Fitting Larger Numbers of Clusters - Feedback Sampling

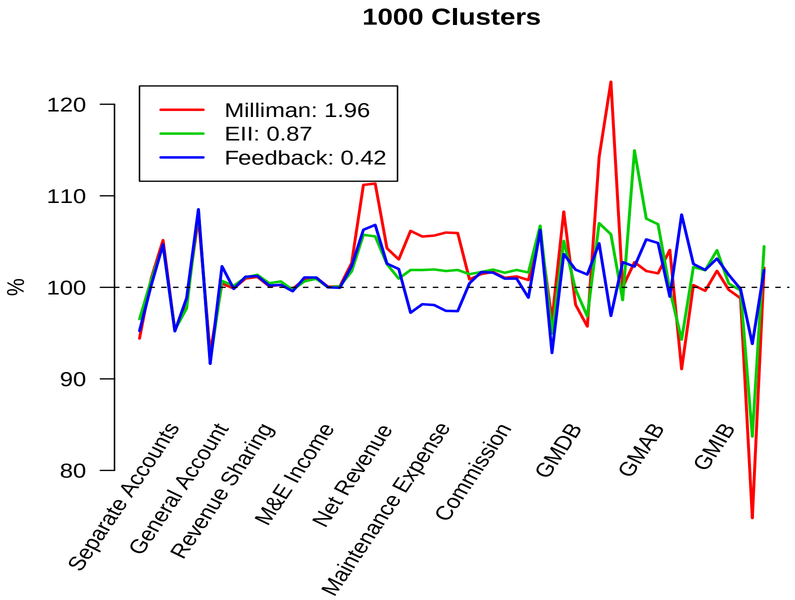
- Take a sample of 2500 objects.
- Partition the sample into a moderate number (e.g. 20-50) of clusters using weighted `mclust`. BIC can be used to select the optimum model type and number of clusters g.
- Treat the resulting cluster centres as g individual objects, scaled up by the sums of the sizes of the objects in each cluster.

Fitting Larger Numbers of Clusters - Feedback Sampling

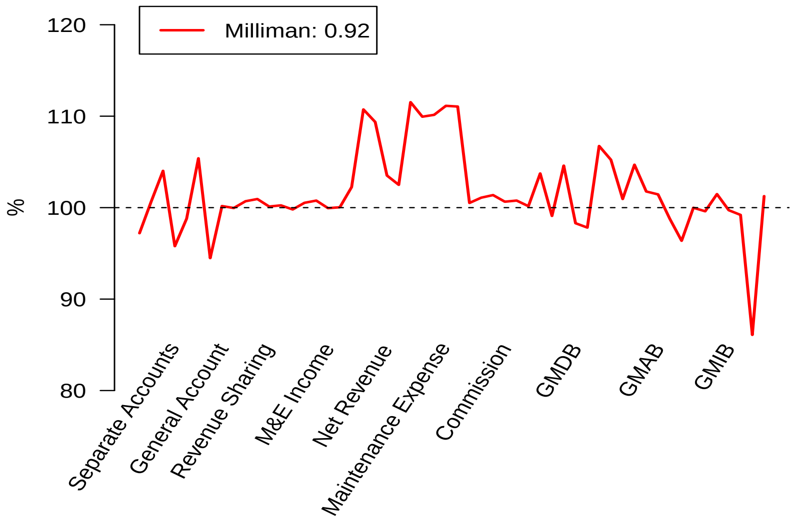
- Replace the sampled objects in the data set with these g scaled-up cluster centres, thus reducing the size of the data set by $(2500 - g)$.
- Repeat until the desired number of objects or cluster centres remain.
- Then simply assign each policy to the cluster whose centre is closest.


1000 Clusters
Milliman: 1.96
%
120
110
100
90
80
Separate Accounts
General Account
Revenue Sharing
M&E Income
Net Revenue
Maintenance Expense
Commission
GMDB
GMAB
GMIB


1000 Clusters
Milliman: 1.96
EII: 0.87
%
120
110
100
90
80
Separate Accounts
General Account
Revenue Sharing
M&E Income
Net Revenue
Maintenance Expense
Commission
GMDB
GMAB
GMIB

