A Model-based Approach to Clustering for Data Compression in Actuarial Applications

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

July 2014

Existing Approach

- We have a dataset of 110,000 policies with 55 '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 becuase it is not feasible to compute a large range of scenarios for the full dataset.

Existing Approach

FIGURE 1

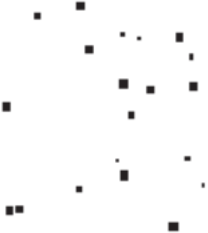


FIGURE 2

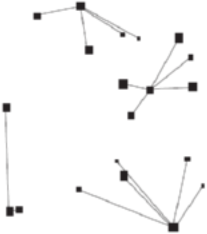
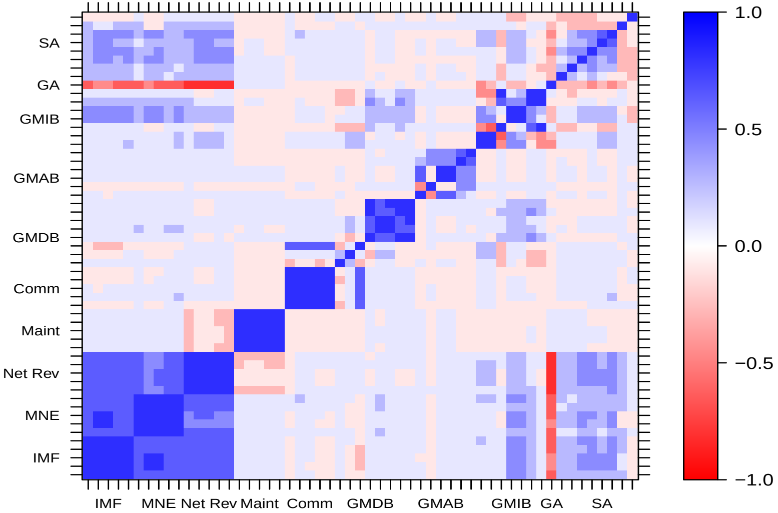


FIGURE 3

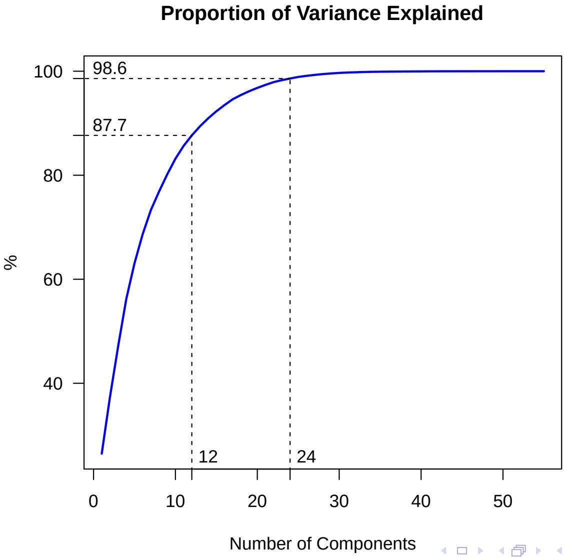
Existing Approach

- Current practice is to use size-weighted hierarchical clustering - iteratively merging the 'least important' policy with its nearest neighbour until the desired number remain.
- If we use a model-based approach to cluster the data, will the resulting representative policies replicate the behaviour of the full dataset more accurately over a range of scenarios?
- Test at various levels of compression - 50, 250, 1000, 2500 and 5000 clusters.

