ZURICH

# Clustering in Ratemaking: Applications in Territories Clustering

**2008 CAS Predictive Modeling Seminar**

6th-7th October 2008

Ji Yao

# Clustering in Ratemaking: Applications in Territories Clustering

## INTRODUCTION

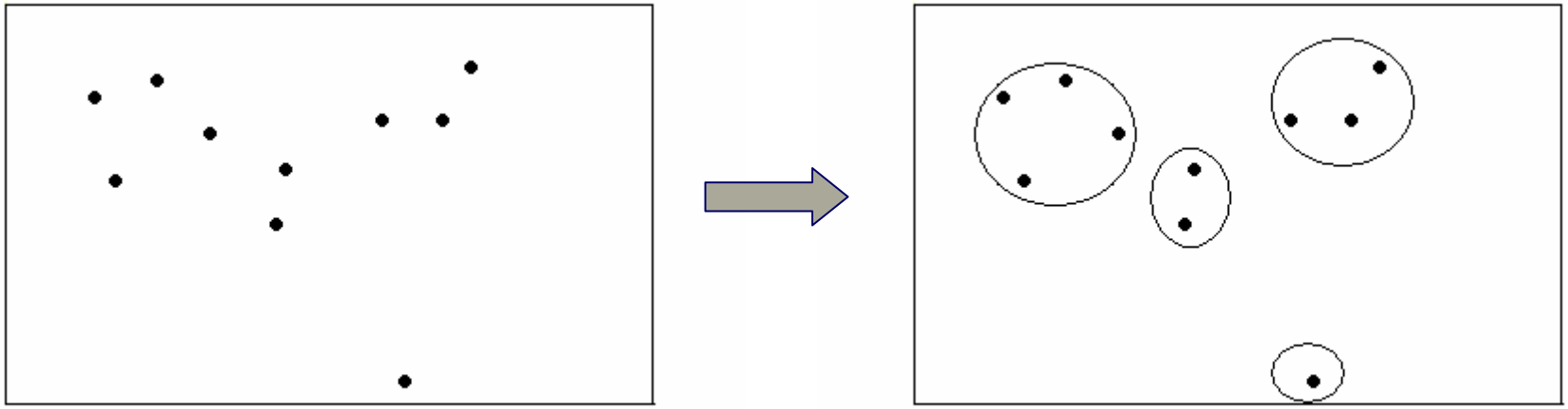
- Introduces clustering and its application in insurance ratemaking
- Reviews clustering methods and their applicability in insurance ratemaking
- Proposes the EAH clustering method and illustrates this method step-by-step using U.K. motor data
- Discusses some other considerations in clustering
- Questions

# Clustering in Ratemaking: Applications in Territories Clustering

## OVERVIEW OF CLUSTERING

- Definition of clustering
  - *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.
- Clustering vs. Discriminant analysis

ssi



# Clustering in Ratemaking: Applications in Territories Clustering

## OVERVIEW OF CLUSTERING

- Purpose of Clustering in Insurance
  - Better understand the data/trends
    - Appropriate grouping
  - Reduce the volatility of data and to make the rates stable over time
  - Reduce the number of levels in rating factors
    - Rates for vehicle
  - Make the rate are reasonable and smooth the rates
    - Rates of adjacent area

# Clustering in Ratemaking: Applications in Territories Clustering

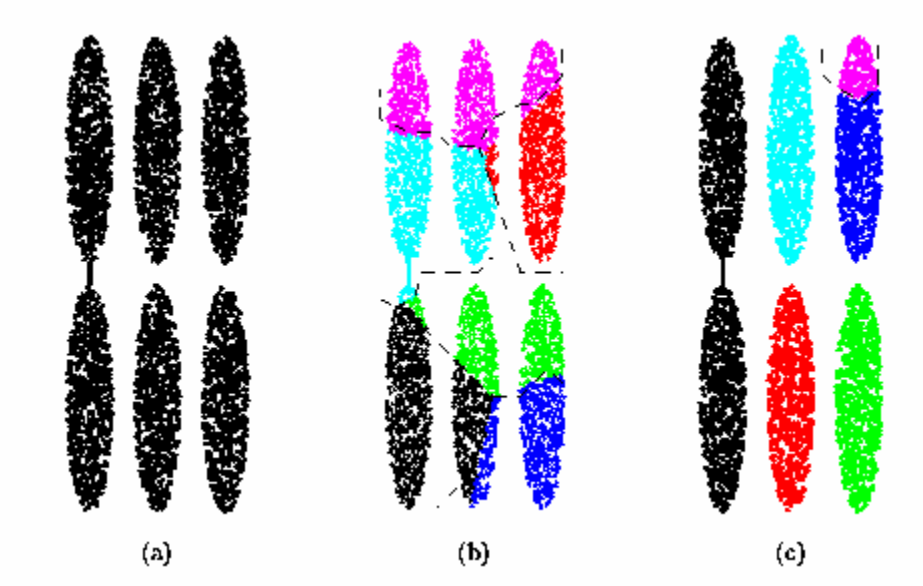
## OVERVIEW OF CLUSTERING

- Use of Clustering in Insurance
  - Geographic
  - Occupation/Trade
  - Vehicle
  - Product list

# Clustering in Ratemaking: Applications in Territories Clustering

## OVERVIEW OF CLUSTERING

- Nature of Insurance Dataset
  - critical in choosing clustering method
  - numerical vs. non-numerical
    - Geographic, occupation, vehicle
  - multi-dimensional
    - Claim experience, plus rating facto
  - large noise
  - not well-separated
    - Conventional clustering method applied to well separated data
  - the change between clusters could be gradual



# Clustering in Ratemaking: Applications in Territories Clustering

## CLUSTERING METHODS -Partitioning Methods

- Partitioning Methods
  - Broadly, this method organizes the data objects into a required number of clusters that optimizes certain similarity measure.
  - Narrowly this is implemented by an iterative algorithm where the similarity measure is based on the distance between data objects.
  - 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.

# Clustering in Ratemaking: Applications in Territories Clustering

## CLUSTERING METHODS-Partitioning Methods

- *K*-Means Method
  - The center of cluster, $m_i$, is defined as the mean of each cluster, $C_i$, that is,

