Clustering in Ratemaking: Applications in Territories Clustering

Ji Yao, Ph.D.

Abstract: Clustering methods are briefly reviewed and their applications in insurance ratemaking are discussed in this paper. First, the reason for clustering and the consideration in choosing clustering methods in insurance ratemaking are discussed. Then clustering methods are reviewed and particularly the problem of applying these methods directly in insurance ratemaking is discussed. An exposure-adjusted hybrid (EAH) clustering method is proposed, which may alleviate some of these problems. Results from EAH approach are presented step by step using the U.K. motor data. The limitations and other considerations of clustering are followed in the end.

Keywords: Clustering, ratemaking, generalized linear modeling, territory analysis, data mining.

1. INTRODUCTION

Clustering is the unsupervised classification of patterns into groups [1]. It is widely studied and applied in many area including computer science, biology, social science, and statistics. A significant number of clustering methods were proposed in literature [1]-[6]. In the context of actuarial study, [7]-[9] studied possible application of clustering in insurance. As to the territory ratemaking, [10] considered the use of geographical information system. However, a thorough analysis of clustering in insurance ratemaking is not known to this author.

The purpose of this paper is two-fold. The first part of the paper introduces the basic idea of clustering and state-of-the-art clustering methods. Due to the large amount of methods, however, it is not intended to give a detailed review of every clustering method in the literature. Rather, the focus is on the key idea of each method and, more importantly, their advantages and disadvantages when applied in insurance ratemaking.

In the second part, a clustering method called exposure-adjusted hybrid (EAH) clustering is proposed. The purpose of this section is not to advocate one certain clustering method, but to illustrate the general approach that could be taken in territory clustering. Because clustering is subjective, it is well recognized that most details should be modified to accommodate the feature of data-set and the purpose the clustering.

The remainder of this paper proceeds as follows: Section 2 introduces clustering and its application in insurance ratemaking; section 3 reviews clustering methods and their applicability in insurance ratemaking; section 4 proposes the EAH clustering method and illustrates this method step-by-step

using U.K. motor data; section 5 discusses some other considerations; and section 6 draws conclusions. Some useful references are listed in Section 7.

2. OVERVIEW OF CLUSTERING

2.1 Introduction to Clustering

The definition of clustering is not unique. Generally, *clustering* is the process of grouping a set of data objects into a cluster or clusters so that the data objects within the cluster are very similar to one another, but are dissimilar to objects in other clusters [3]. Usually a *similarity measure* is defined and the clustering procedure is to optimize this measure locally or globally.

It is important to understand the difference between clustering and discriminant analysis. In discriminant analysis, we have a set of pre-classified samples, which could be used to train the algorithm to learn the description of each class. For example, we have a set of claims, some of which are fraud claims. These fraud cases are used to train the algorithm to find a rule that predicts the probability of fraud claims in future cases. However, in the case of clustering, these pre-classified samples are not available. So all these rules have to be derived solely from data, indicating that clustering is subjective in nature.

With so many clustering methods available in literature [1]-[6], it is a very difficult task to choose the appropriate method. Two considerations are the purpose of clustering and the feature of dataset.

2.2 Purpose of Clustering in Insurance

There are many reasons to use clustering in insurance ratemaking. The first reason is to better understand the data. After grouping data object into clusters, the feature of each cluster is clearer and more meaningful. For example, it is useful to cluster similar occupations and analyze their claim experience together.

The second reason is to reduce the volatility of data and to make the rates stable over time. Because the amounts of data are usually limited over a certain period, historical data in each segment may show high volatility. In ratemaking, if analysis is only based on experience of each single segment, the resulting rates will be volatile as well. Appropriate clustering alleviates this problem.

The third reason is to reduce the number of levels in rating factors. For example, in ratemaking for vehicles, it is possible to have rates for each individual vehicle type, probably because enough historical data have been collected over a long period. However, this may be difficult to implement and usually similar vehicles will be clustered together.

And lastly, the fourth reason is to make the rate are reasonable and smooth the rates. For example, in territory ratemaking there may be marketing, regulatory, or statute limitations requiring that adjacent territories have similar rates. Some clustering methods may reduce the probability that the rate of one location is substantially different than a neighboring area.

2.3 Nature of Insurance Dataset

The nature of data is also critical in choosing clustering method. The type of insurance data is usually numerical, such as claim frequency and severity. Some information that is originally expressed in non-numerical format, such as postcode, can be translated into a numerical format. However, in some cases such translation may be not possible; one example is occupation. The focus of this paper is on numerical data or data that could be translated numerical format.

Insurance data usually is multi-dimensional; some are related to risk characteristics and some are related to the rating factors that need to be clustered. For example, in territory clustering there may be one dimension of claim frequency and two dimensions of longitude and latitude. However, in most cases the dimension would not be too high. It is well understood that high-dimension clustering is very different to low-dimension [1]-[6]. The focus of this work is on low-dimension.

The data usually have a large noise because of the uncertainty of insurance results. Ideally, we should use expected claim frequency or expected loss in clustering. However, only the observed claim experience is available for analysis. This uncertainty should be considered in designing the similarity measure.

The insurance data also may not be well-separated and the change between clusters could be gradual. For example, in territory clustering, the difference in risk characteristics between two adjacent areas usually is quite small. So the task is to find the boundaries where the difference is relatively large. This indicates that some methods that require data well-separated may be not suitable.

3. CLUSTERING METHODS

There are many clustering methods in literature and detailed reviews are in references [1]-[6]. In this section, each method is briefly introduced, focusing on its applicability in insurance ratemaking.

3.1 Partitioning Methods

Broadly speaking, partitioning method organizes the data objects into a required number of clusters that optimizes certain similarity measure. However, it is usually narrowly defined as a method that is implemented by an iterative algorithm where the similarity measure is based on the distance between

data objects. In the context of insurance ratemaking, the distance could be the difference in claim frequency/severity or the difference in the numerical rating factors or a combination of these two.

Generally, the algorithm of partitioning methods is as follows:

(i) choose initial data objects randomly as a center or a representation of clusters;

(ii) calculate the membership of each data object according to the present center or a representation of clusters;

(iii) update the center or representation of clusters that optimizes the total similarity measure;

(iv) repeat step (ii) if there is a change in the center or representation of clusters; otherwise stop.

There are different methods to be used depending on how the similarity measure is chosen and how the center or the representation of clusters is defined.

3.1.1 *K*-Means Method

The center of the cluster m_i, is defined as the mean of each cluster C_i, that is,

$$m_i = \frac{1}{n_i} \sum_{\mathbf{x} \in C_i} \mathbf{x}$$

where $\mathbf{x}$ is the data object and is n_i the number of data objects within the cluster C_i. The total similarity measure is the squared-error function around the center of each cluster, i.e.

$$f = \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} \left| \mathbf{x} - m_i \right|^2 ,$$

where k is the number of clusters.

This method is easy to understand and apply. The time complexity of this method is also lower than *k*-Medoids method. Generally it is one of the most popular clustering methods. However, it is very sensitive to noise and outliers because the mean of data objects in each cluster is used to represent each cluster. This is a big problem for insurance data as outliers are always expected. It is also difficult to choose the appropriate number of clusters. This may not be critical, however, because in insurance ratemaking the number of cluster may be determined by other factors such as IT limitations. The silhouette coefficient [1]-[4] is also introduced to solve this problem. The results of *K*-Means method tend to be sphere-shaped because the squared-error function is used as similarity measure. This applies to most methods that use distance as a similarity measure. This drawback is quite critical in territory clustering, as the nature cluster is not necessarily sphere-shaped. This method does not work very well when the density of data changes. Finally, the efficiency of this method is greatly affected by the initial setting and sometimes it may only converge to a local optimal. In practice, this may be solved by running the program several times with different initial

settings.

