Data Clustering with Actuarial Applications

Guojun Gan Emiliano Valdez

Department of Mathematics
University of Connecticut
Storrs, CT, USA

2017 Advances in Predictive Analytics (APA) conference,
Waterloo, Canada
December 1, 2017

Outline

- Data clustering
- An application

Data clustering

- The process of dividing a set of objects into homogeneous groups
- Originated in anthropology and psychology in the 1930s (Driver and Kroeber, 1932; Zubin, 1938; Tryon, 1939), data clustering is now one of the most popular tools for exploratory data analysis

Data clustering is a major task of data mining

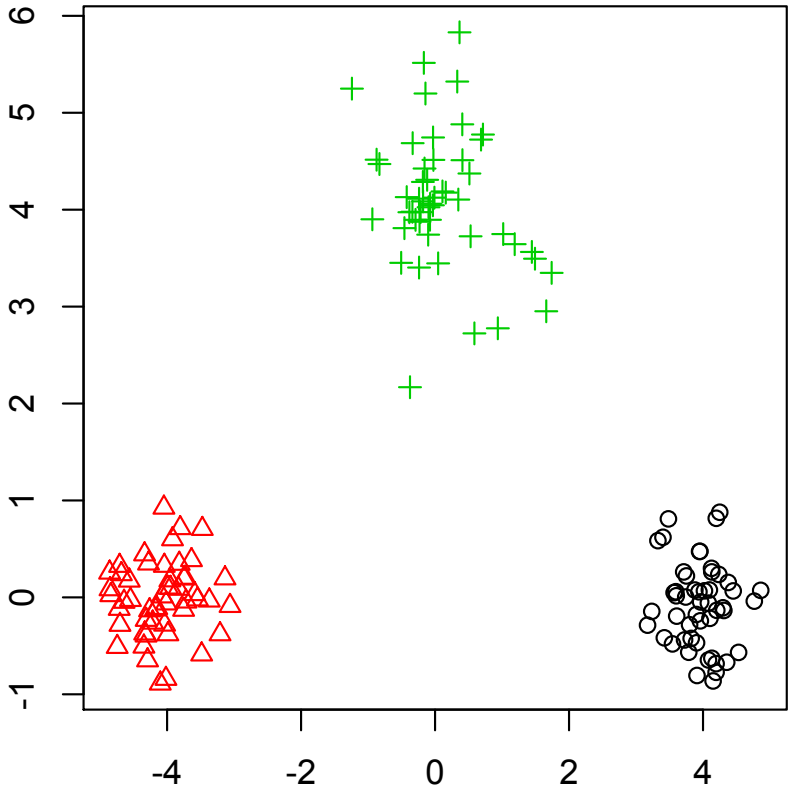
Unsupervised learning	**Supervised learning**
Data clustering	Classification
Association rules	Numerical prediction