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

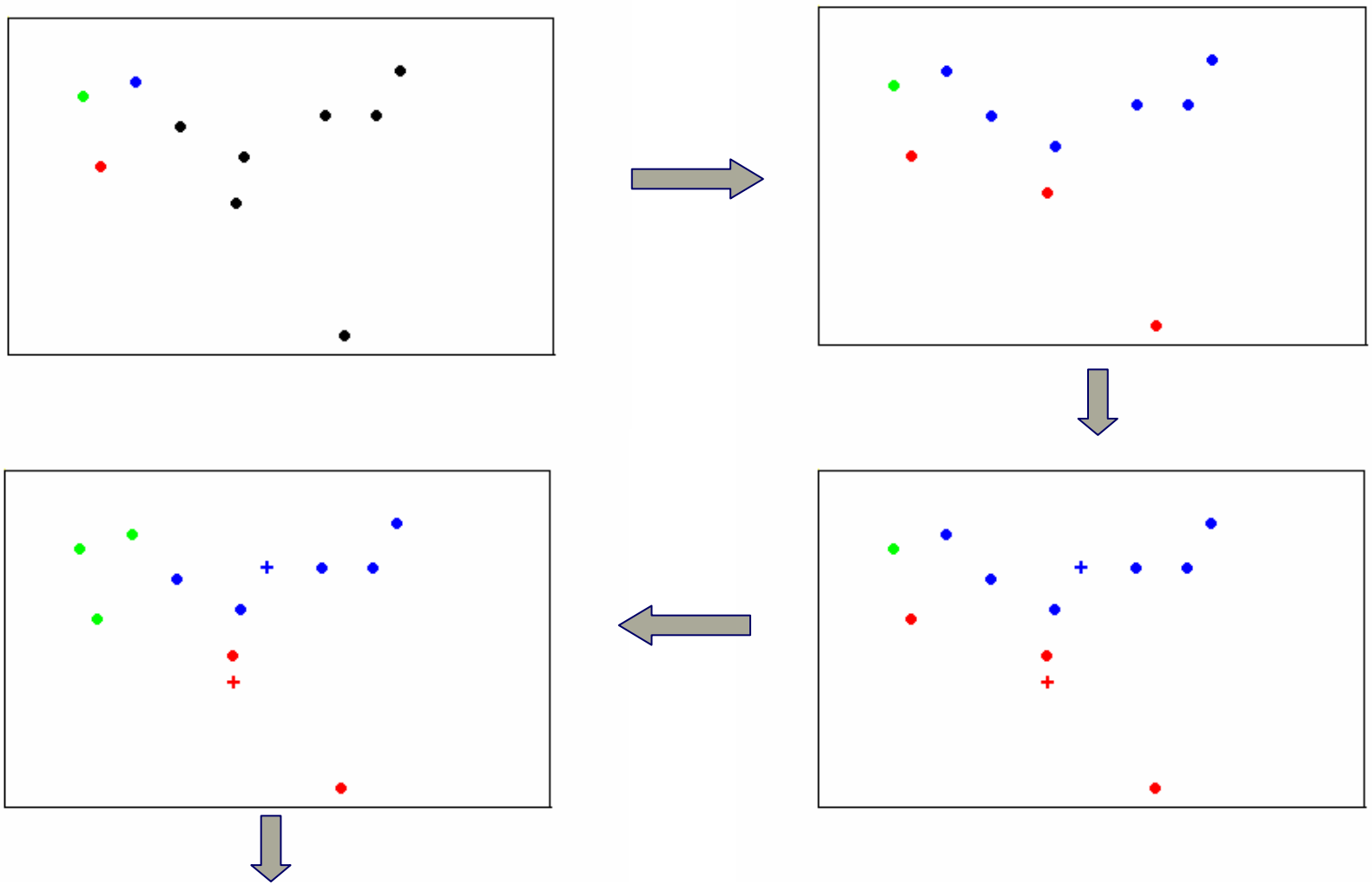
  - Similarity function is the square-error function

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

- Example
  - Looking for 3 clusters

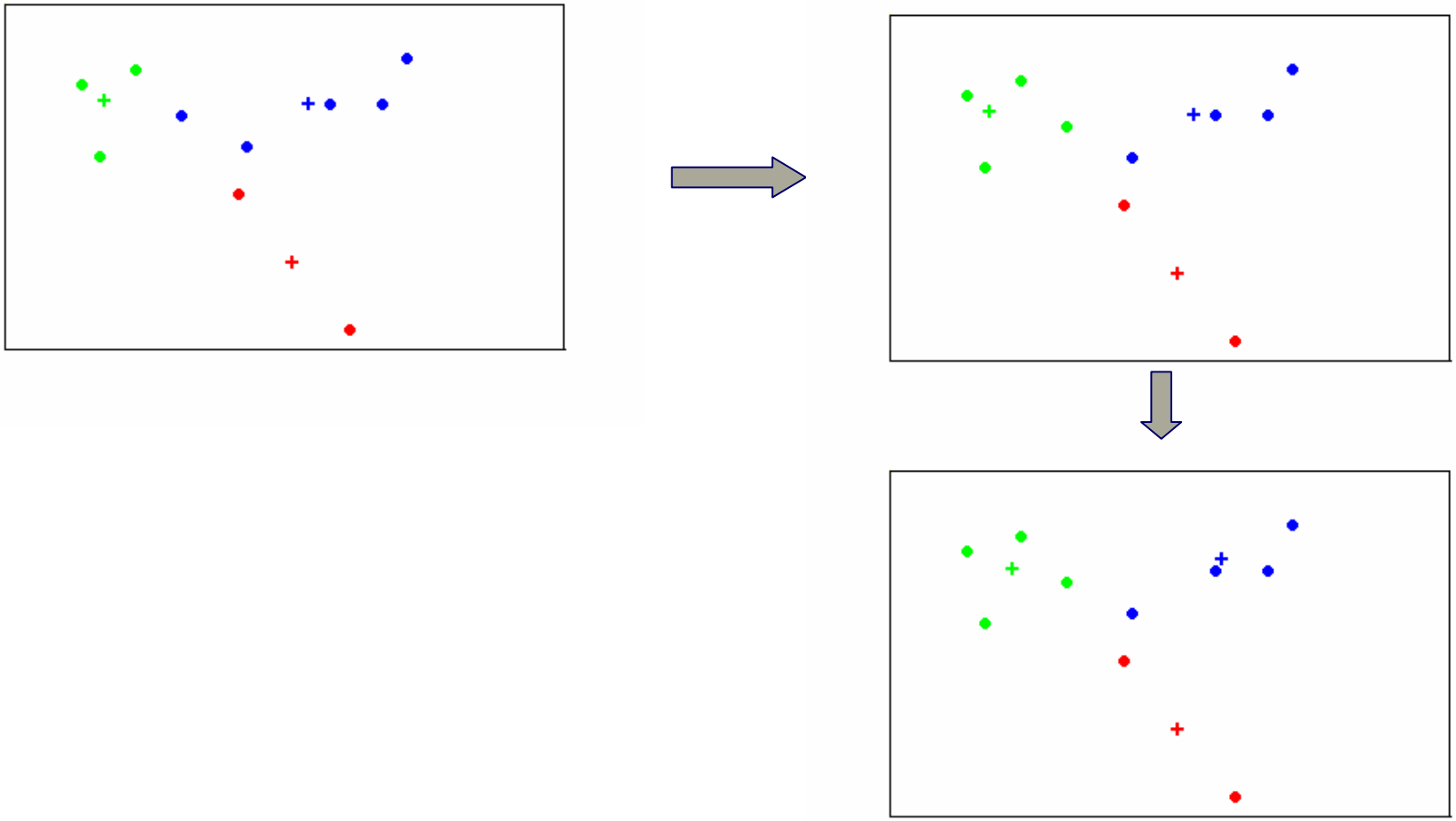
# Clustering in Ratemaking: Applications in Territories Clustering

## CLUSTERING METHODS-Partitioning Methods



# Clustering in Ratemaking: Applications in Territories Clustering

## CLUSTERING METHODS-Partitioning Methods



# Clustering in Ratemaking: Applications in Territories Clustering

## CLUSTERING METHODS-Partitioning Methods

- Advantage
  - easy to understand and apply
  - time complexity of is lower than most other methods
  - most widely used
- Disadvantage
  - sensitive to noise and outliers
  - difficult to choose the appropriate number of clusters
  - tend to be sphere-shaped
  - affected by the initial setting
  - only converge to a local optimal

# Clustering in Ratemaking: Applications in Territories Clustering

CLUSTERING METHODS-Partitioning Methods

- *K*-Medoids Method
  - defines the most centrally located data object of cluster *Ci* as the cluster center to calculate the squared-error function
- Advantage
  - less sensitive to noise and outliers
- Disadvantage
  - much higher run time to find the “most centrally” located data
  - Other similar problem with *k*-Means method

# Clustering in Ratemaking: Applications in Territories Clustering

CLUSTERING METHODS-Partitioning Methods

- Expectation Maximization (EM)
  - represents each cluster by a probability distribution
- Advantage
  - time complexity is lower than *K*-Medoids method
- Disadvantage
  - most of the problem *K*-Means suffers
  - choice of probability distribution