3.1.2 *K*-Medoids Method

K-Medoids method is similar to *K*-Means method but it defines the most centrally located data object of cluster C_i as the cluster center to calculate the squared-error function. Because of this, this method is less sensitive to noise and outliers. However, the procedure to find the most centrally located object requires a much higher run time than *K*-Means method [1]-[4]. This basic method is named partition around medoids (PAM) method. Clustering large application (CLARA) and clustering large applications based upon RANdomized search (CLARANS) were later proposed to reduce the time complexity [1]-[4]. However, these methods are still subject to other problems as *K*-Means method.

3.1.3 Expectation Maximization (EM)

Rather than representing each cluster by a point, this method represents each cluster by a probability distribution. In step (i), each cluster is represented by a default probability distribution. In step (ii), the probabilities of each data object belonging to every cluster C_i are calculated by the probability distribution representing cluster C_i. Then every data object is assigned to the cluster that gives the highest probability. In step (iii), the probability distribution is then re-calculated for each cluster based on new members of each cluster. If there is change in probability distribution that represents each cluster, then go to step (ii); otherwise, stop the iteration.

The time complexity of EM is lower than *K*-Medoids, but it has most of the problem *K*-Means suffers. What's more, the choice of probability distribution gives rise to more complexity.

3.2 Hierarchical Methods

Hierarchical method creates a hierarchical decomposition of the given set of data objects forming a dendrogram, a tree that splits the dataset recursively into smaller subsets. So the number of clusters is not chosen at the early stage of analysis in this method.

3.2.1 AGglomerative NESting (AGNES) and DIvisia ANAlysis (DIANA)

Both of these methods are earlier hierarchical clustering methods, where AGNES is bottom-up method and DIANA is top-down method. In AGNES, clustering starts from sub-clusters that each includes only one data object. The distances between any two sub-clusters are then calculated and the two nearest sub-clusters are combined. This is done recursively until all sub-clusters are merged into one cluster that includes all data objects. In DIANA, clustering starts from one cluster that includes all data objects. Then it iteratively chooses the appropriate border to split one cluster into

two smaller sub-clusters that are least similar.

Slightly different from the object-to-object definition of similarity measure in partitioning methods, the similarity measure in hierarchical method should be cluster-to-cluster. Different similarity measures of two clusters can be defined and common ones are

1. Min distance: $d_{\min}(C_i, C_j) = \min_{p_i \in C_i, p_j \in C_j} d(p_i, p_j)$, where $d(\cdot,\cdot)$ is a similarity measure of two data objects p_i, p_j and C_i, C_j are two clusters;
2. Max distance: $d_{\max}(C_i, C_j) = \max_{p_i \in C_i, p_j \in C_j} d(p_i, p_j)$;
3. Average distance: $d_{\text{avg}}(C_i, C_j) = \frac{1}{n_i n_i} \sum_{p_i \in C_i, p_j \in C_i} d(p_i, p_j)$,

where n_i and n_j are the size of clusters C_i and C_j, respectively.

The concept is easy to understand and apply. The resulting clusters are less sphere-shaped than partitioning methods, but still have that tendency because distance is used as similarity measure. The number of clusters is also chosen at a later stage, which is better than partitioning methods. The performance regarding noise and outlier depends on the similarity measure chosen. However, the most critical problem of the AGNES and DIANA methods is that the over-simplified similarity measure often gives erroneous clustering results, partly because the hierarchical method is irreversible. Another problem is that the complexity of time, which depends on the number of data objects, is much higher than that of *K*-Means method.

3.2.2 Balanced Iterative Reducing and Clustering using Hierarchies (BIRTH)

The method's key idea is to compress the data objects into small sub-clusters in first stage and then perform clustering with these sub-clusters in the second stage. In the second stage, the AGNES or DIANA methods could be used, while in actuarial literature BIRTH is specifically named after the method that uses a tool called clustering feature (CF) tree [3].

One advantage is that it greatly reduces the effective number of data objects that need to cluster and reduces the time complexity.

However, it still tends to have spherical shape clustering because similarity measure has the same definition as AGNES or DIANA.

3.2.3 Clustering Using REpresentatives (CURE) and CHAMELEON

The key idea of CURE is to use a fixed number of well-scattered data objects to represent each cluster and shrink these selected data objects towards their cluster centers at a specified rate. Then the two clusters with the closest "distance" will be merged. The "distance" can be defined in any way as in Section 3.2.1.

Compared with AGNES and DIANA, CURE is more robust to outliers and has a better performance when clusters have non-spherical shape. However, all parameters, such as number of representative data points of a cluster and shrinking speed, have a significant impact on the results, which makes this method difficult to understand and apply.

In CHAMELEON method, instead of distance, more sophisticated measures of similarity such as *inter-connectivity* and *closeness* are used. CHAMELEON also uses a special graph partitioning algorithm to recursively partition the whole data objects into many small unconnected sub-clusters.

CHAMELEON is more efficient than CURE in discovering arbitrarily shaped clusters of varying density. However, the time complexity, which is on order of square of number of data objects, is quite high [1]-[4].

The ability to create arbitrarily shaped clusters makes these two methods quite attractive in territory clustering. However both methods are too complicated to apply and therefore not further developed in this paper.

3.3 Density-Based Methods

Most partitioning and hierarchical methods use the similarity measure based on distance. However, density could be used as similarity measure. For example, an intuitive understanding of this method is that satellite towns around a big city can often be clustered with the big city while rural areas are not clustered.

3.3.1 Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

This method defines the density of a data object as the number of data objects within a certain distance of the data object. If the data object's density is high, the data object is very similar to its neighbors and should be clustered with those neighboring data objects. This is exactly the basic idea in DBSCAN method. After calculating the density of every data object, clusters are generated by several rules—the basic idea is to expand every cluster as long as the density of the neighboring data object is higher than the threshold. Outliers are discarded and not grouped to any clusters.

The advantage of this method is that it could find arbitrary shape of clusters. However, the efficiency of this method largely depends on parameters chosen by the user, so it requires a high level of expertise to apply this method successfully. Also it does not work very well for a large or high-dimensional dataset, because the time complexity is very high in finding all those neighboring data objects and any intermediate results are not an approximation to final results.

3.3.2 Ordering Points To Identify the Clustering Structure (OPTICS)

Rather than producing a clustering of data objects for certain chosen parameters as in DBSCAN, this method produces a cluster ordering for a wide range of parameter settings. The user then can do clustering interactively by using the cluster ordering results.

Other than finding arbitrary shape clusters, this method solves the problem of dependency on parameters as in DBSCAN. However, it still has other problems as DBSCAN does.

3.3.3 DENsity-based CLUstEring (DENCLUE)

This method is efficient for large datasets and high-dimensional noisy datasets. It can also find arbitrarily shaped clusters, which makes it suitable for insurance ratemaking. However, there are many parameters to set and it may be difficult for the non-expert to apply. Details of this algorithm are discussed in [1]-[4].

3.4 Grid-Based Methods

These methods quantize the space into a finite number of cells that form a grid structure on which all of the clustering operations are performed. The basic grid-based algorithm defines a set of grid-cells, assigns data objects to the appropriate grid cell, and computes the density of each cell. After cells with densities below a certain threshold are eliminated, clusters are generated by combining adjacent groups of cells with similar densities or minimizing a given objective function.

The advantage of these methods is fast processing time, which is typically independent of the number of data objects and only dependent on the number of cells in each dimension of the quantized space. However, its disadvantage is that the shape of the cluster is limited by the shape of grid. But this problem can be reasonably overcome by smaller grid, which has become feasible because of the rapid development of computer. So this is a promising clustering method for insurance ratemaking.