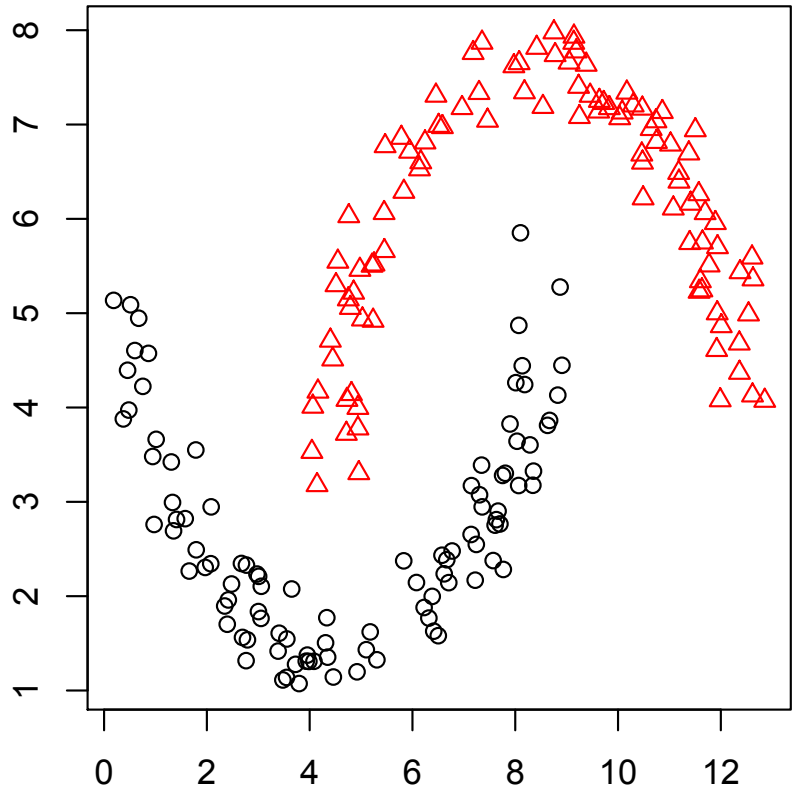
Definition of clusters

Bock (1989) also suggested the following criteria for data points in a cluster:

1. Share the same or closely related properties;
2. Have small mutual distances;
3. Have “contacts” or “relations” with at least on other data point in the cluster;
4. Can be clearly distinguishable from the data points that are not in the cluster.

Examples of clusters





Data types

	V_1	V_2	$\cdots$	V_d
$\mathbf{x}_1$	x_{11}	x_{12}	$\cdots$	x_{1d}
$\mathbf{x}_2$	x_{21}	x_{22}	$\cdots$	x_{2d}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
$\mathbf{x}_n$	x_{n1}	x_{n2}	$\cdots$	x_{nd}

Table: A dataset in a tabular form.

Types of variables:

- Discrete
- Continuous

Dissimilarity measures

A distance measure D is a binary function that satisfied the following conditions (Anderberg, 1973):

1. $D(\mathbf{x}, \mathbf{x}) \geq 0$ (Nonnegativity);
2. $D(\mathbf{x}, \mathbf{y}) = D(\mathbf{y}, \mathbf{x})$ (Symmetry);
3. $D(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$ (Reflexivity);
4. $D(\mathbf{x}, \mathbf{z}) \leq D(\mathbf{x}, \mathbf{y}) + D(\mathbf{y}, \mathbf{z})$ (Triangle inequality),

where $\mathbf{x}, \mathbf{y}$, and $\mathbf{z}$ are arbitrary data points.

$$D_{min}(\mathbf{x}, \mathbf{y}) = \left(\sum_{j=1}^{d} |x_j - y_j|^p \right)^{\frac{1}{p}}, \tag{1}$$