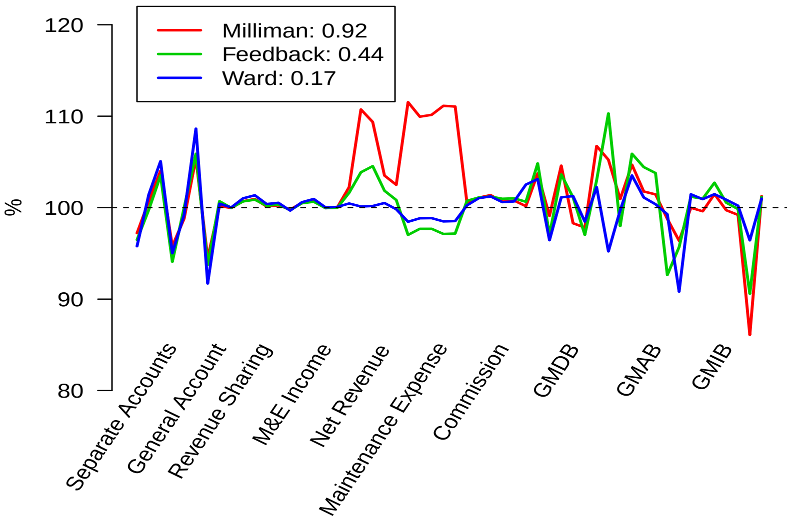

1000 Clusters
Milliman: 1.96
EII: 0.87
Feedback: 0.42
120
110
100
90
80
%
Separate Accounts
General Account
Revenue Sharing
M&E Income
Net Revenue
Maintenance Expense
Commission
GMDB
GMAB
GMIB

2500 Clusters

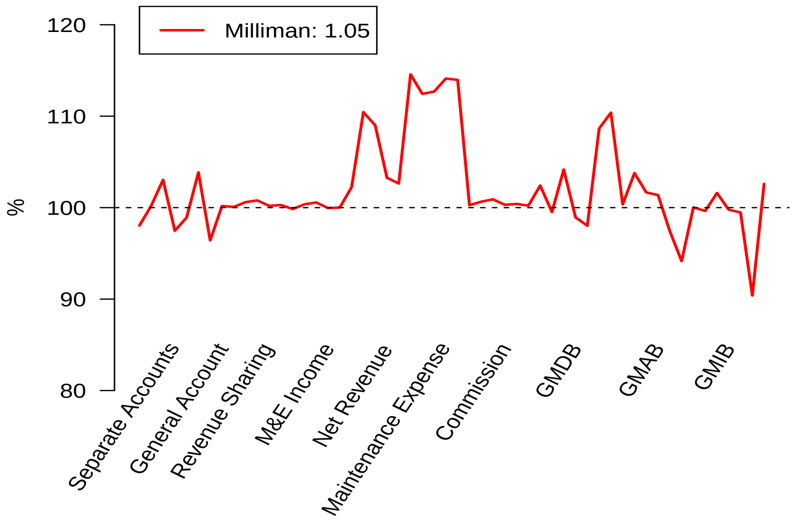



2500 Clusters
Milliman: 0.92
Feedback: 0.44
120
110
100
90
80
%
Separate Accounts
General Account
Revenue Sharing
M&E Income
Net Revenue
Maintenance Expense
Commission
GMDB
GMAB
GMIB