STING, WaveCluster, and CLIQUE are three advanced grid-based methods, that differ in how information about data objects is stored in each grid or what cluster principle is used. STING explores statistical information, WaveCluster uses wavelet transform to store the information, and CLIQUE discovers sub-clusters using the a priori principle [1]-[4].

3.5 Kernel and Spectral Methods

Both of these are relatively new methods. Although they originated from different backgrounds, recent studies indicate that there is a possible connection between these two methods [5], [6].

The key idea of kernel method is to map the data into high-dimensional space called *feature space*, so

that non-linear feature in the low-dimensional space becomes linear in the feature space. The conventional clustering methods introduced in previous sections are then applied in the feature space.

The main tools for spectral clustering methods are graph Laplacian matrices [6] and associated eigenvectors, which are widely studied in spectral graph theory. The original data is first transformed into the *similarity matrix*, which is defined as the matrix of similarity measure and its eigenvectors. Then the conventional clustering methods, such as *K*-Means method, are applied on the similarity matrix or eigenvectors.

Although these methods appear to be easy to implement [6], they actually are not that easy for the non-expert to use. What's more, they seems to give no more advantages than other methods in the context of insurance ratemaking. Thus, these two methods will not be further explored.

4. EXPOSURE-ADJUSTED HYBRID (EAH) CLUSTERING METHOD

The choice of clustering method depends on the feature of the data and the purpose of clustering. Most of the methods introduced in Section 3 could be used in an appropriate situation. However, in insurance ratemaking, another consideration is that the method be easy to understand and use. Based on this philosophy, this paper is focused on how to modify the partitioning and hierarchical methods to accommodate the needs of insurance ratemaking.

The proposed exposure-adjusted hybrid (EAH) method is a combination of the partitioning and hierarchical methods. This method also adjusts the similarity measure by exposure to take account of the volatility of insurance data. The whole procedure, which has been customized to territory clustering, is as follows:

1. Use the generalized linear model (GLM) technique to model the claim experience;
2. Calculate the residual of the GLM results as the pure effect of territory;
3. Use the partitioning method to generate small sub-clusters that contain highly similar data points;
4. Use the hierarchical method to derive the dendrogram clustering tree;
5. Choose an appropriate number of clusters and get corresponding clusters;
6. Repeat steps 3-5 with different initial setting to find a relatively consistent pattern in clusters;
7. Use the territory clustering results to re-run GLM and compare the results with that of Step 1. If there is large difference in the resulting relativities from GLM, then start again from Step 1; otherwise stop.

4.1 Comments on the EAH Method

The purpose of steps 1 and 2 is to calculate the "pure" effect of territory. Because of the correlation between rating factors, the effect of territory cannot be calculated by simple one-way analysis. A common approach is to use the generalized linear model (GLM) [11]. However, because the output from territory clustering usually will be fed into GLM again to calculate the final relativities, there are two possible approaches:

Approach One: Include all rating factors other than territory in the first GLM and consider the residual as the pure effect of territory. Perform the clustering analysis and have the resulting clusters fed into second GLM, which includes all rating factors.

Approach Two: Include all rating factors, including a high-level group of territories in the first GLM and consider the residual as the pure effect of territory. Do the clustering analysis and have the resulting cluster fed into second GLM.

The problem with Approach One is that the relativities of other rating factors will change between the first and the second GLM because the second GLM includes a new rating factor. Approach Two has the same problem although to a less extent. In both case, it is necessary to compare the relativities of all the other rating factors between the two GLM. If it changes significantly the whole procedure should be repeated from Step 1, and this iteration stops when the relativities don't change much between the first and the second GLM. Because in most cases Approach Two has a lesser number of iterations, it is better to take this approach if possible.

Steps 3 to 6 are the clustering procedures. They could be replaced by any other methods discussed in Section 3. However, whichever methods are used, the definition of measure similarity must be modified to accommodate the feature of insurance data. Usually, each data point has at least two types of data: one is the measure of geographical information and the other is the measure of claim potential/risk characteristics. The most common measure for geographical information is Euclidean distance:

$$g\left(x_i, y_i, x_j, y_j\right) = \left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2,$$

Where (x_i, y_i) are longitude and latitude of data object *i* and (x_j, y_j) are those of the data object *j*. However, because of the curvature of the Earth's surface, some other definitions could be used. One such formula is the Haversine formula, which gives the shortest distance over the Earth's surface between two locations.

As for the claim potential/risk characteristics, one can use claim frequency, severity or burning cost. However, since the claim severity can be quite volatile in most cases, claim frequency is most commonly used. The similarity measure can be defined as Euclidean distance $(\mu_1 - \mu_2)^2$ where μ_1 and

μ_2 are claim frequencies. However, while Euclidean distance should be calculated by the expected claim frequency, only actual claim frequency is available for analysis. The uncertainty of the data must be considered in the definition of the similarity measure. One solution is that, if it is assumed that every risk in both territories have same variance σ^2, then the observed actual claim frequency μ_1 and μ_2 are approximately normal distributed with variance σ^2/E_1 and σ^2/E_2, where E_1 and E_2 are exposures in each territory, respectively. This assumption could be justified by the Central Limit Theorem. So the variance of μ_1- μ_2 is $\sigma^2/E_1 + \sigma^2/E_2$, which is used to adjust the Euclidean distance

$$f\left(\mu_1, E_1, \mu_2, E_2\right) = -\frac{\left(\mu_1 - \mu_2\right)^2}{\left(1/E_1 + 1/E_2\right)},$$

where σ^2 is dropped as it will be merged into the weight parameter introduced next.

Another question is how to combine the two measures. The solution proposed in this paper is to use the weighed sum of two similarity measures:

$$g\left(\cdot\right) + w \cdot f\left(\cdot\right) \quad .$$

This weight *w* has to be chosen tentatively and subjectively.

In step 3, the user chooses the number of small sub-clusters. Because of the high time complexity of hierarchical clustering method in step 4, this number cannot be too high. On the other hand, if the number of small sub-clusters is too low, the performance of the EAH method will deteriorate to a partitioning method. Numbers around one hundred could be used but it also depends on the purpose of clustering and the dataset feature.

The choice of the number of clusters in step 5 is also largely subjective and usually affected by other considerations, such as IT limitations or market practice. However, the general rule is not to put the threshold at the place where there is only a small change in the similarity measure. This will be illustrated later in a case study.

4.2 Case Study

In this case study we consider the territory clustering for ratemaking in motor insurance. The data we have are the geographical information in the form of postcode, other rating factors, exposures, and actual claim numbers.

In the U.K., it is a normal practice to use the postcode as a rating factor. The postcode is in a hierarchical structure and there are about 2 million postcodes. This amount is too large to analyze, so the data is first aggregated at the postcode district level, which has about 2,900 districts. All these postcode districts are then translated into longitude and latitude.

The difference between GLM-predicted claim frequency and the actual claim frequency is the

residual that will be clustered. This is plotted in Fig. 1. The color of black and blue shows the area where actual claim frequency is worse than predicted and the color of green and yellow shows where actual claim frequency is better. Other colors mean that actual frequency is similar to GLM prediction. Although overlapping, it is quite clear that there are clusters: the London and Midlands areas are the worst risks while the North and Wales are much better.