Clustering algorithms

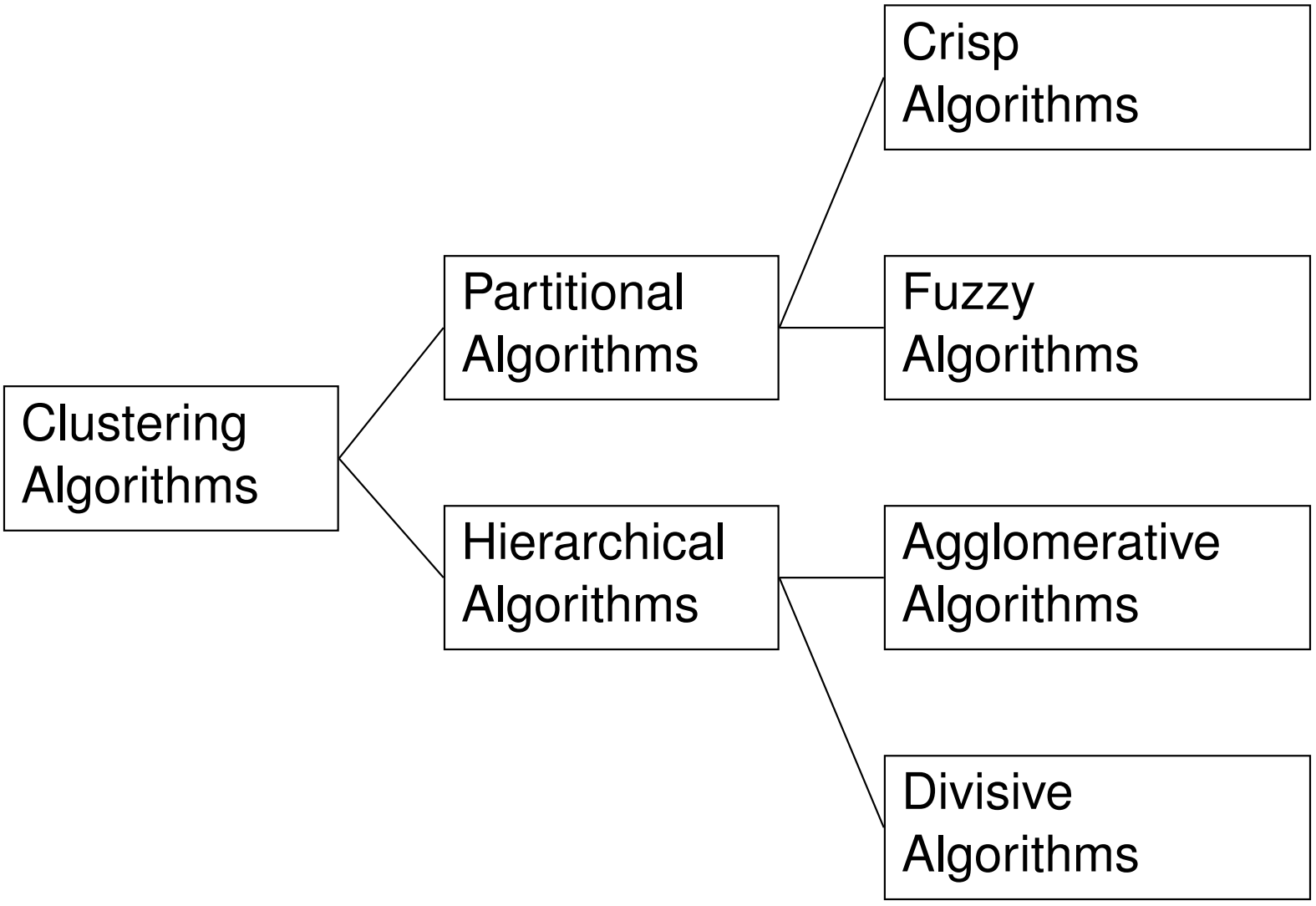


Figure: Taxonomy of clustering algorithms.

Cluster validity

- Internal validity indices evaluate the clustering results based only on quantities and features inherited from the underlying dataset.
- External validity indices evaluate the clustering results based on a prespecified structure imposed on the underlying dataset.
- Relative validity indices evaluate the results of a clustering algorithm against the results of a different clustering algorithm or the results of the same algorithm but with different parameters.

k-means

Given a set of n data points $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, the k-means algorithm tries to divide the dataset into k clusters by minimizing the following objective function:

$$P(U, Z) = \sum_{l=1}^{k} \sum_{i=1}^{n} u_{il} \|\mathbf{x}_i - \mathbf{z}_l\|^2, \tag{2}$$

where k is the desired number of cluster specified by the user, $U = (u_{il})_{n \times k}$ is an $n \times k$ partition matrix, $Z = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k\}$ is a set of cluster centers, and $\|\cdot\|$ is the L^2 norm or Euclidean distance. The partition matrix U satisfies the following conditions:

$$u_{il} \in \{0, 1\}, \quad i = 1, 2, \ldots, n, \; l = 1, 2, \ldots, k, \tag{3a}$$

$$\sum_{l=1}^{k} u_{il} = 1, \quad i = 1, 2, \ldots, n. \tag{3b}$$

Truncated fuzzy c-means

Let T be an integer such that $1 \leq T \leq k$ and let $\mathcal{U}_T$ be the set of fuzzy partition matrices U such that each row of U has at most T nonzero entries.
The TFCM algorithm (Gan et al., 2016) aims to find a truncated fuzzy partition matrix U and a set of cluster centers Z that minimize the following objective function:

$$P(U, Z) = \sum_{i=1}^{n} \sum_{l=1}^{k} u_{il}^{\alpha} \left(\|\mathbf{x}_i - \mathbf{z}_l\|^2 + \epsilon \right), \tag{4}$$

where $\alpha > 1$ is the fuzzifier, $U \in \mathcal{U}_T$, $Z = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k\}$ is a set of cluster centers, $\|\cdot\|$ is the L^2-norm or Euclidean distance, and ϵ is a small positive number used to prevent division by zero.

Hierarchical k-means

Hierarchical k-means (Nister and Stewenius, 2006) uses a divisive approach to apply the traditional k-means with small k's repeatedly until the desired number of clusters is reached.

Algorithm 1: Pseudo-code of hierarchical k-means.