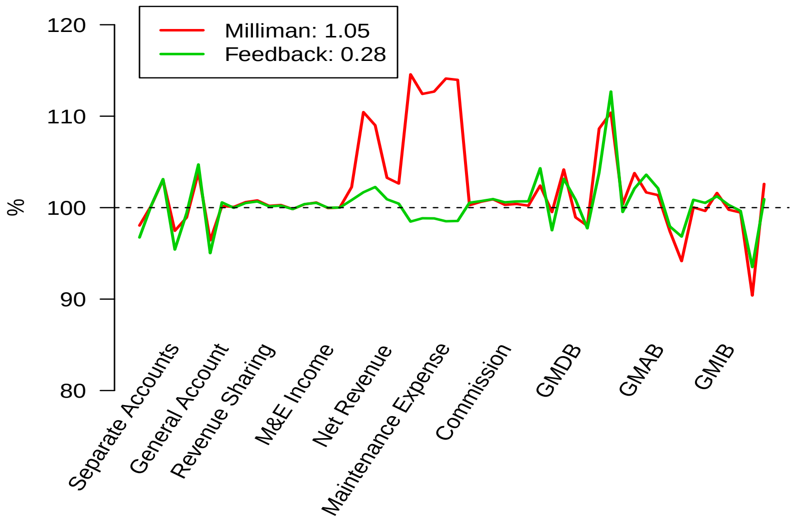
2500 Clusters

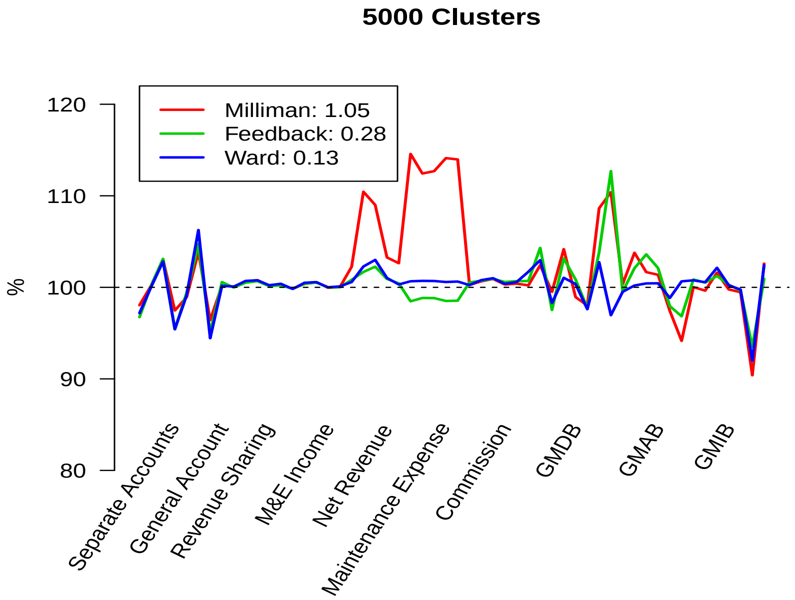


5000 Clusters

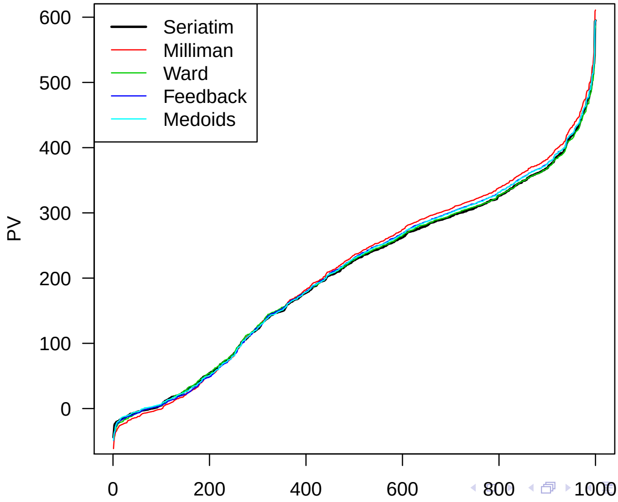


5000 Clusters

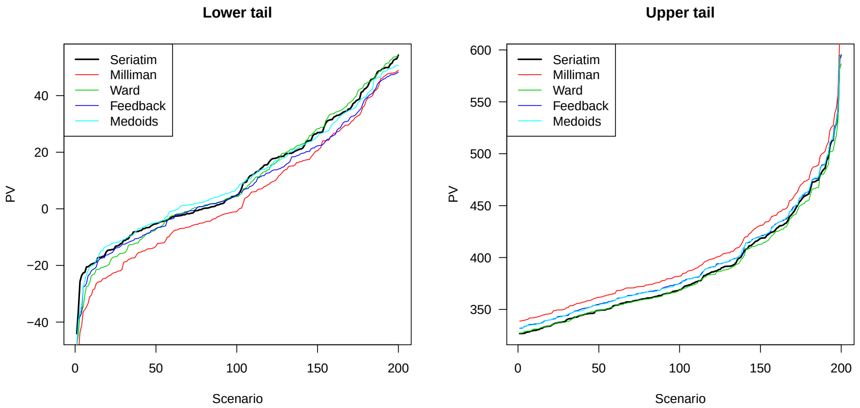



5000 Clusters
Milliman: 1.05
Feedback: 0.28
Ward: 0.13
120
110
100
90
80
%
Separate Accounts
General Account
Revenue Sharing
M&E Income
Net Revenue
Maintenance Expense
Commission
GMDB
GMAB
GMIB