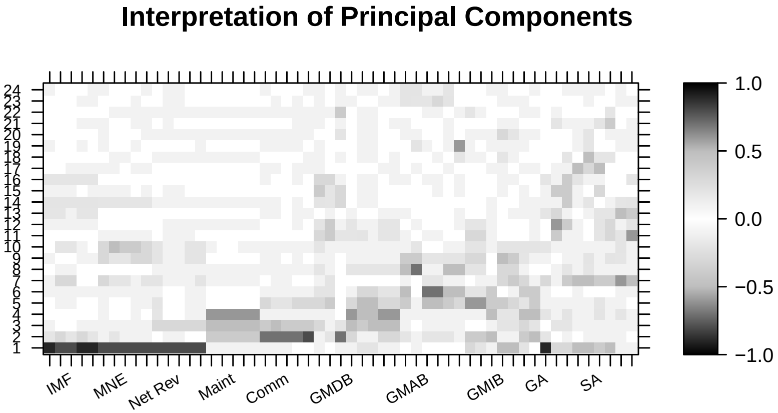
Weighted Correlation of Location Variables



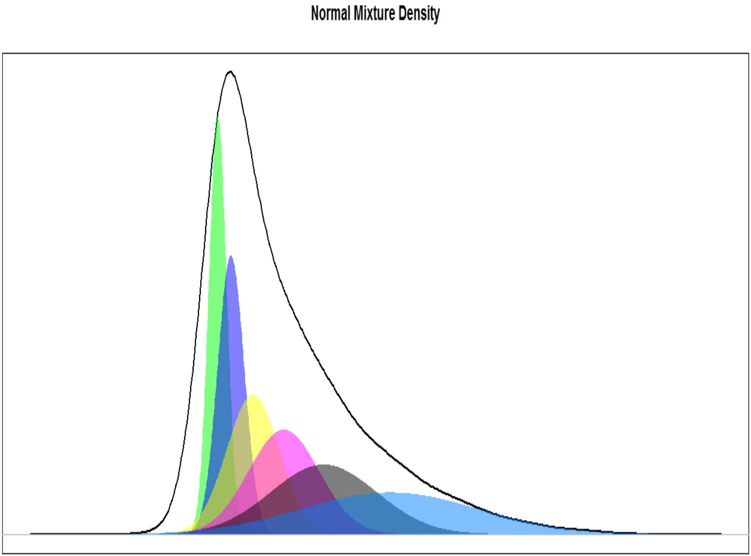
PCA


Proportion of Variance Explained
%
100
80
60
40
98.6
87.7
12
24
0
10
20
30
40
50
Number of Components

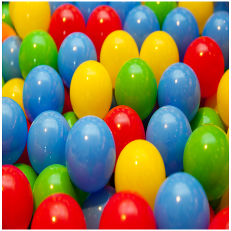
PCA


Interpretation of Principal Components
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
IMF
MNE
Net Rev
Maint
Comm
GMDB
GMAB
GMIB
GA
SA
1.0
0.5
0.0
−0.5
−1.0