Input: A dataset X, k
Output: k clusters

1 Apply the k-means algorithm to divide the dataset into two clusters;
2 **repeat**
3 | Apply the k-means algorithm to divide the largest existing cluster into two clusters;
4 **until** *The number of clusters is equal to* k;
5 Return the k clusters;

An application in variable annuity valuation

The metamodeling approach consists of the following major steps:

1. selecting a small number of representative contracts
2. using Monte Carlo simulation to calculate the fair market values (or other quantities of interest) of the representative contracts
3. building a regression model (i.e., the metamodel) based on the representative contracts and their fair market values
4. using the regression model to value the whole portfolio of variable annuity contracts

A synthetic portfolio with 190,000 variable annuiyt policies

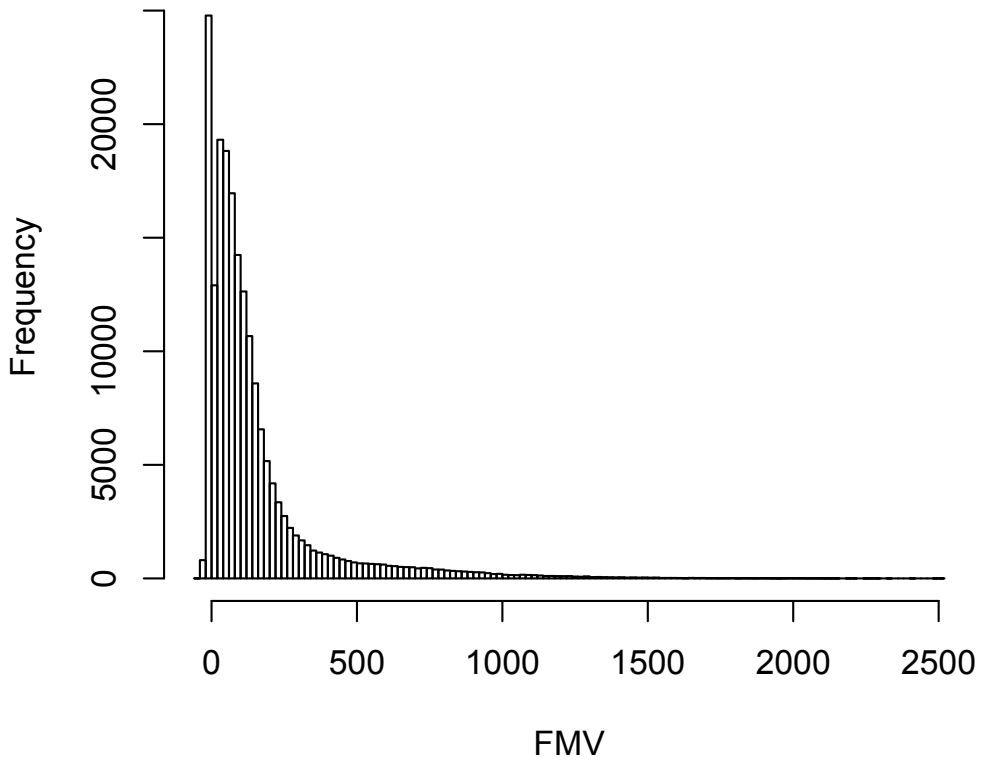


Figure: A histogram of the fair market values. The fair market values are in 1000s.

Clustering results

Table: Performance of the TFCM algorithm and the hierarchical k-means on the VA data. The runtime is in seconds.

	Hkmean (340)	Hkmean (680)	TFCM (340)	TFCM (680)
RWCSS	0.90	0.76	0.82	0.66
Runtime	130.02	136.19	2,647.11	5,544.81

The RWCSS measure is defined as follows:

$$RWCSS = \frac{\sum_{l=1}^{k} \sum_{\mathbf{x} \in C_l} \sum_{j=1}^{d} \left(x_j - z_{lj}\right)^2}{\sum_{\mathbf{x} \in X} \sum_{j=1}^{d} \left(x_j - \bar{x}_j\right)^2}, \tag{5}$$

where C_l denotes the lth cluster, $\mathbf{z}_l$ is the center of the lth cluster, $\bar{\mathbf{x}}$ is the center of the whole dataset X.

Predictive modeling results I

Table: Accuracy and runtime (in seconds) of the ordinary kriging model based on different clustering results.