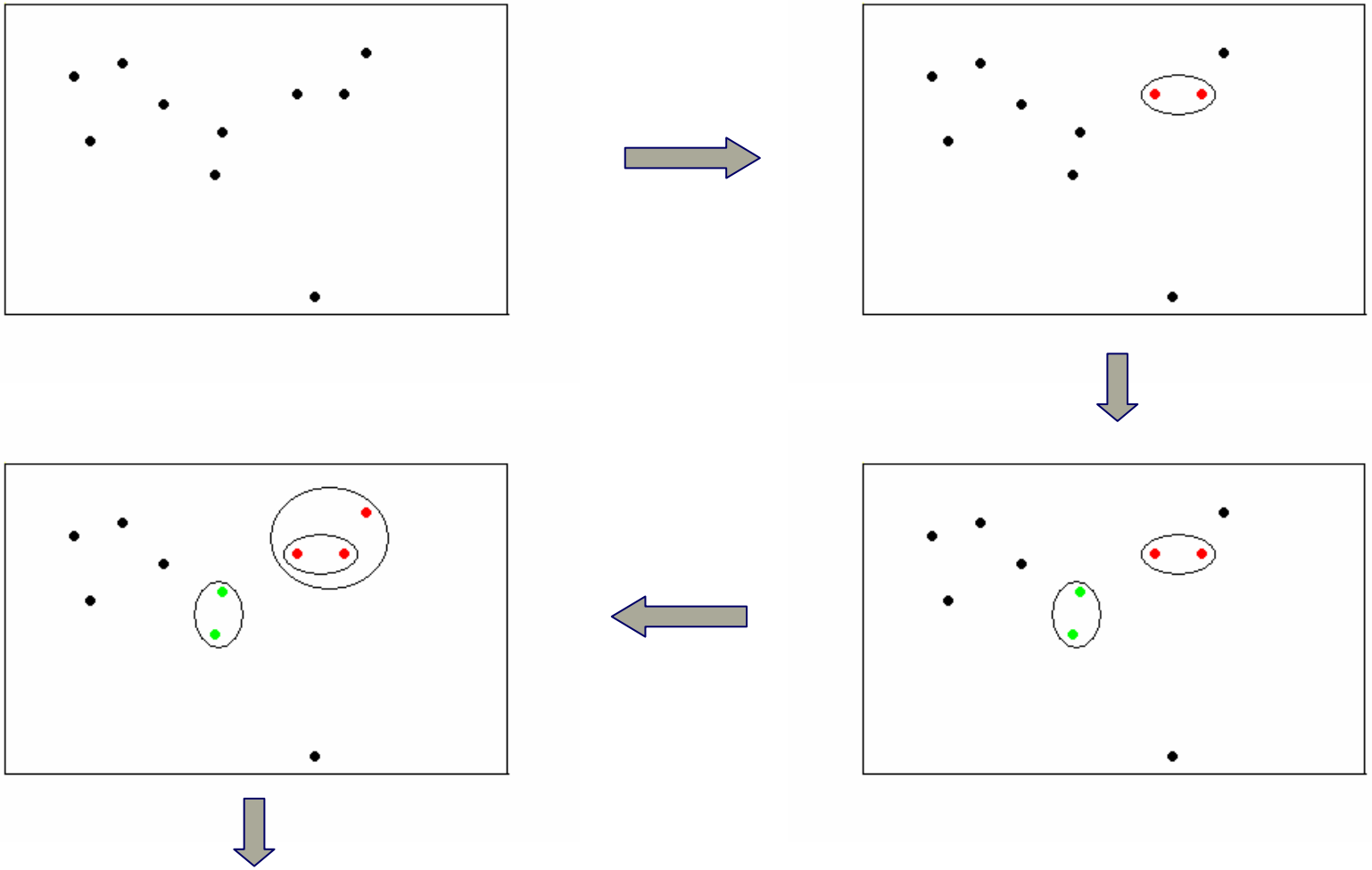
# Clustering in Ratemaking: Applications in Territories Clustering

CLUSTERING METHODS-Hierarchical Methods

- AGglomerative NESting (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.
- Need to define the cluster-to-cluster similarity measure. Common ones are
  1. Min distance
  2. Max distance
  3. Average distance
- Example

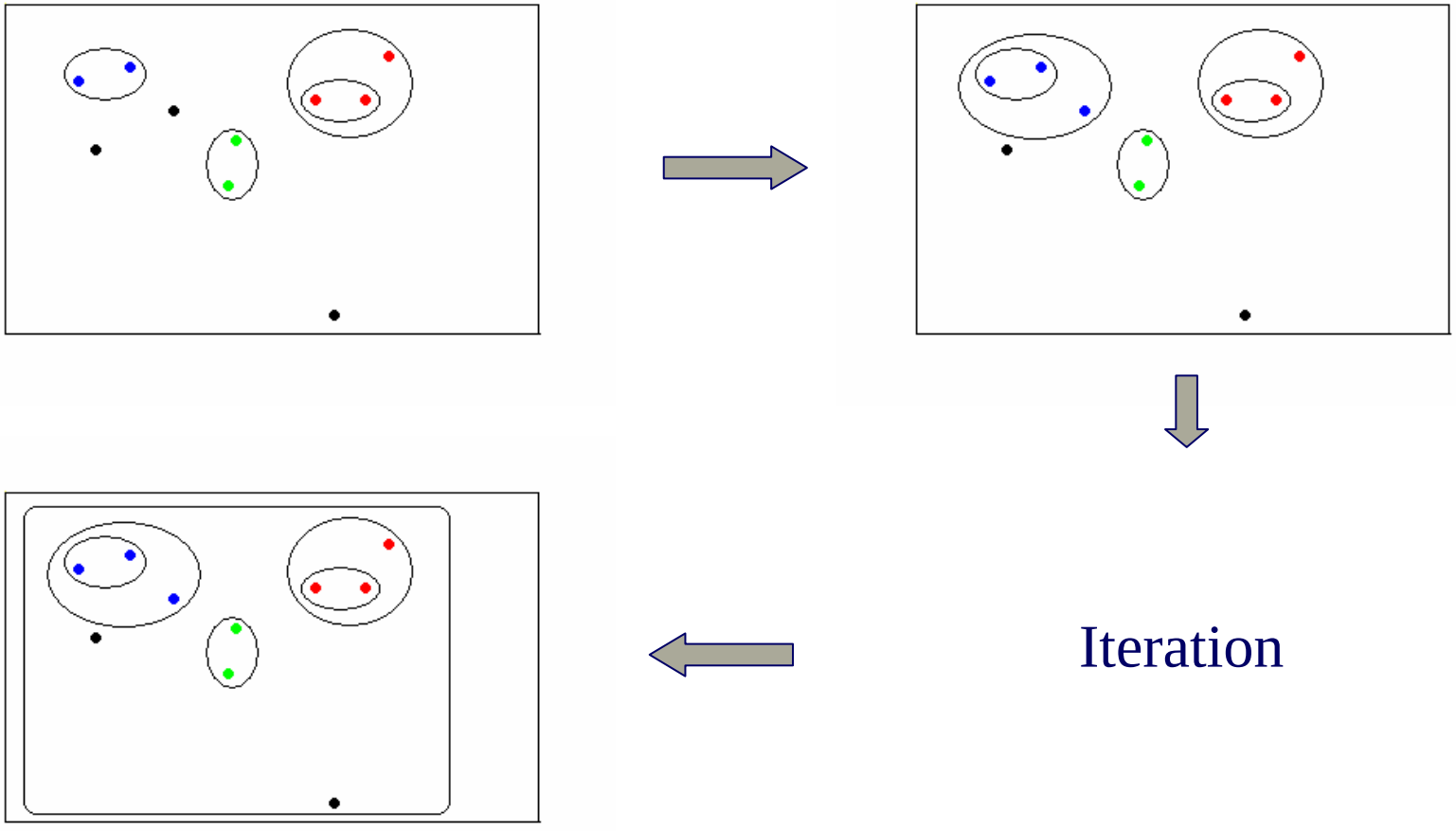
# Clustering in Ratemaking: Applications in Territories Clustering