Model-based Clustering



Model-based Clustering

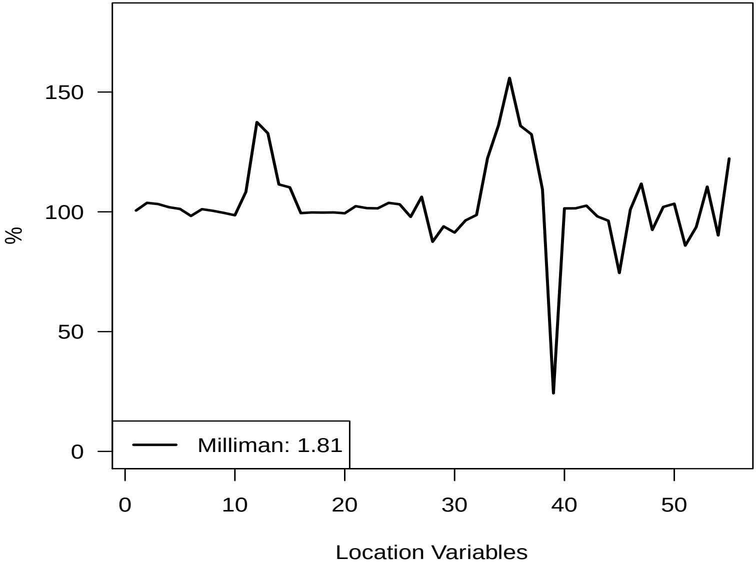


EII

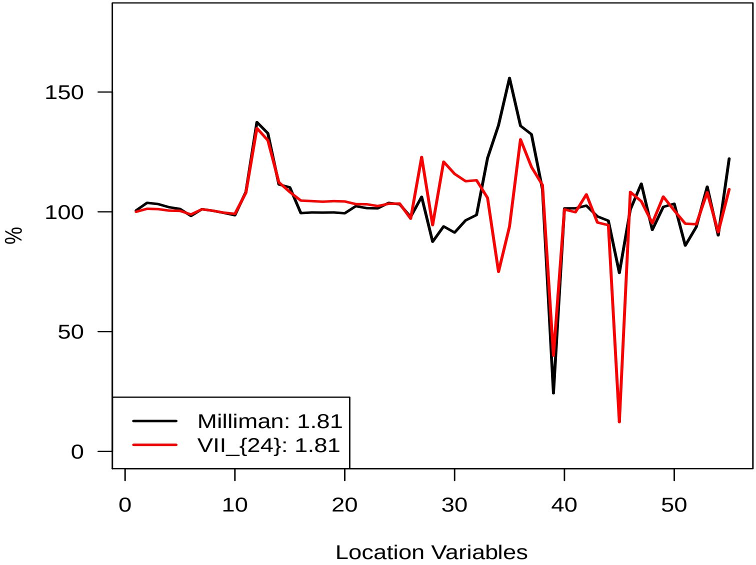
VII

VVV

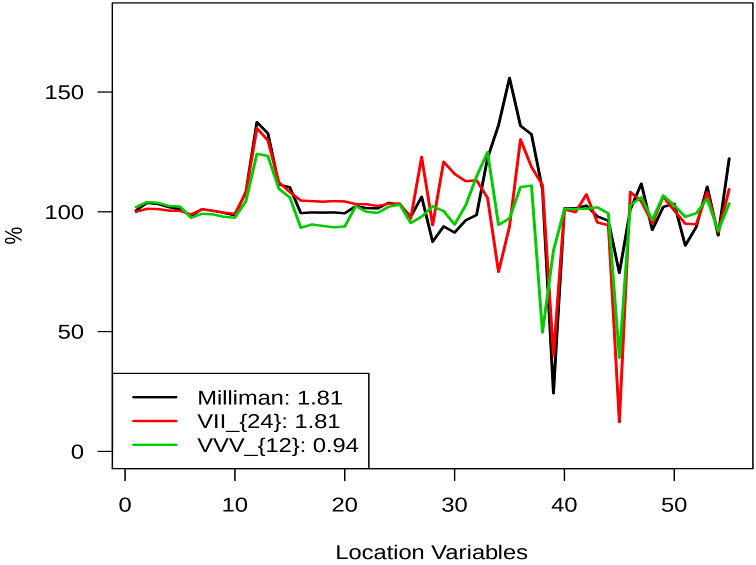
50-Cluster Solutions



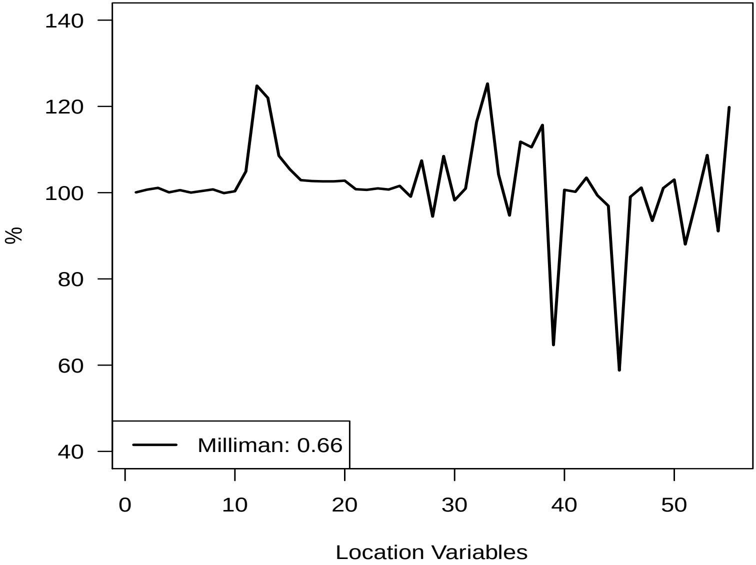
50-Cluster Solutions



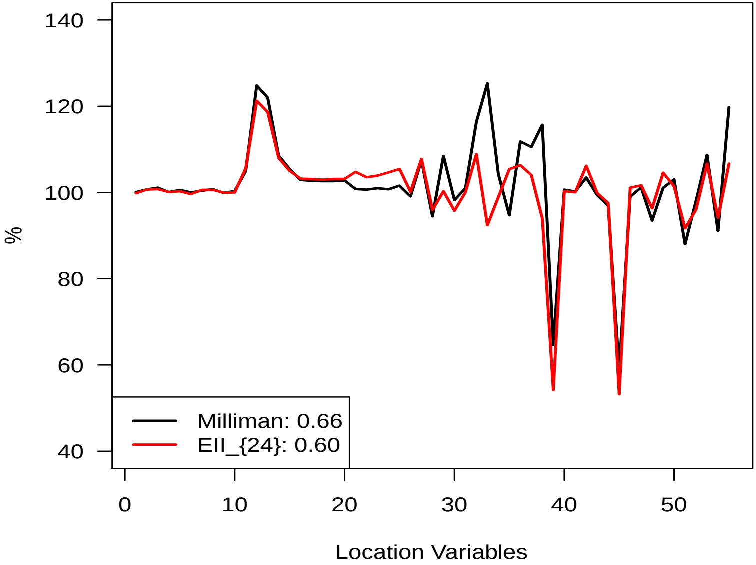
50-Cluster Solutions



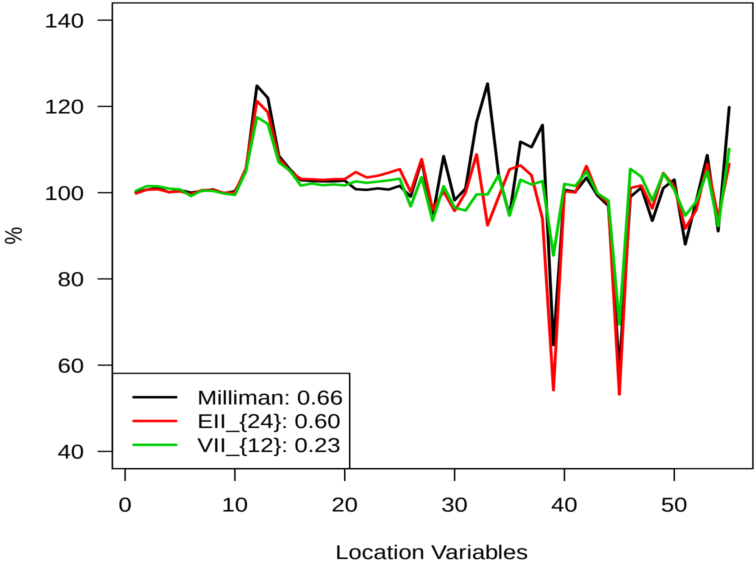
250-Cluster Solutions



250-Cluster Solutions