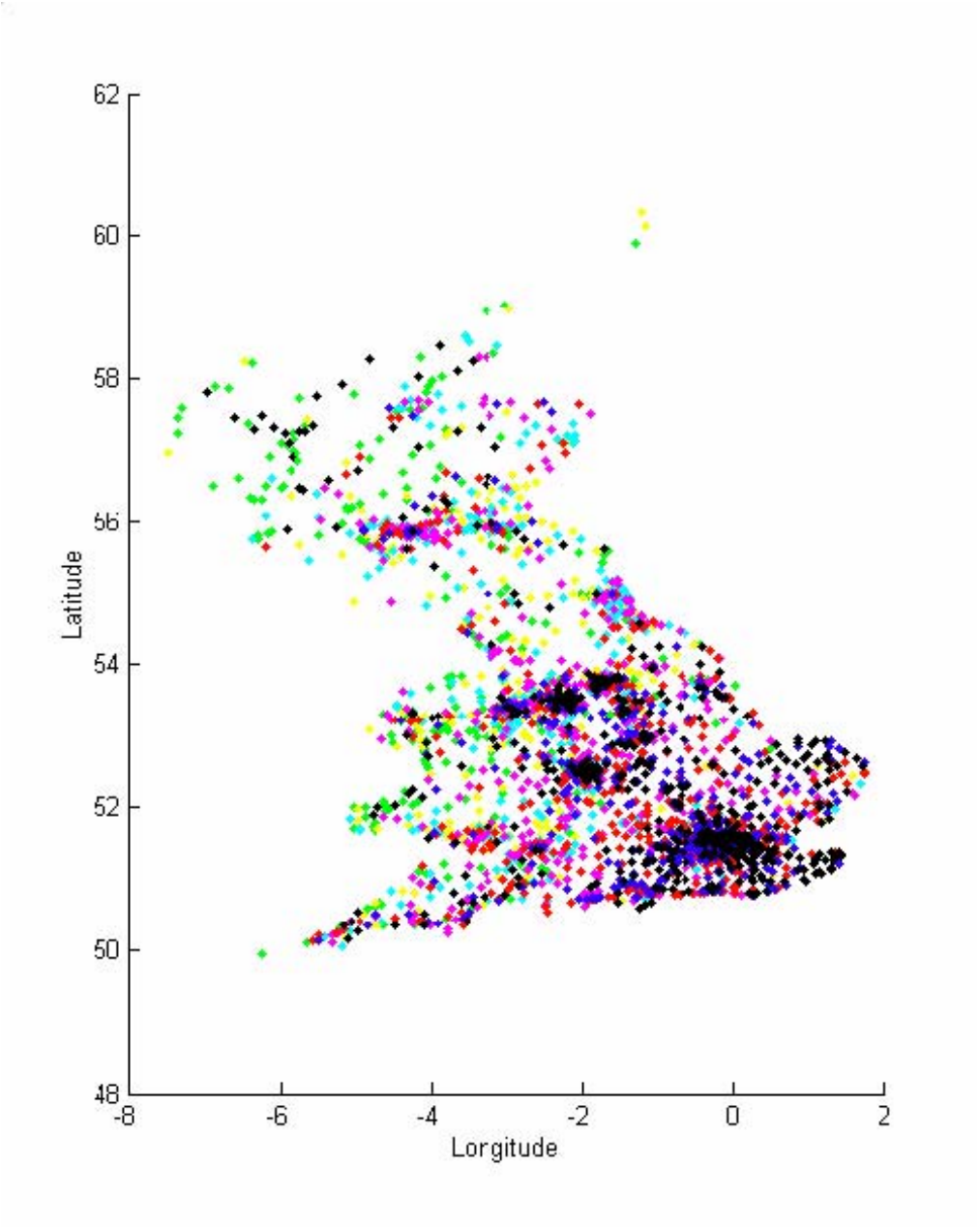

Fig. 1 Residual of GLM results that will be clustered

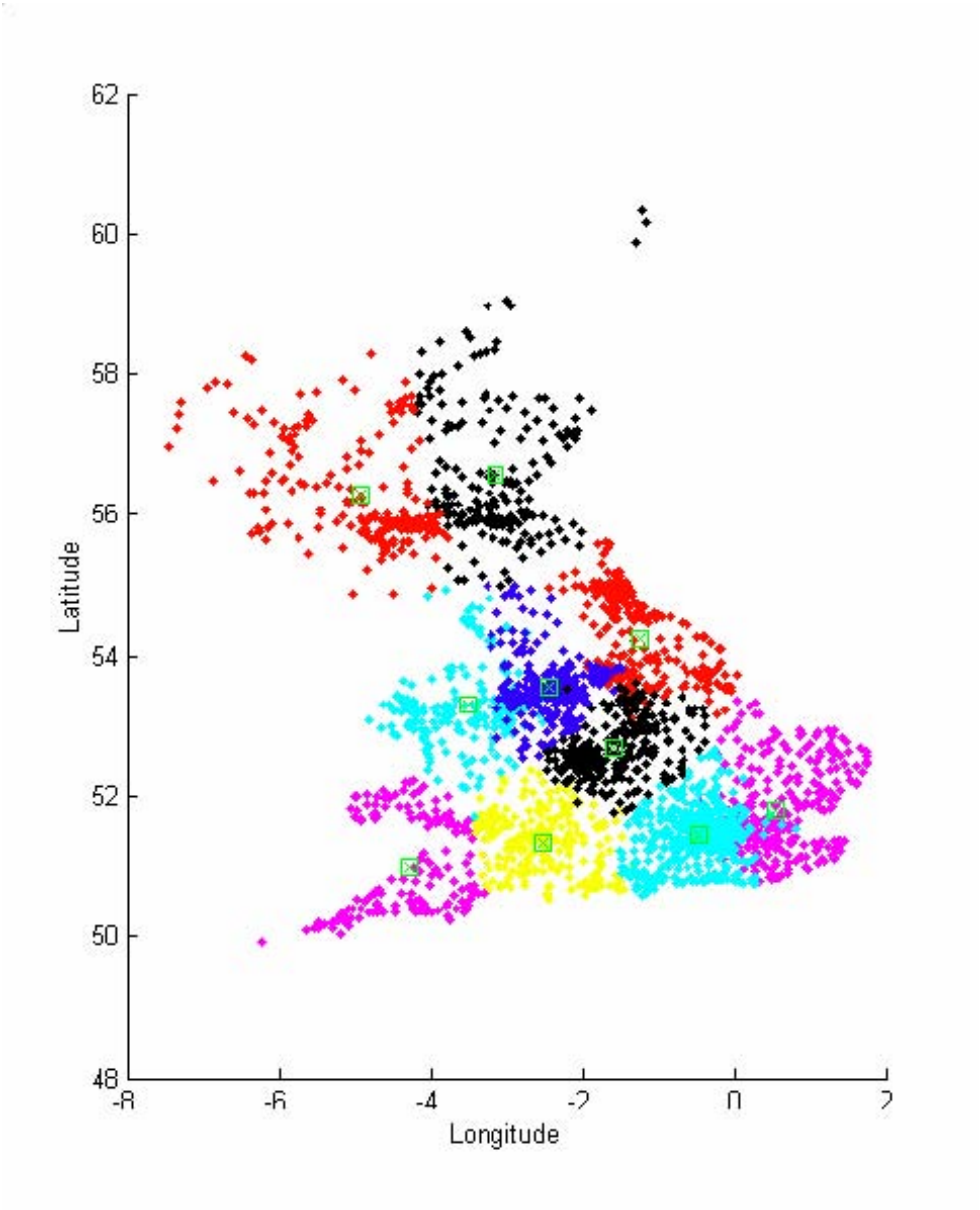

Fig. 2. Results of *K*-Means clustering method with $w = 1$.

4.2.1 Results of *K*-Means Method

The *K*-Means method is first applied and the results are plotted in Fig. 2-5 for different settings. In all cases, 10 clusters are generated and Fig. 2 plots the clustering results for the weighting parameter $w = 1$. Fig. 3 gives the result for the same weighting parameter $w = 1$ but with different initial settings. Although similar in the South, the results are significantly different in the North. It is very difficult to determine which one is better by looking at the initial dataset in Fig. 1.