## CLUSTERING METHODS-Hierarchical Methods



# Clustering in Ratemaking: Applications in Territories Clustering

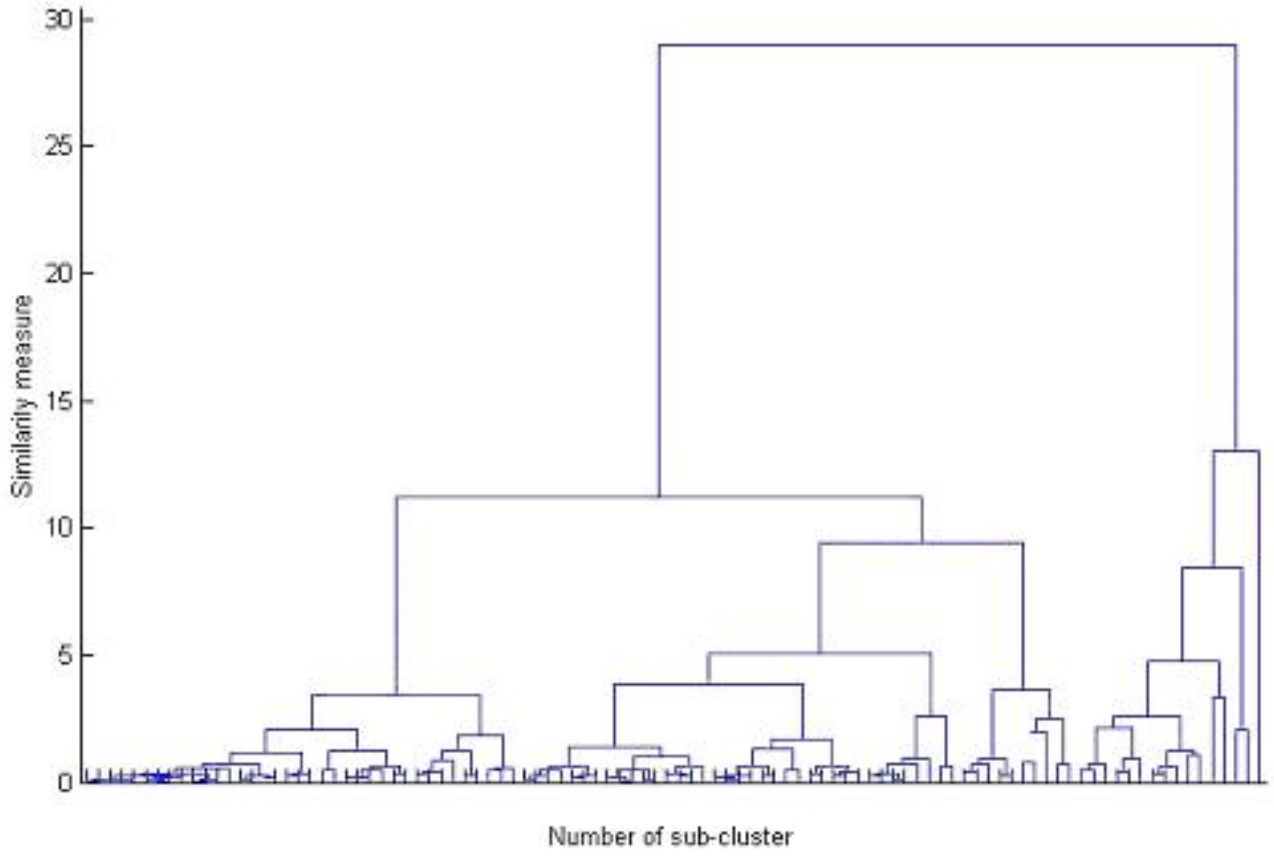
## CLUSTERING METHODS-Hierarchical Methods



# Clustering in Ratemaking: Applications in Territories Clustering

CLUSTERING METHODS-Hierarchical Methods

- The result is a dendrogram, looks like this



# Clustering in Ratemaking: Applications in Territories Clustering

CLUSTERING METHODS-Hierarchical Methods

- DIvisia ANAlysis (DIANA)
  - reverse to AGNES
  - 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.
  - result is also presented in dendrogram

# Clustering in Ratemaking: Applications in Territories Clustering

CLUSTERING METHODS-Hierarchical Methods

- Advantage
  - easy to understand and apply
  - are less sphere-shaped than partitioning methods
  - number of clusters is also chosen at a later stage
- Disadvantage
  - the over-simplified similarity measure often gives erroneous clustering results
  - irreversible
  - high complexity of time

# Clustering in Ratemaking: Applications in Territories Clustering

CLUSTERING METHODS-Hierarchical Methods

- Balanced Iterative Reducing and Clustering using Hierarchies (BIRTH)
  - compress the data objects into small sub-clusters in first stage and then perform clustering with these sub-clusters in the second stage
- advantage
  - greatly reduces the effective number of data objects that need to cluster
  - reduces the time complexity.
- Disadvantage
  - spherical shape clustering

# Clustering in Ratemaking: Applications in Territories Clustering

CLUSTERING METHODS-Hierarchical Methods

- Clustering Using REpresentatives (CURE)
  - 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.
- Advantage
  - robust to outliers and has a better performance when clusters have non-spherical shape
- Disadvantage
  - all parameters, such as number of representative data points of a cluster and shrinking speed, have a significant impact on the results

# Clustering in Ratemaking: Applications in Territories Clustering

CLUSTERING METHODS-Hierarchical Methods

- CHAMELEON method
  - more sophisticated measures of similarity such as *inter-connectivity* and *closeness* are used
  - uses a special graph partitioning algorithm to recursively partition the whole data objects into many small unconnected sub-clusters .
- Advantage
  - more efficient than CURE in discovering arbitrarily shaped clusters of varying density
- Disadvantage
  - the time complexity is quite high

# Clustering in Ratemaking: Applications in Territories Clustering

CLUSTERING METHODS-Density-Based Methods

- Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
  - Basic idea:

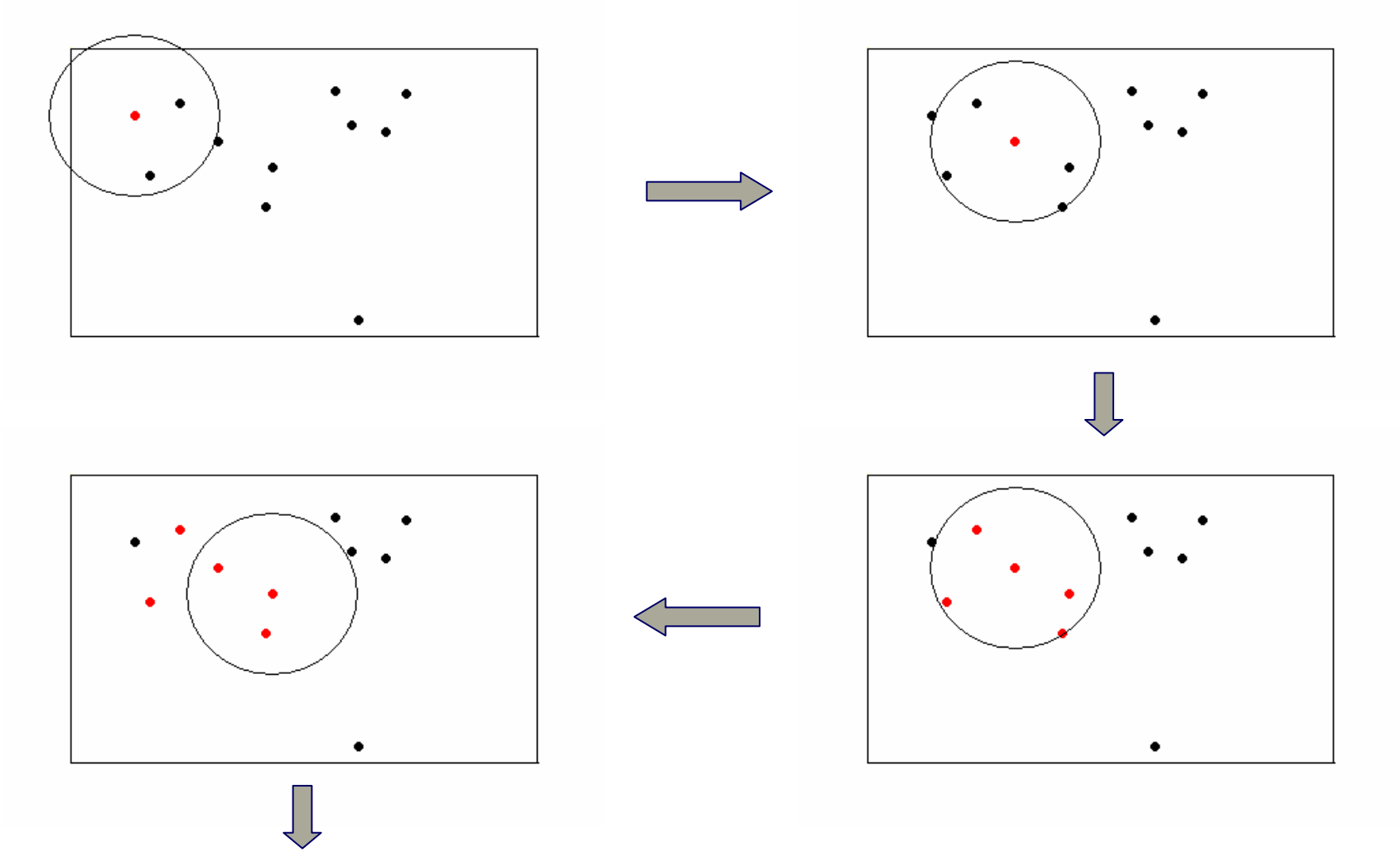
    1. Defines the density of a data object as the number of data objects within a certain distance of the data object.
    2. If the density of a data object is larger then a threshold, this object is termed “core”.
    3. Expand every cluster as long as the neighboring data object is a “core” object.
    4. Outliers are discarded and not grouped to any clusters

- Example

  Threshold=3

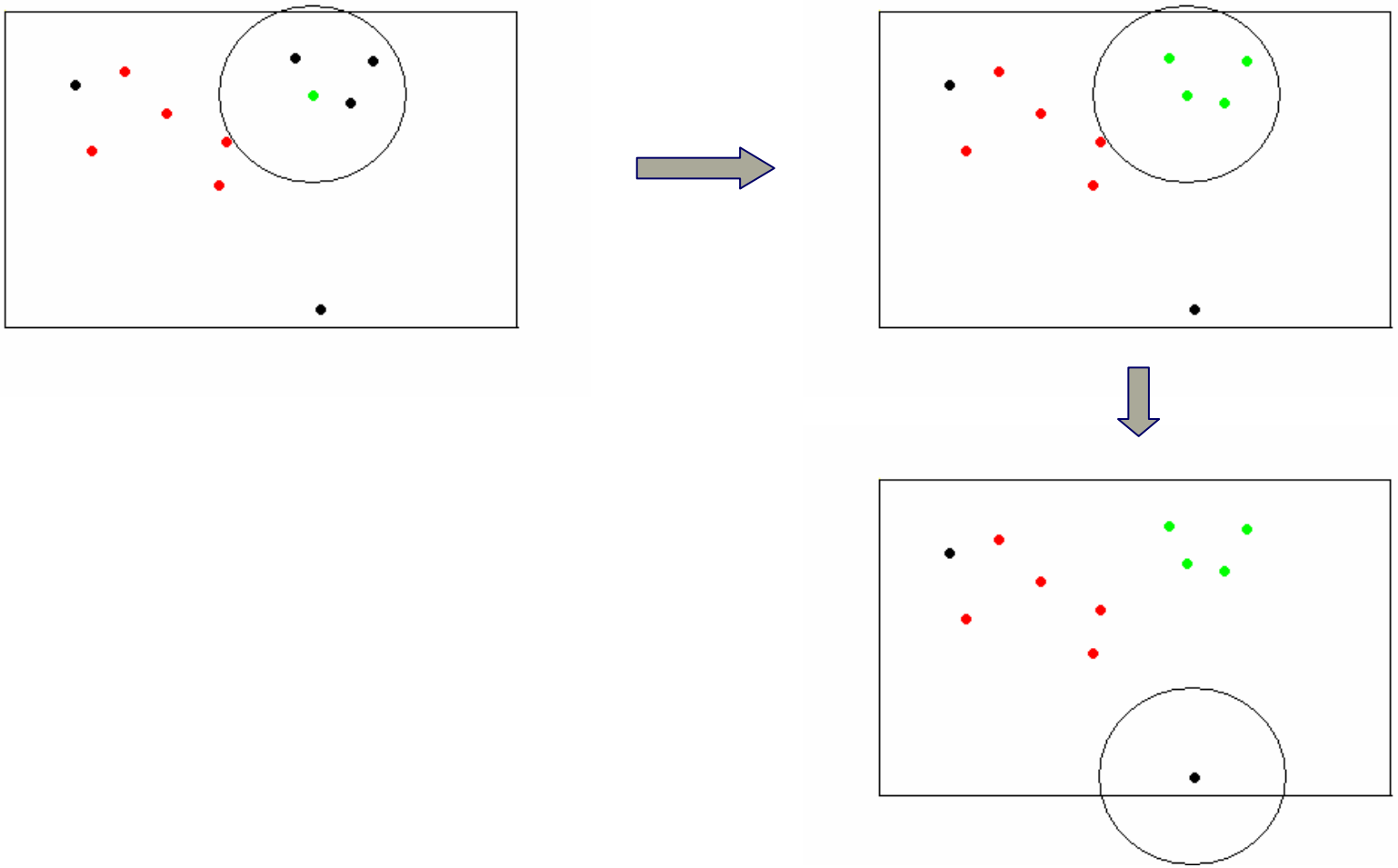
# Clustering in Ratemaking: Applications in Territories Clustering

## CLUSTERING METHODS-Density-Based Methods



# Clustering in Ratemaking: Applications in Territories Clustering

## CLUSTERING METHODS-Density-Based Methods



# Clustering in Ratemaking: Applications in Territories Clustering

CLUSTERING METHODS-Density-Based Methods

- Advantage
  - could find arbitrary shape of clusters
- Disadvantage
  - efficiency of this method largely depends on parameters chosen by the user
  - not work very well for a large or high-dimensional dataset

# Clustering in Ratemaking: Applications in Territories Clustering

CLUSTERING METHODS- Density-Based Methods

- Ordering Points To Identify the Clustering Structure (OPTICS)
  - This method produces a cluster ordering for a wide range of parameter settings
  - Key Idea
    - For each data object, find distance to the nearest “core” object, i.e. find the minimum distance that this data object could be clustered rather than discarded as noise.
    - Ordering the data object from the minimum distance
  - Advantage
    - Solves the problem of dependency on parameters as in DBSCAN