	Hkmean (340)	Hkmean (680)	TFCM (340)	TFCM (680)
PE	-0.02	0.02	0.01	0.02
R^2	0.82	0.92	0.81	0.92
Runtime	329.50	787.11	334.62	808.99

Predictive modeling results II

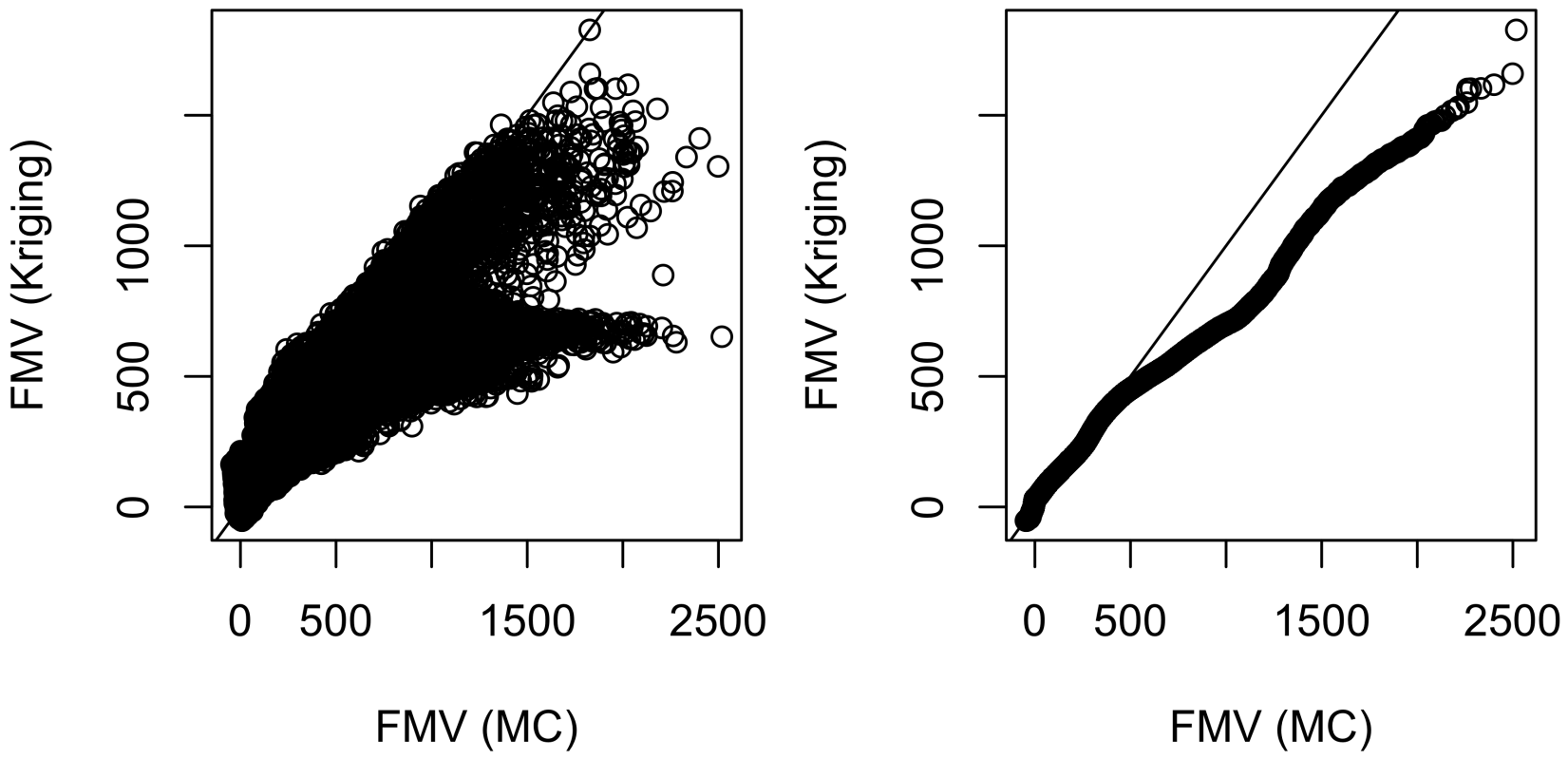


Figure: Scatter and QQ plots of the ordinary kriging model based on the clustering result from hierarchical k-means with $k = 340$.