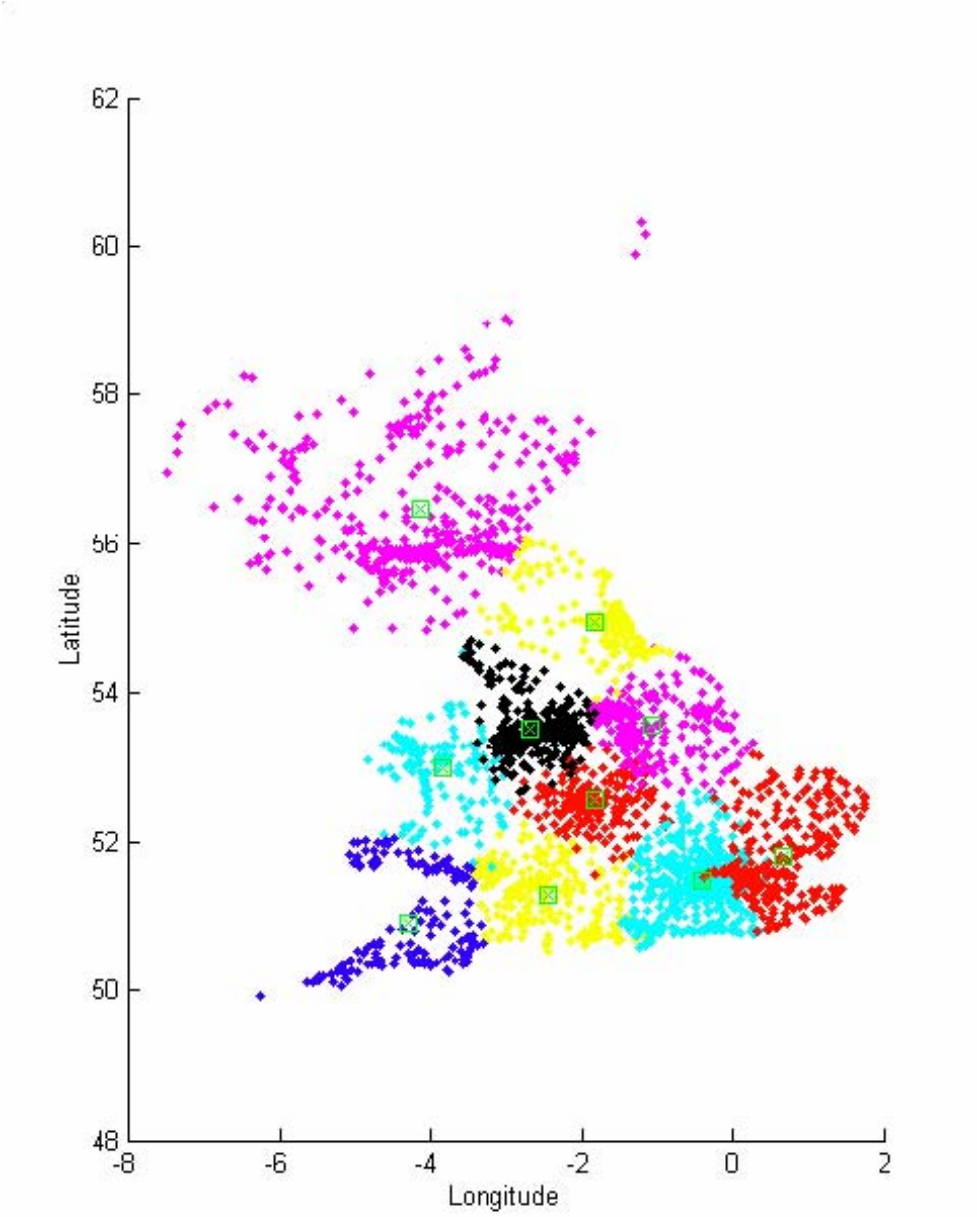

Fig. 3 Results of *K*-Means clustering method with $w = 1$ and different initial setting from Fig. 2.

The result of $w = 0.1$ is plotted in Fig. 4 and $w = 10$ is in Fig. 5. They each have the same initial settings as Fig. 2. The larger the parameter w is, the more weight that is put on the similarity measure in the claim experience and the less weight is put on the geographical closeness. Fig. 4 shows a much clearer border than Fig. 2, while Fig. 5 shows more overlapping. Probably, the result in Fig. 5 is not acceptable, but the choice between Fig. 2 and Fig. 4 is quite subjective.

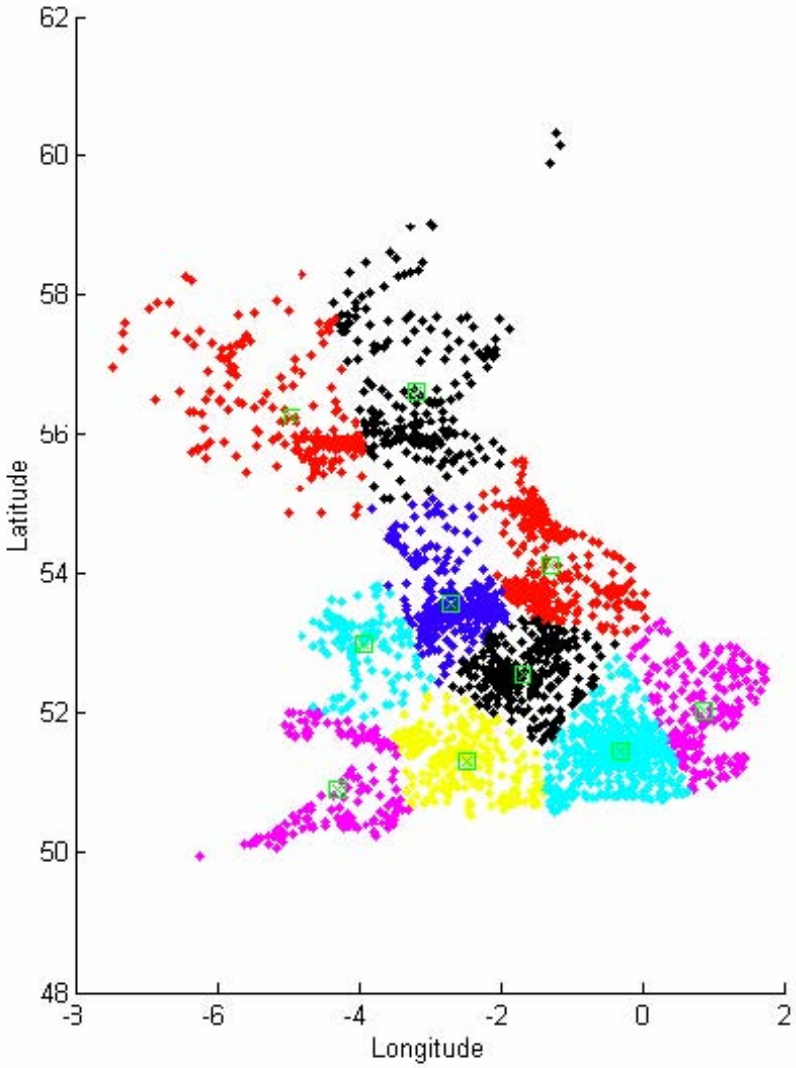

Fig. 4. Results of *K*-Means clustering method with $w = 0.1$.