# Clustering in Ratemaking: Applications in Territories Clustering

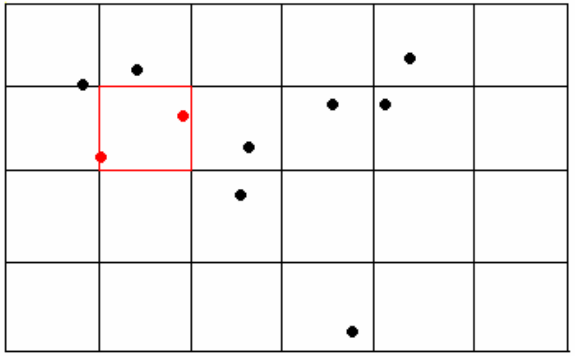
CLUSTERING METHODS- Density-Based Methods

- DENsity-based CLUstEring (DENCLUE)
  - This method is efficient for large datasets and high-dimensional noisy datasets;
  - Many parameters to set and it may be difficult for the non-expert to apply;

# Clustering in Ratemaking: Applications in Territories Clustering

## CLUSTERING METHODS- 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.
- Some features of cells are then used for clustering
- Combined with other methods
- Example



# Clustering in Ratemaking: Applications in Territories Clustering

CLUSTERING METHODS- Grid-Based Methods

- Advantage
  - fast processing time
- Disadvantage
  - shape of the cluster is limited by the shape of grid ->smaller grid
- Advanced methods
  - STING: explores statistical information
  - WaveCluster: uses wavelet transform to store the information
  - CLIQUE: discovers sub-clusters using the a priori principle

# Clustering in Ratemaking: Applications in Territories Clustering

## CLUSTERING METHODS- Kernel and Spectral Methods

- Kernel and Spectral Methods
  - relatively new methods
  - not easy for the non-expert to use and understand
  - give no more advantages than other methods in actuarial application

# Clustering in Ratemaking: Applications in Territories Clustering

## EXPOSURE-ADJUSTED HYBRID (EAH) CLUSTERING METHOD

- Other than the disadvantage mentioned for clustering method, what problem is expected in territory clustering?
  - What to cluster?
    - Claim frequency, severity, burning cost
    - What number to use? ->GLM
  - Volatility in data;
    - Adjusted to exposure
  - How to combine geographic and claim experience?
    - Weighted distance measure

# Clustering in Ratemaking: Applications in Territories Clustering

## EXPOSURE-ADJUSTED HYBRID (EAH) CLUSTERING METHOD

- The whole procedure is:
  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.

# Clustering in Ratemaking: Applications in Territories Clustering

## EXPOSURE-ADJUSTED HYBRID (EAH) CLUSTERING METHOD

- Exposure adjusted distance measure

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

  based on Normal distributed assumption

- Geographic information
  - Euclidean distance

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

  - Haversine formula to take account of curve of earth surface

# Clustering in Ratemaking: Applications in Territories Clustering

## EXPOSURE-ADJUSTED HYBRID (EAH) CLUSTERING METHOD

- Weighted distance measure

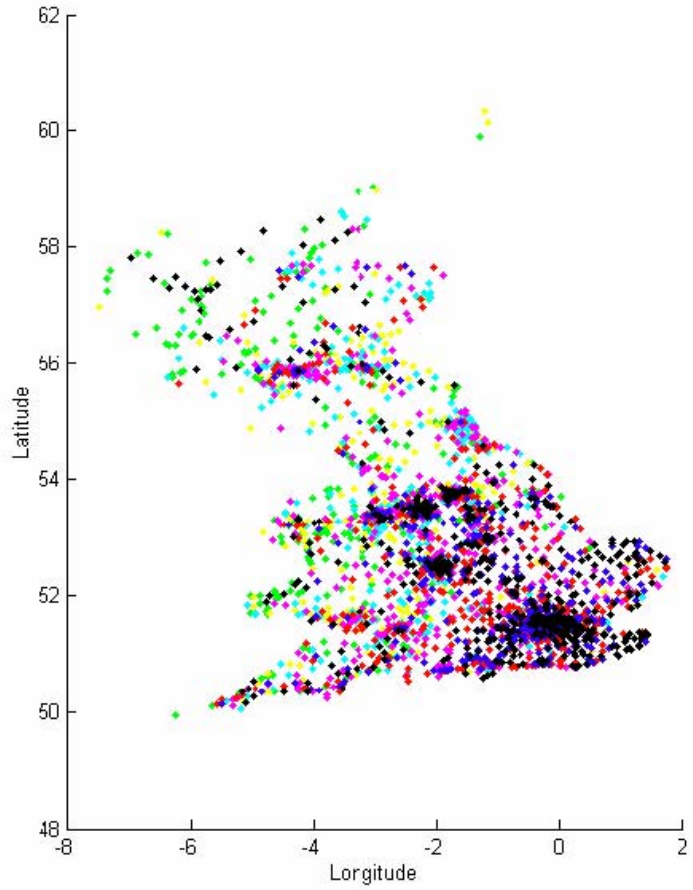
$$g(\cdot) + w \cdot f(\cdot)$$

- Higher weight means more emphasis on claim history

# Clustering in Ratemaking: Applications in Territories Clustering

## EXPOSURE-ADJUSTED HYBRID (EAH) CLUSTERING METHOD

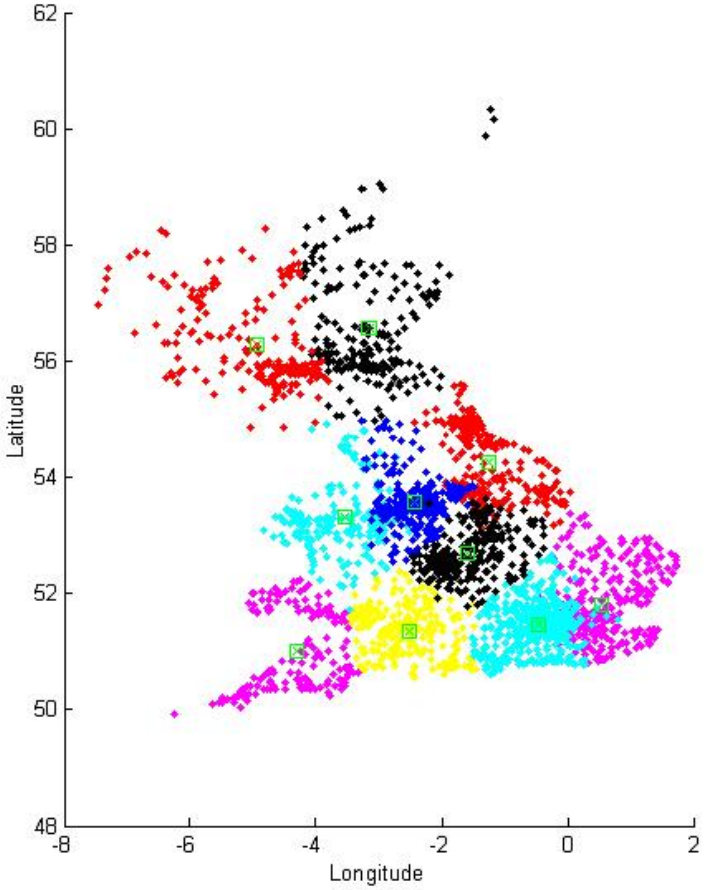
- Case study
  - Use modified UK motor data for illustration purpose only
  - The left graph show the adjusted claim experience by GLM