Predictive modeling results III

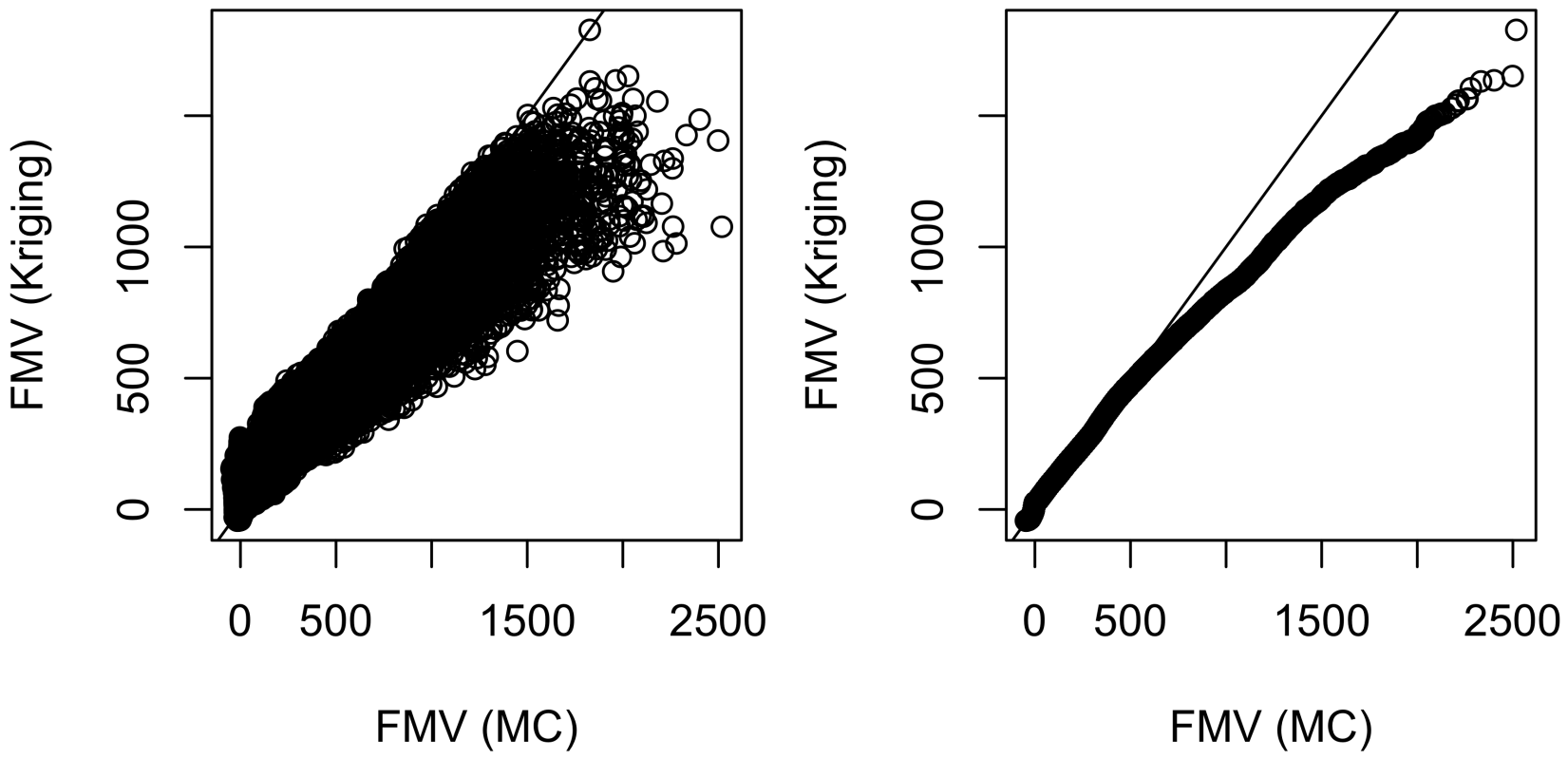


Figure: Scatter and QQ plots of the ordinary kriging model based on the clustering result from hierarchical k-means with $k = 680$.