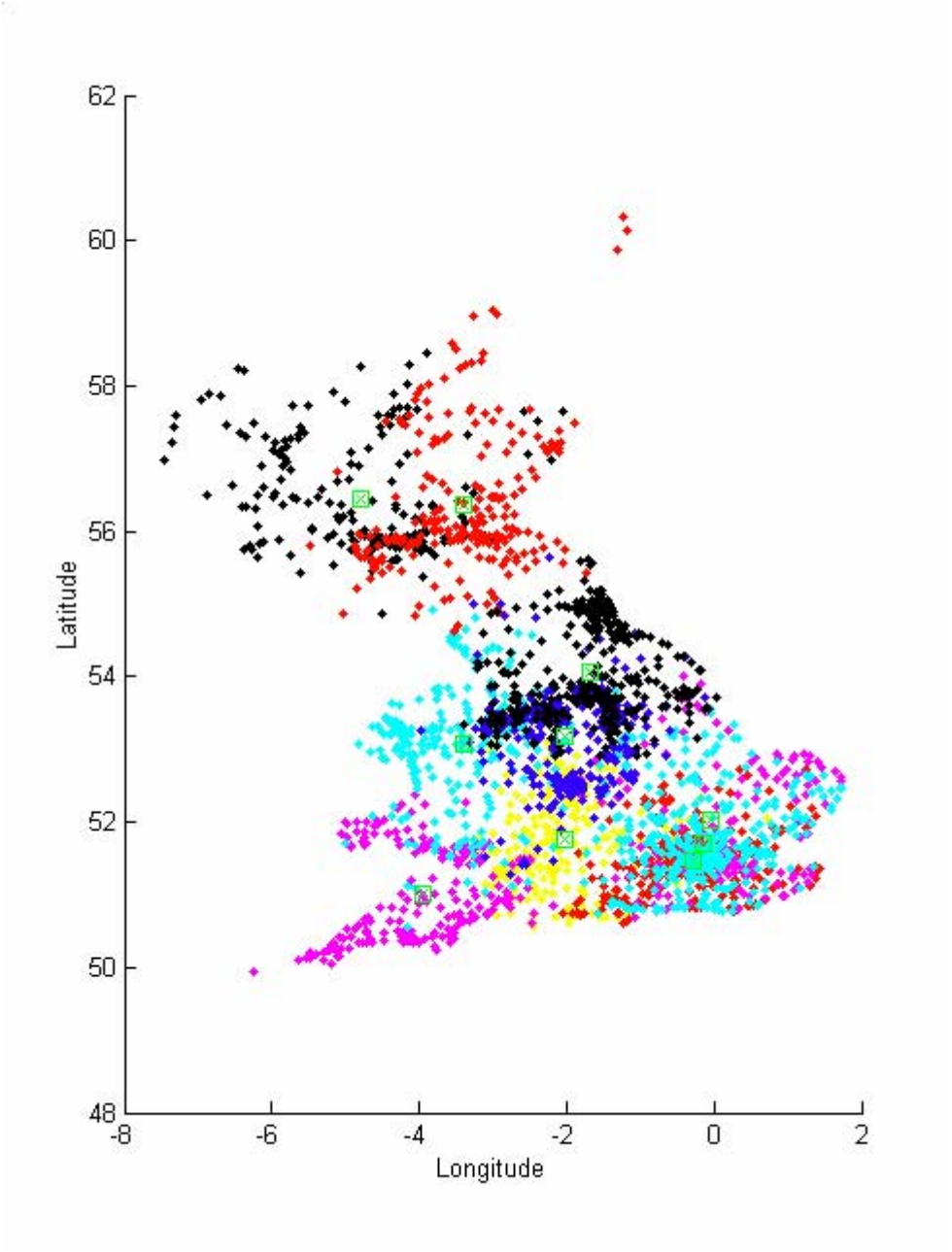

Fig. 5. Results of *K*-Means clustering method with $w = 10$.

These results highlight some features of *K*-Means methods. The sensitivity to initial settings is a big problem and the choice of parameters is also difficult. However, the dependence on parameters may be not a big problem as Fig. 2 looks very similar to Fig. 4.

4.2.2 Results of EAH Method

Then the results from each steps of the EAH method are presented. In step 3, 200 small sub-

clusters are generated. The output of the center of each cluster is shown in Fig. 6.

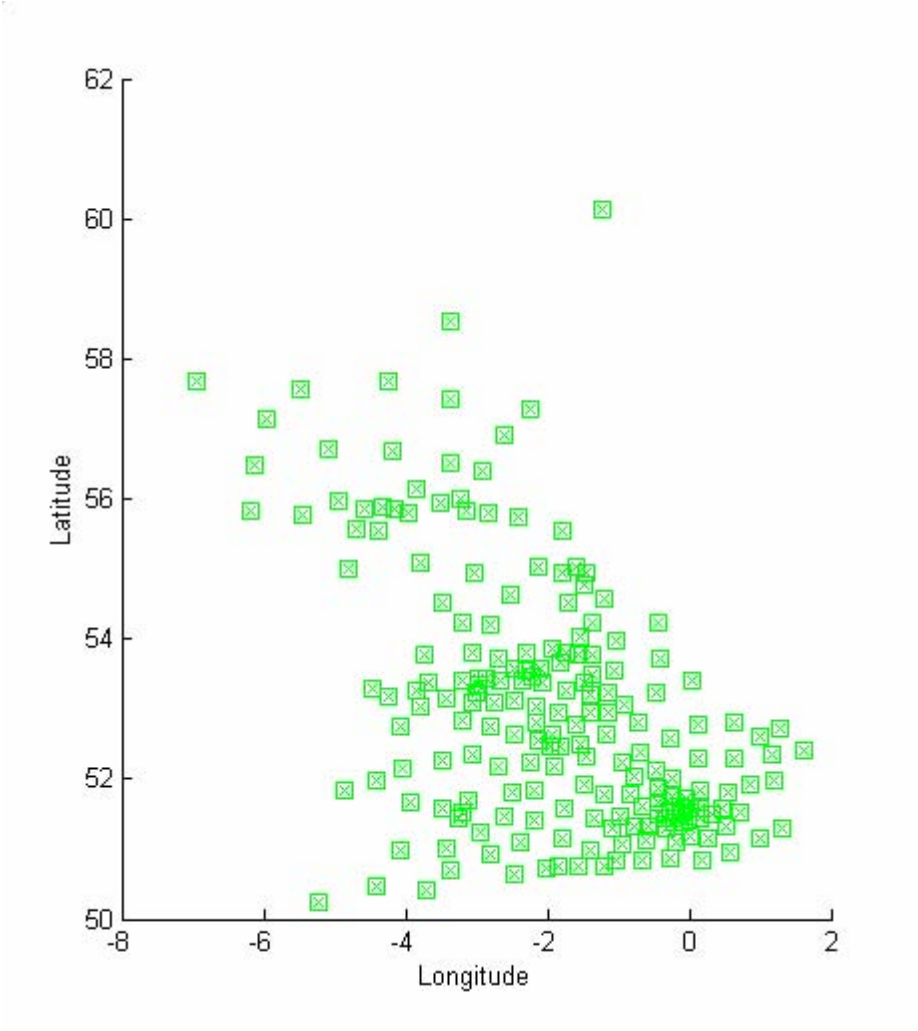

Fig. 6. Output from step 3 of 200 small clusters.

In step 4 of the hierarchical clustering method AGNES, the average distance defined in Section 3.2 is used in the weighted similarity measure proposed in Section 4.1. The resulting dendrogram is shown in Fig. 7, where the numbers of sub-clusters are not shown on the *x*-axis because there are too many to be shown in a readable format. The *y*-axis is the value of similarity measure that two sub-clusters are merged, which is termed *merging point*. The first 20 merging points are listed in Table 1. The first value is 29.0071, which means that if any two clusters with a similarity measure less than 29.0071 can be merged there will be only one cluster. Similarly, if any two clusters with similarity measures less than 4 can be merged, then there will be 8 clusters (because 4 is between 3.8875 and 4.7887).