Out-of-sample results for 2500 clusters - PV of Net GMIB Costs



Out-of-sample results for 2500 clusters - PV of Net GMIB Costs



Kolmogorov-Smirnov Test

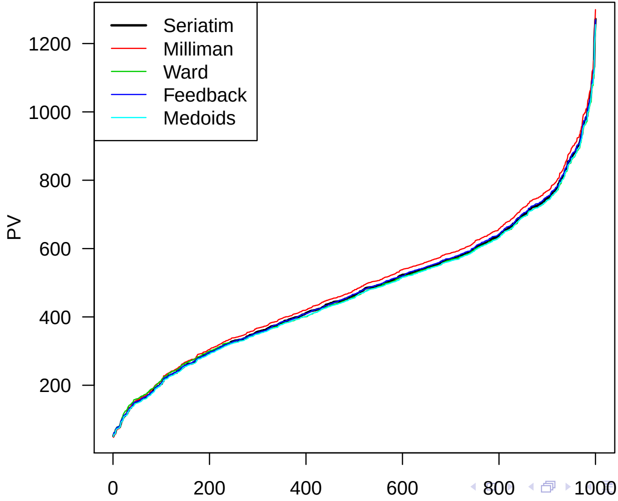
- The two-sample Kolmogorov-Smirnov test compares the distributions of data from two samples.
- Null hypothesis: both come from the same distribution.
- The test statistic, and hence the p-value, quantifes the maximum absolute difference between the two empirical sample CDFs over the range of values in the samples.
- The closer the p-value is to 1, the more similar the two samples are.

Out-of-sample results for 2500 clusters - PV of Net GMIB Costs

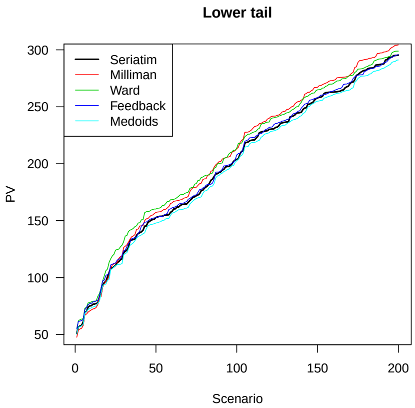
Table : *P-value from Kolmogorov-Smirnov tests for present value of net GMIB cost.*

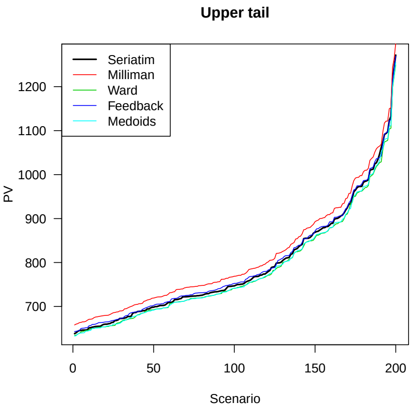
Seriatim	1.000
Milliman	0.181
Ward	1.000
Feedback	0.794
K-medoids	0.888

Out-of-sample results for 2500 clusters - PV of Net M&E Fee Income



Out-of-sample results for 2500 clusters - PV of Net M&E Fee Income





Out-of-sample results for 2500 clusters - PV of Net M&E Fee Income

Table : *P-value from Kolmogorov-Smirnov tests for present value of net GMIB cost.*

Seriatim	1.000
Milliman	0.241
Ward	0.980
Feedback	1.000
K-medoids	0.954

Conclusion

- Freedman & Reynolds (2008)'s original approach is not necessarily the optimum method for clustering when compressing actuarial data.
- A model-based approach appears promising as an alternative, particularly when the number of clusters is small.
- Ward's minimum variance hierarchical clustering method and k-medoids clustering both outperform Milliman's method for large numbers of clusters.

Further work?
Optimizing the Approach for General Insurance

- So far we have only clustered data based on continuous numerical variables. What about nominal and ordinal variables such as gender and car type?
- McParland & Gormley (2014) developed **clustMD** to perform model-based clustering for such mixed data.
- It would be possible to integrate the mixed data methodology with the size-weighted nature of the actuarial data original approach for continuous variables.
- Nominal and ordinal variables can then be modelled directly, or used to power the feedback sampling approach in randomly selecting data subsets for clustering.

Key References

- Reynolds, C. & Freedman, A. (2008) Cluster Analysis: A Spatial Approach to Actuarial Modelling, accessed September 2013, `http://www.milliman.com/uploadedFiles/insight/research/life-rr/cluster-analysis-a-spatialrr08-01-08.pdf`
- Raftery, A.E. & Fraley, C. (2002) Model-based Clustering, Discriminant Analysis and Density Estimation *Journal of the American Statistical Association* 97, 611-631.