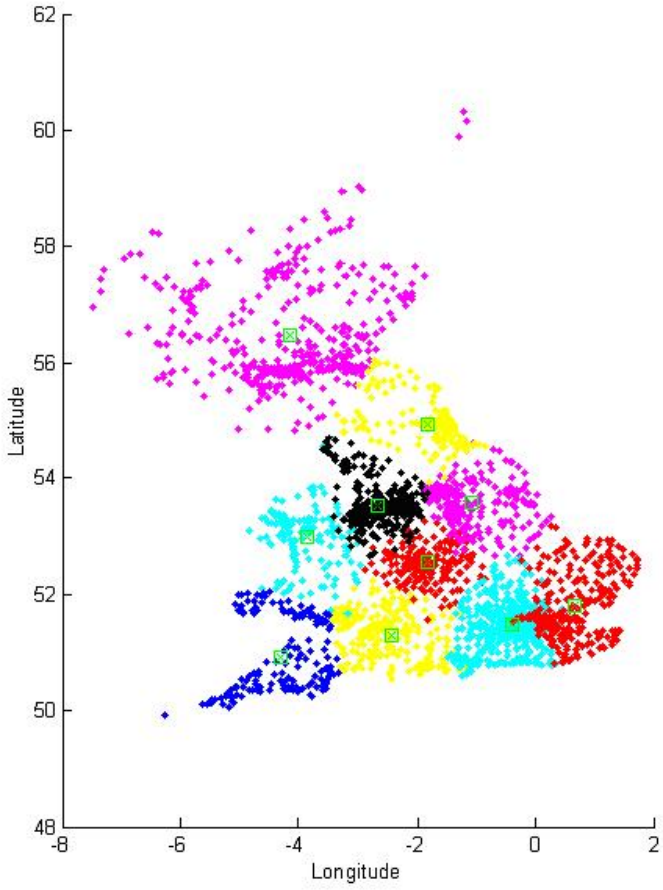


# Clustering in Ratemaking: Applications in Territories Clustering

## EXPOSURE-ADJUSTED HYBRID (EAH) CLUSTERING METHOD

- Results of K-Means method: different initial setting

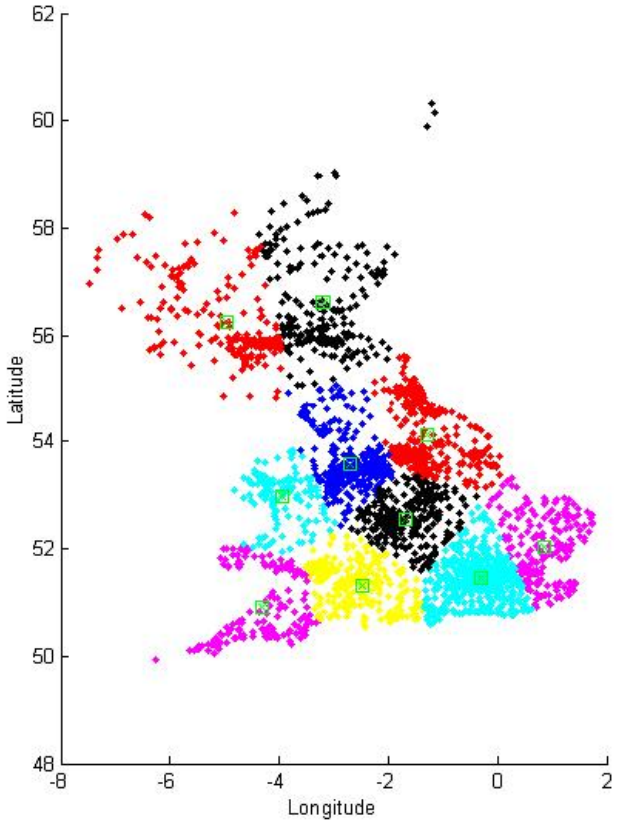


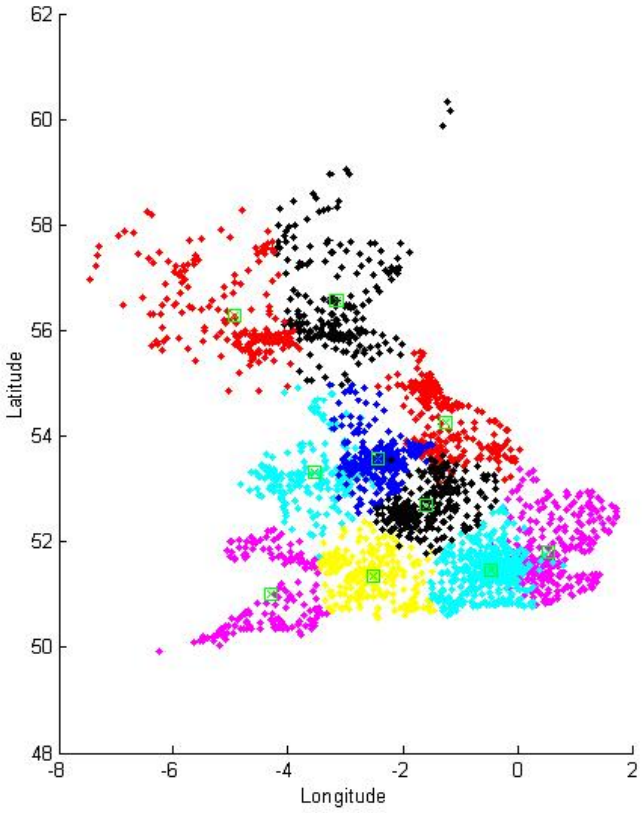
# Clustering in Ratemaking: Applications in Territories Clustering

## EXPOSURE-ADJUSTED HYBRID (EAH) CLUSTERING METHOD

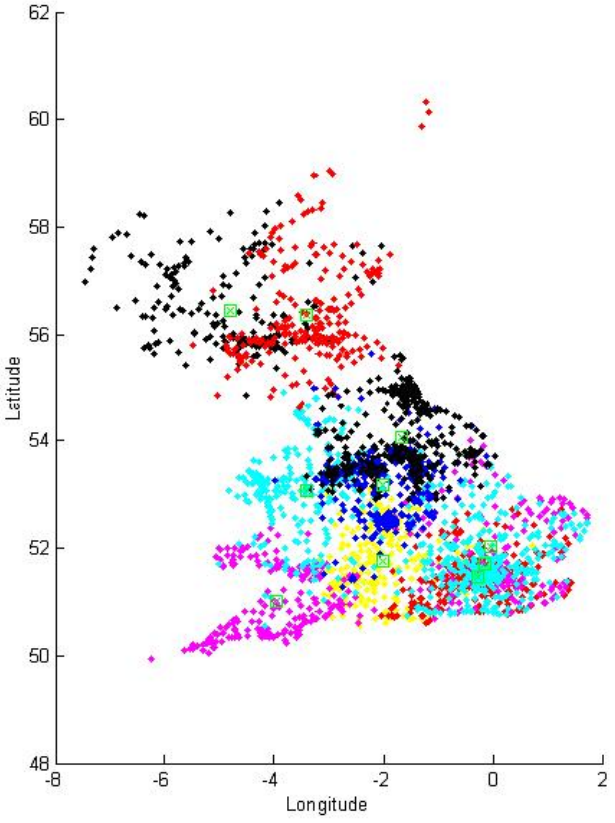
- Results of K-Means method: different weight