250-Cluster Solutions



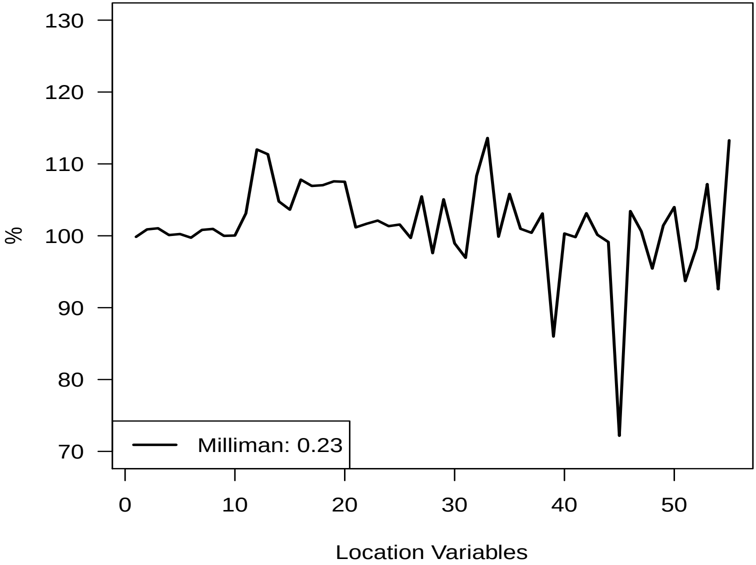
Fitting Larger Numbers of Clusters

- 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 24 location variables would require over a million parameters.
- Feedback Sampling is an approach we have developed that takes advantage of the size-weighted nature of the data to obtain an EII solution.

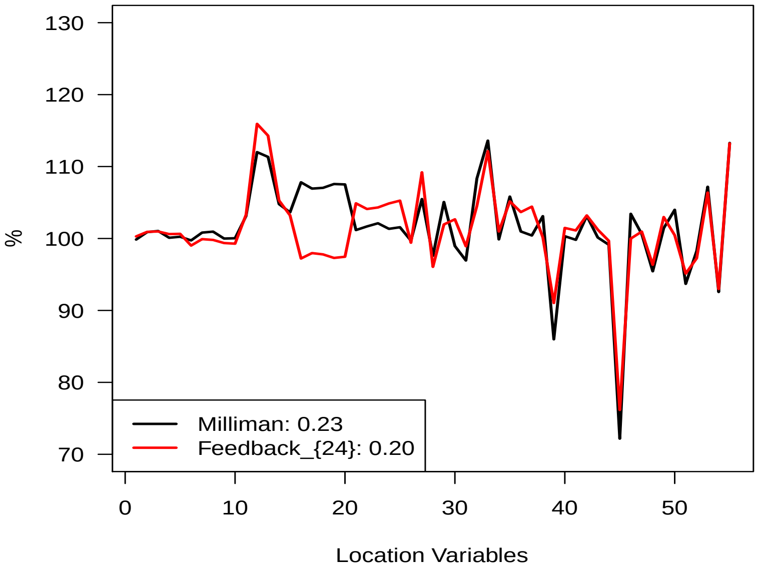
Fitting Larger Numbers of Clusters - Feedback Sampling

1. Randomly sample 2020 policies and fit a 20-cluster model.
2. Treat the resulting cluster centres as 20 individual policies, scaled up by the sums of the sizes of the policies in each.
3. Replace the 2020 policies in the full dataset with these 20 scaled-up cluster centres.
4. Repeat until the desired number of cluster centres remain.
5. Then simply assign each policy to the cluster whose centre is closest.