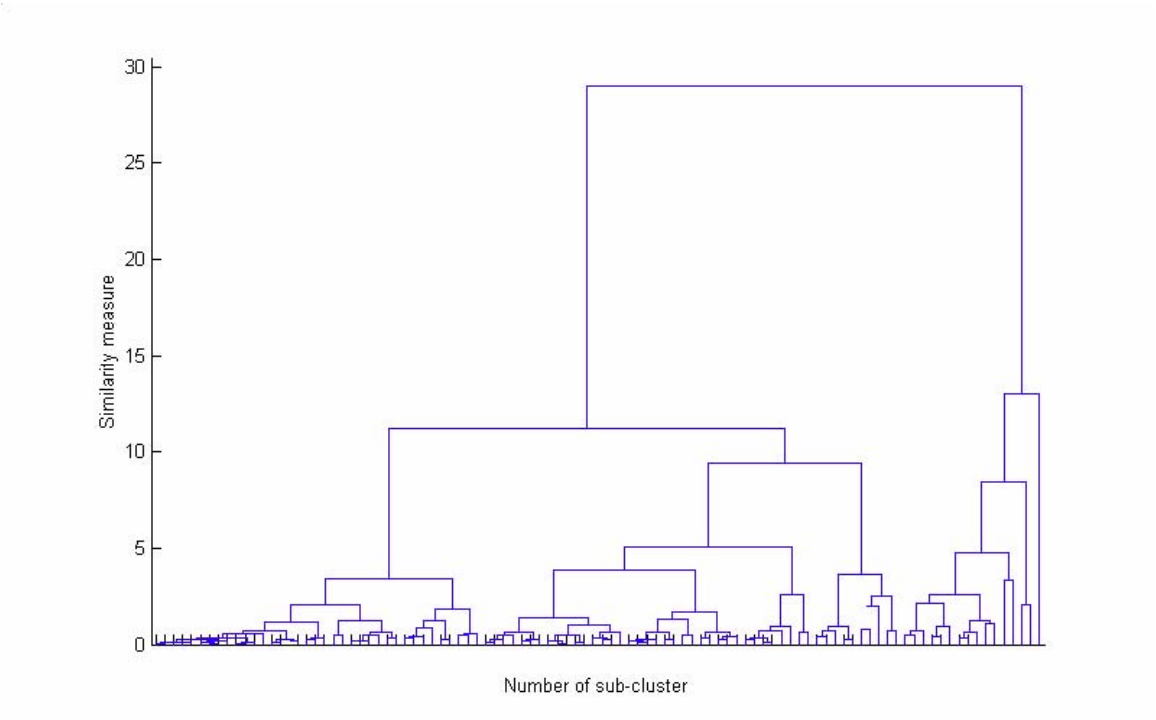

Fig.7 Dendrogram of clustering

In step 5, the number of clusters has to be chosen based on the dendrogram output. The general idea is not to put the threshold at the place where the change in similarity measure is small. If the change in similarity measures is quite small (for example, such as the gaps between numbers 10 and 11, 12 and 13, or 13 and 14 in Table 1) it is not very clear which two sub-clusters should be merged, making the results less reliable. Based on this rule, it is better to have 8 or 12 clusters in this case. The result of 12 clusters is plotted in Fig. 8.

Table 1. First 20 merging points.

Number	Similarity Measure	Change in Similarity Measure
1	29.0071	16.0212
2	12.9859	1.805
3	11.1809	1.7825
4	9.3984	0.9691
5	8.4293	3.3747
6	5.0546	0.2659
7	4.7887	0.9012
8	3.8875	0.2687
9	3.6188	0.2026
10	3.4162	0.0561
11	3.3601	0.7895
12	2.5706	0.0205
13	2.5501	0.0402
14	2.5099	0.3665
15	2.1434	0.1108
16	2.0326	0.0013
17	2.0313	0.034
18	1.9973	0.177
19	1.8203	0.1282
20	1.6921	0.2814

The whole procedure could be re-run from step 3 with different initial settings in the *K*-Means method. Another possible result is plotted in Fig. 9. There is still an apparent difference between Fig. 8 and Fig. 9, which means that this method still converges to local optimal. However, the difference is much smaller than that between Fig. 2 and Fig. 3, in this case.