Predictive modeling results IV

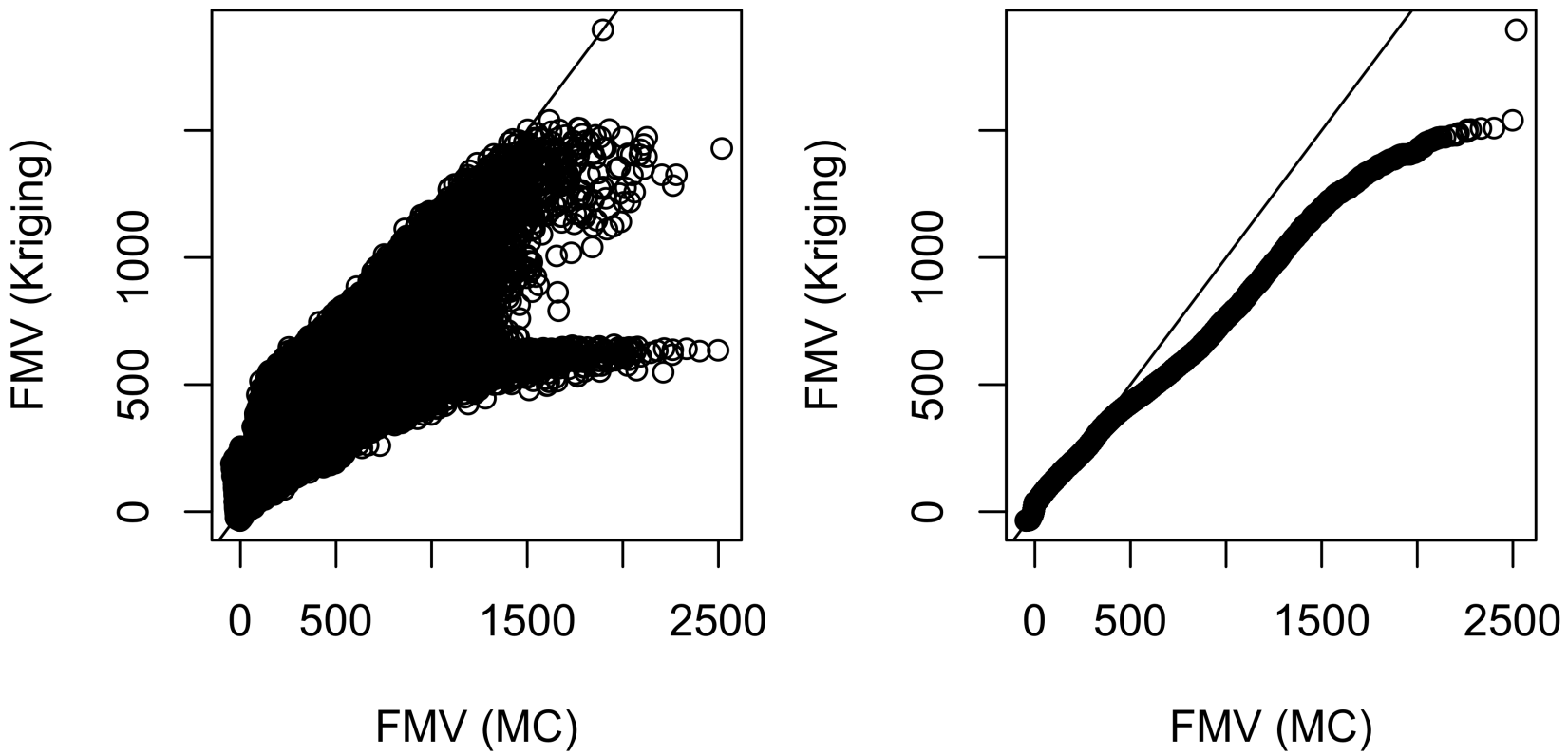


Figure: Scatter and QQ plots of the ordinary kriging model based on the clustering result from TFCM with $k = 340$.