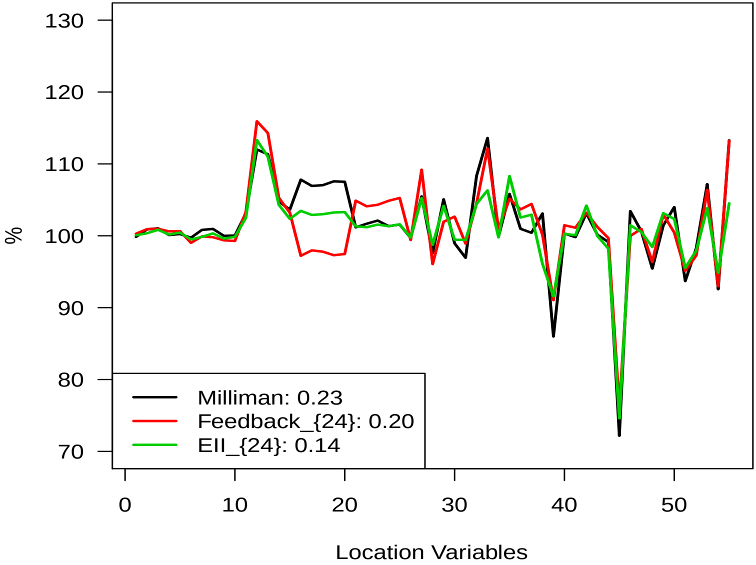
1000-Cluster Solutions



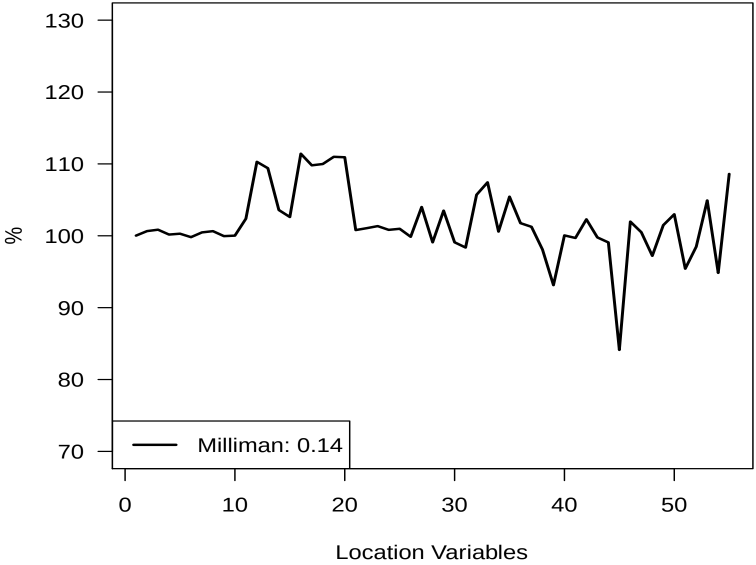
1000-Cluster Solutions



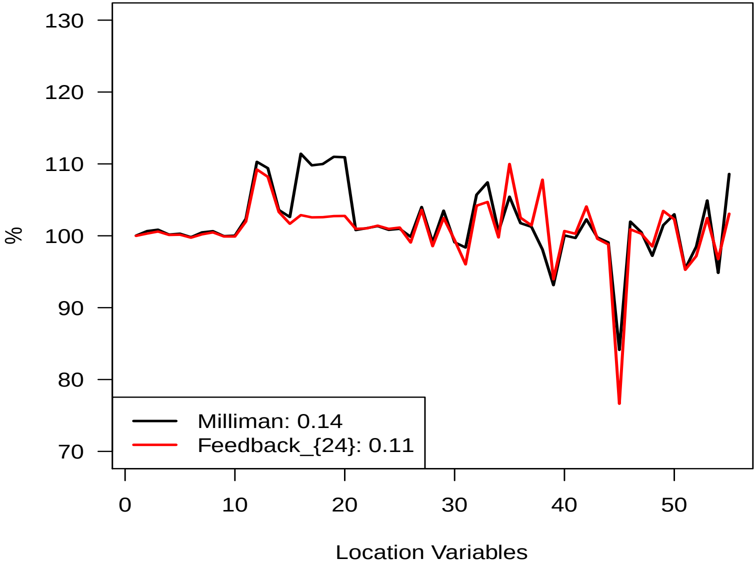
1000-Cluster Solutions



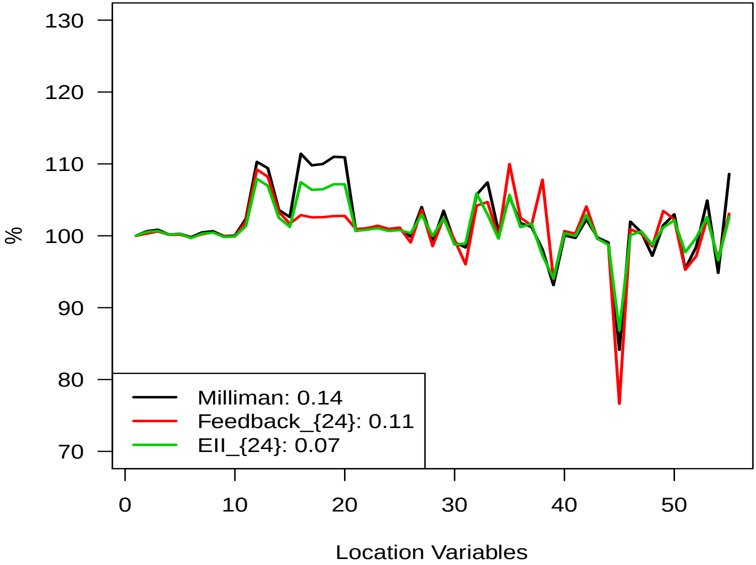
2500-Cluster Solutions



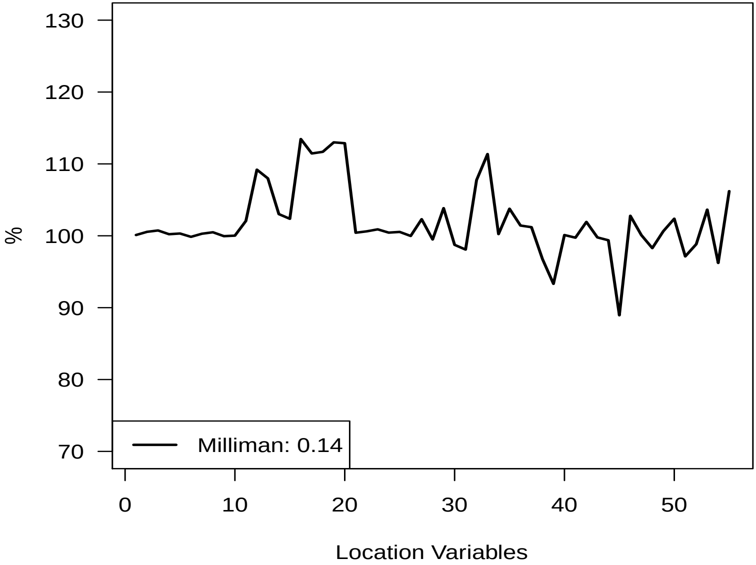
2500-Cluster Solutions



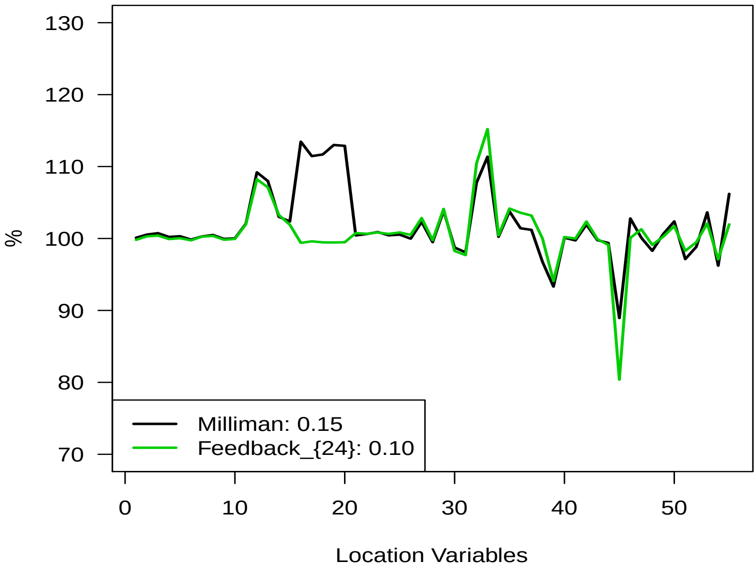
2500-Cluster Solutions



5000-Cluster Solutions



5000-Cluster Solutions



Conclusion & Further Work

- A model-based approach appears promising as an alternative to the non-parametric, hierarchical clustering method for compressing actuarial data.
- Testing results - how do the model-based compressed datasets perform in simulating values over a large range of scenarios, relative to both the hierarchically compressed datasets and to the full dataset?