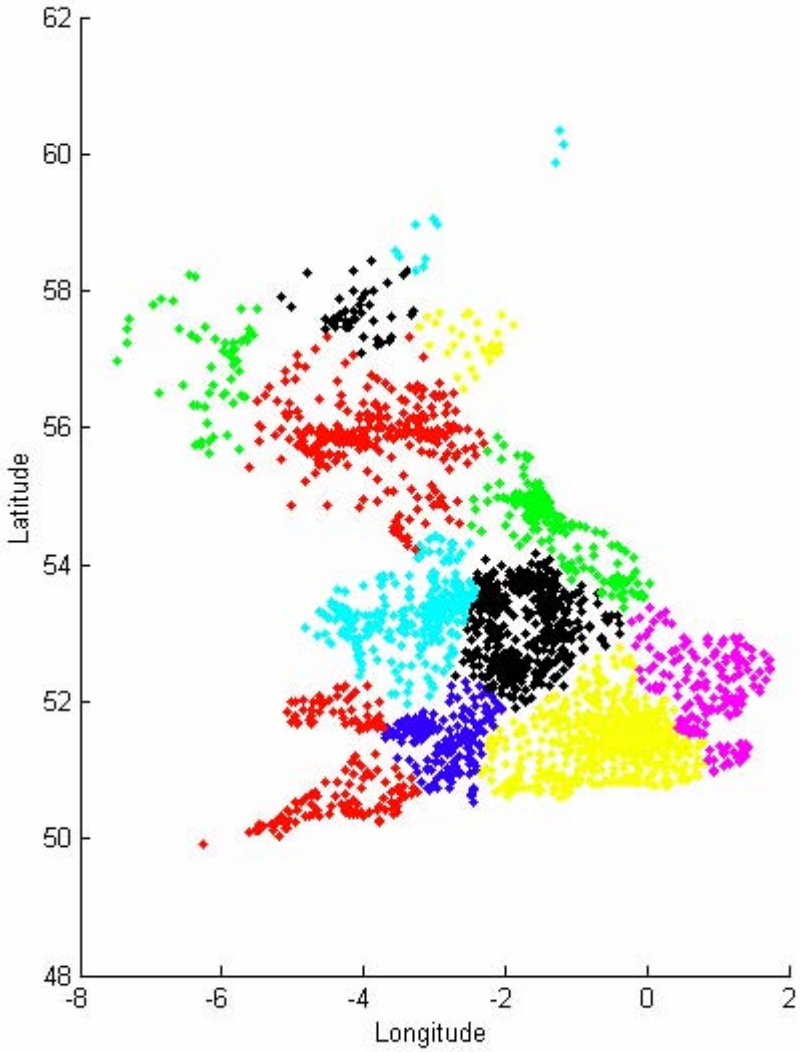

Fig.8 Clustering result by EAH method with 12 clusters

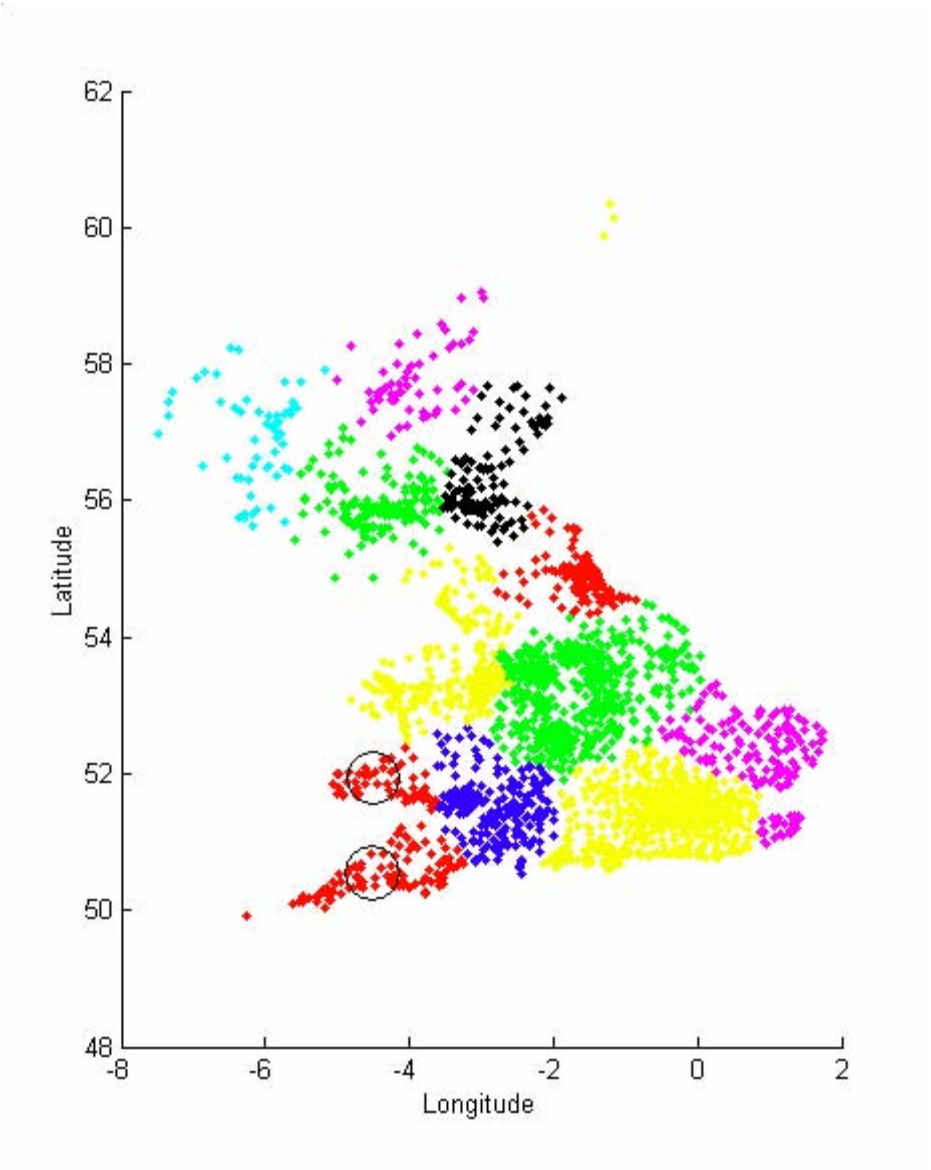

Fig. 9 Clustering result by EAH method with 12 clusters using different initial setting to Fig. 8

5. OTHER CONSIDERATIONS

In this section some other considerations in clustering are introduced and briefly explained. One common problem is the existence of obstacles and constraints in clustering. For example, in Fig. 9 the two circled areas are not adjacent because of the gulf. However, when distance is used to define the similarity measure, they are very close. One solution is to introduce a distance function that includes this information: for example, the distance between any two points across the gulf doubles the normal Euclidean distance. There are other methods in [1]-[4].

Whether to use claim frequency, severity or burning cost is also interesting. As explained in Section

4, there is strong argument for using claim frequency. However, in the case where claim severity is different between territories, it may be more reasonable to use claim severity or burning cost. In such a situation, the variance adjustment to the Euclidean distance could be different and it could also affect the magnitude of weighting parameter w.

Finally, checking the effectiveness of clustering is also difficult. As illustrated in the case study, it is very difficult to compare the results from different clustering methods or the same method with different initial settings. One solution is to repeat the clustering procedure by a large number of times and find the consistent pattern. Another is to check with external information, such as industry benchmarks. The third option is to split the data into half, using half of the data to do clustering analysis and the other half to test whether the same pattern appears.

6. CONCLUSIONS

Clustering is an important tool in data mining for insurance ratemaking. However, choosing clustering methods is a difficult task because there are a large number of clustering methods in literature and there is no conclusion as to which method is always best. The philosophy suggested in this paper is to use the simplest method possible, as long as there is no critical drawback.

Broad review of clustering methods shows that using partitioning methods without proper modifications are not suitable for insurance ratemaking. Hierarchical methods have much better performance but they are limited by highly complex calculations, so they struggle with a large dataset. Advanced methods could improve the efficiency of clustering but may be difficult to understand and apply. So from a practical point of view, this paper emphasizes modifying the partitioning and hierarchical methods to accommodate the needs of insurance ratemaking.

In the proposed exposure-adjusted hybrid (EAH) clustering method, the exposure-adjusted similarity measure is used to take account of the uncertainty of insurance data and the *K*-Means method is applied first to generate sub-clusters to reduce the time used in the hierarchical methods. Case study results show that this method could alleviate some problems of basic partitioning and hierarchical methods.

By its unsupervised nature of clustering, there is no definite choice for best clustering method; other methods introduced in this paper could give reasonable solutions in appropriate situations. However, it is hoped that various considerations mentioned in this paper could provide some practical help to users of clustering in insurance ratemaking.

Acknowledgment

The author would like to thank the reviewers for comments that greatly improved the paper, in technical aspects as well as in spelling and grammar.

7. REFERENCES

[1] Jain, A.K., M.N. Murty, and P.J. Flynn. "Data clustering: A review." *ACM Computing Surveys* 31, no. 3:264-323.
[2] Xu, R. and D. Wunsch. "Survey of clustering algorithms." *IEEE Transactions on Neural Networks* 16, no. 3:645-678.
[3] Han, J. M. Kamber, and A.K.H. Tung. "Spatial clustering methods in data mining: A survey." In: Miller, H., Han, J. (Eds.), *Geographic Data Mining and Knowledge Discovery* (London: Taylor and Francis, 2001).
[4] P. Berkhin. "Survey of clustering data mining techniques," Technical Report, Accrue Software, 2002.
[5] Filippone, M., F. Camastra , F. Masulli, S. Rovetta. 2008. "A survey of kernel and spectral methods for clustering," *Pattern Recognition* 41, no. 1:176-190.
[6] Von Luxburg, U. 2007. "A Tutorial on Spectral Clustering." *Statistics and Computing* 17, no. 4:395–416.
[7] Pelessoni, R. and L. Picec. 1998. "Some applications of unsupervised neural networks in rate making procedure." Presented at the 1998 General Insurance Convention & ASTIN Colloquium, 549-567.
[8] Sanche, R. and K. Lonergan. "Variable reduction for predictive modeling with clustering." *Casualty Actuarial Society Forum*, Winter 2006, 89-100.
[9] Guo, L. "Appling data mining techniques in property/casualty insurance." *Casualty Actuarial Society Forum*, Winter 2003, 1-25.
[10] Christopherson, S. and D. L. Werland. "Using a geographic information system to identify territory boundaries." *Casualty Actuarial Society Forum*, Winter 1996, 191-212.
[11] Anderson, D., S. Feldblum, C. Modlin, D. Schirmacher, E. Schirmacher, and N. Thandi. "A practitioner's guide to generalized linear models." *Casualty Actuarial Society Discussion Paper Program*, 2004, 1-116

Biography of the Author

Ji Yao is an actuarial analyst for Zurich Financial Services in the United Kingdom. He works in the commercial lines pricing department. Before his position at Zurich, he had two and a half years of personal lines experience. He has a BEng degree in electronic information from Shanghai's Jiao Tong University in the People's Republic of China and a Ph.D. in mathematics and statistics from the University of Birmingham, U.K. He is a Fellow of the Royal Statistical Society.