Predictive modeling results V

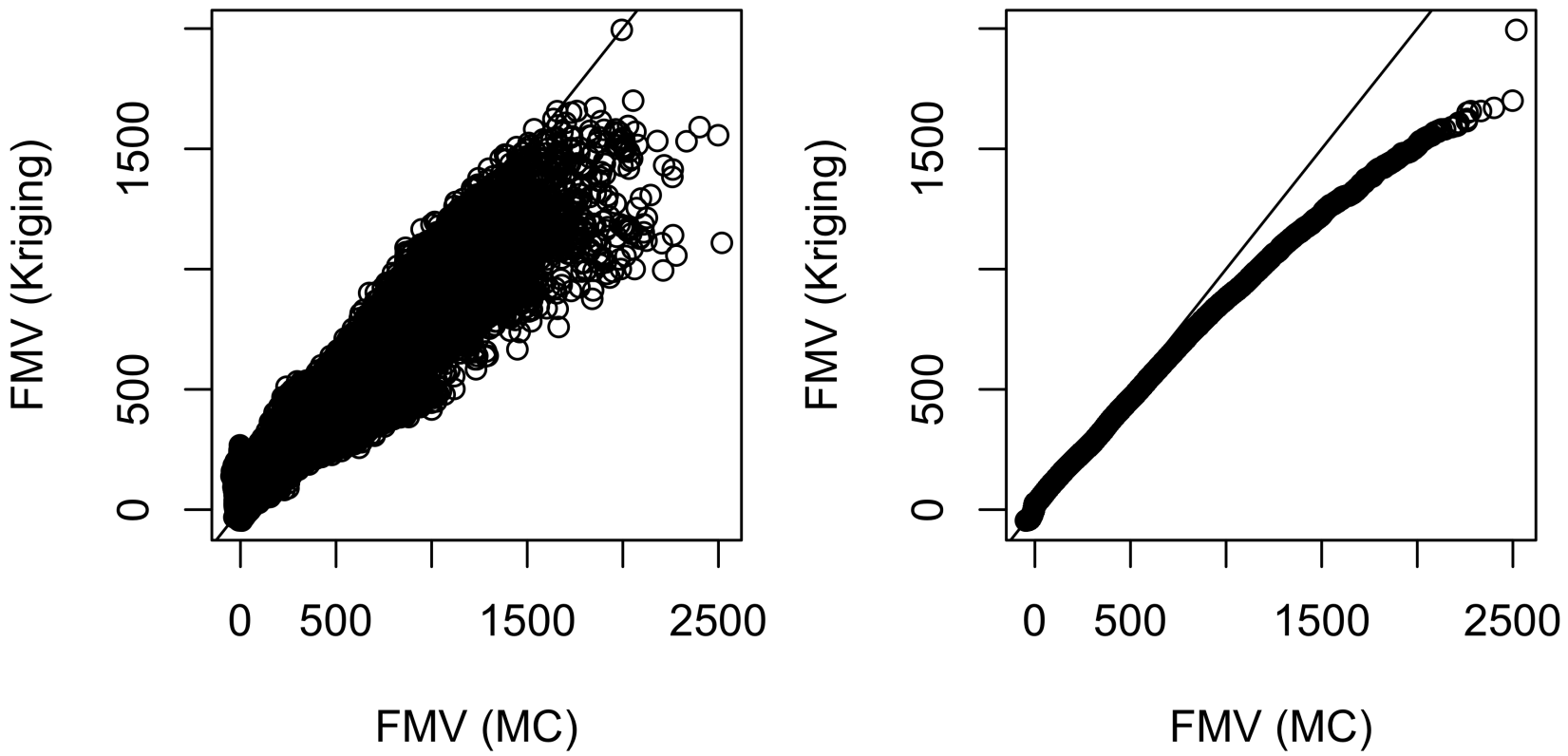


Figure: Scatter and QQ plots of the ordinary kriging model based on the clustering result from TFCM with $k = 680$.

Acknowledgements

This work is supported by a CAE (Centers of Actuarial Excellence) grant [1] from the Society of Actuaries.

[1] http://actscidm.math.uconn.edu

References I

Anderberg, M. (1973). *Cluster Analysis for Applications*. Academic Press, New York.

Bock, H. (1989). Probabilistic aspects in cluster analysis. In Opitz, O., editor, *Conceptual and Numerical Analysis of Data*, pages 12–44, Augsburg, FRG. Springer-Verlag.

Driver, H. E. and Kroeber, A. L. (1932). Quantitative expression of cultural relationships. *University of California Publications in American Archaeology and Ethnology*, 31(4):211–256.

Gan, G., Lan, Q., and Ma, C. (2016). Scalable clustering by truncated fuzzy *c*-means. *Big Data and Information Analytics*, 1(2/3):247–259.

Nister, D. and Stewenius, H. (2006). Scalable recognition with a vocabulary tree. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2161–2168.

Tryon, R. C. (1939). *Cluster analysis; correlation profile and orthometric (factor) analysis for the isolation of unities in mind and personality*. Edwards brother, Inc., Ann Arbor, MI.

Zubin, J. (1938). A technique for measuring like-mindedness. *Journal of Abnormal and Social Psychology*, 33(4):508–516.