W=0.1




W=1

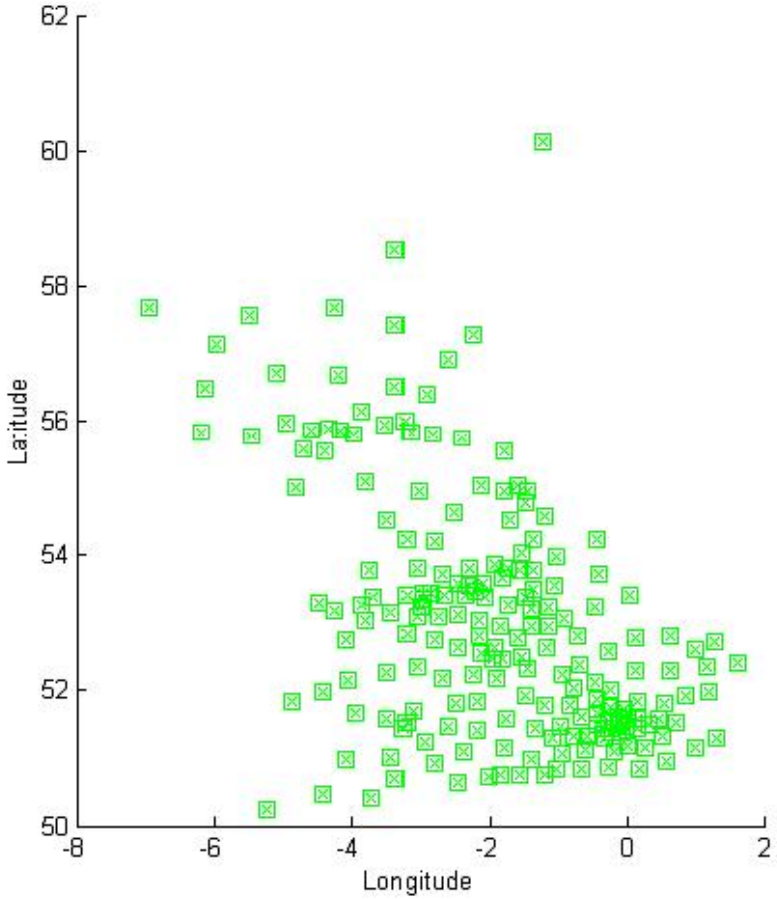


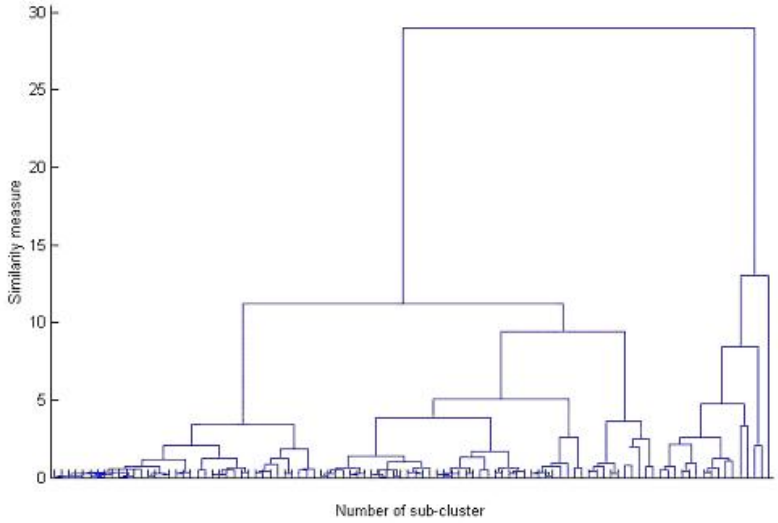

W=10

# Clustering in Ratemaking: Applications in Territories Clustering

## EXPOSURE-ADJUSTED HYBRID (EAH) CLUSTERING METHOD

- Results of EAH method

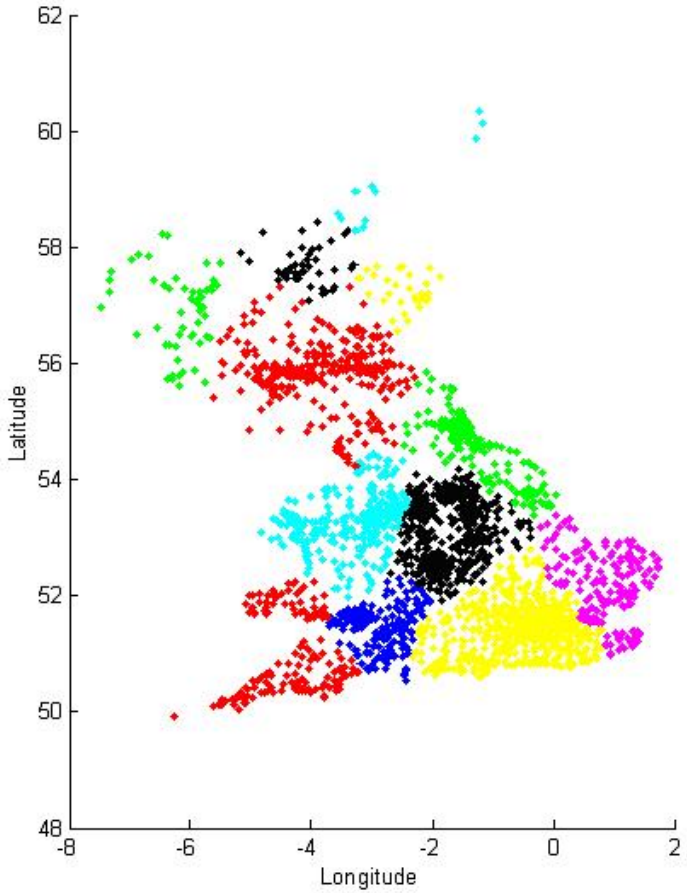


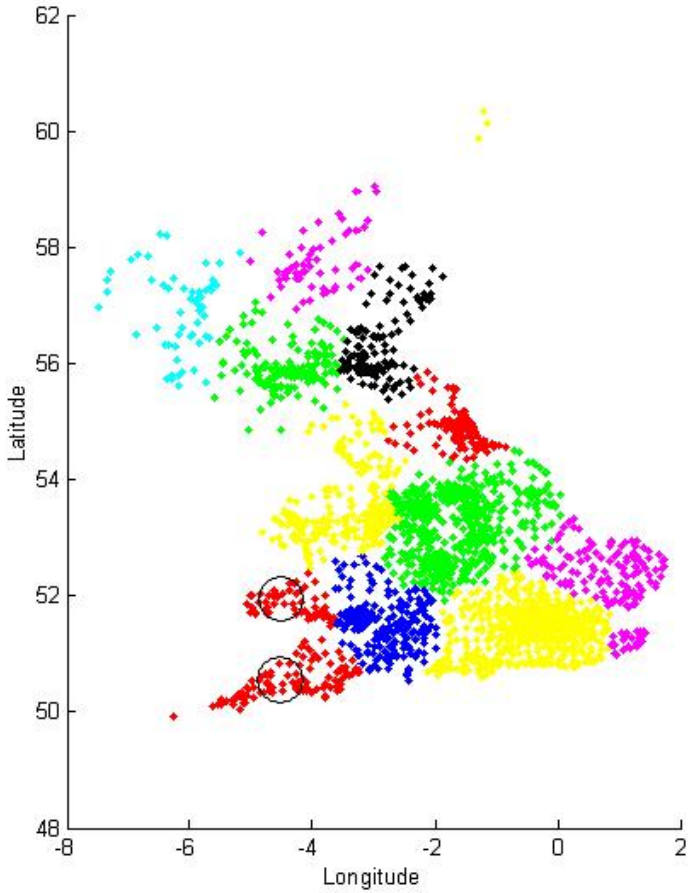



# Clustering in Ratemaking: Applications in Territories Clustering

## EXPOSURE-ADJUSTED HYBRID (EAH) CLUSTERING METHOD

- Results of EAH method-different initial setting



# Clustering in Ratemaking: Applications in Territories Clustering

MORE CONSIDERATION

- Existence of obstacles and constraints in clustering
- Change distance measure if severity or burning cost are used
- Validation of clustering results

# Questions?

# Thank You