2017 Predictive Analytics Symposium

Session 10, Clustering Techniques

Moderator:
Geoffrey R. Hileman, FSA, MAAA

Presenters:
Matthias Kullowatz
Marjorie A. Rosenberg, FSA

An application of K-means cluster analysis with missing values

Matthias Kullowatz, MS

Session 10: Clustering Techniques

September 14, 2017

SOCIETY OF ACTUARIES

Motivation

SOCIETY OF ACTUARIES

Clustering examples

- Taxonomy (e.g. organisms)
- Optimal geographic positioning (e.g. ambulances)
- Market segmentation

Background

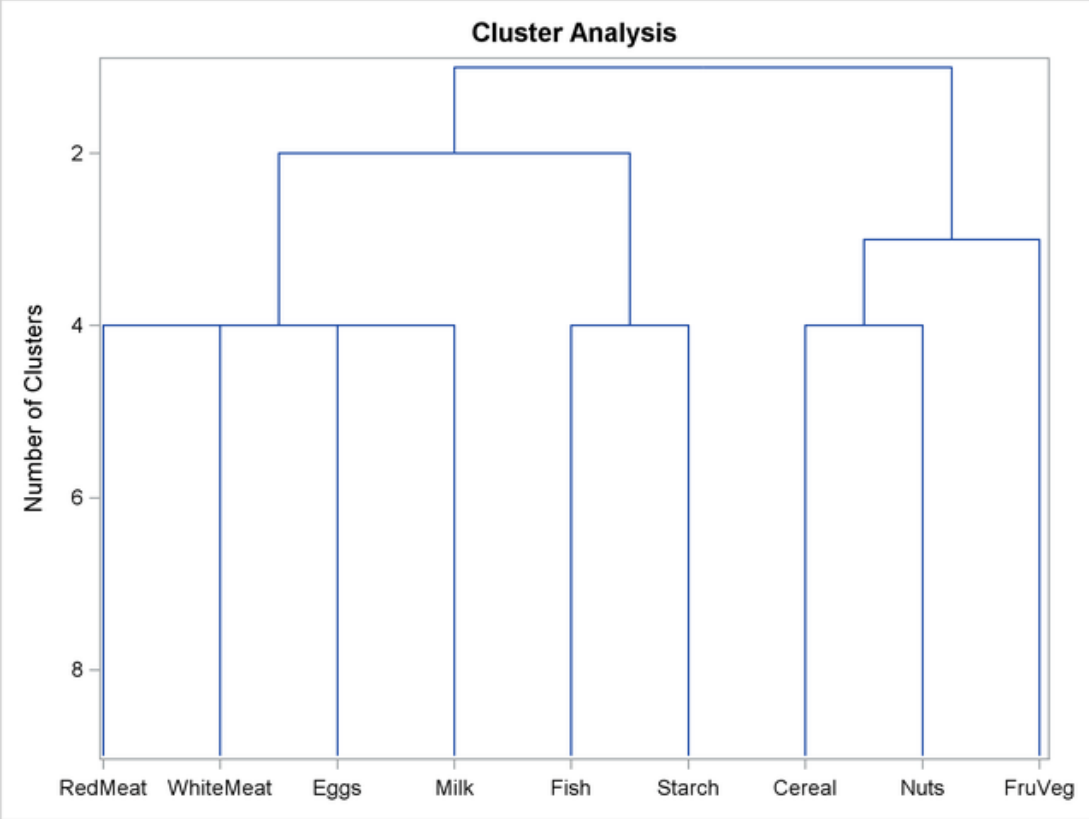
- Use enriched dataset to predict variable annuity policyholder behavior (with GLWB rider)
- Data
  - Credit info
  - Lifestyle info
  - Mortgage info
  - Census Bureau info
- Behaviors
  - Lapse
  - GLWB election timing
  - GLWB utilization efficiency

Benefits of clustering

- Produce a more implementable model
- More stable representation of complex data for long-term projections
- Create recognizable clusters to tell intuitive stories

What is (and isn't) k-means clustering?

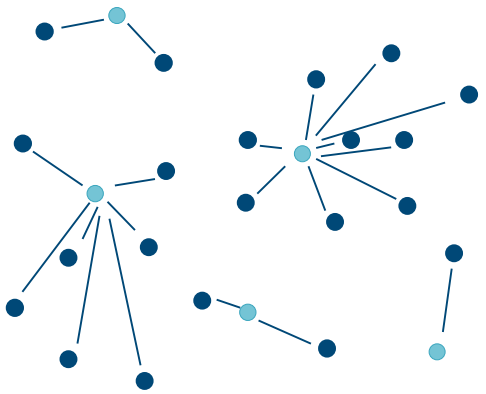
- Exclusive
- Unsupervised
- Not hierarchical



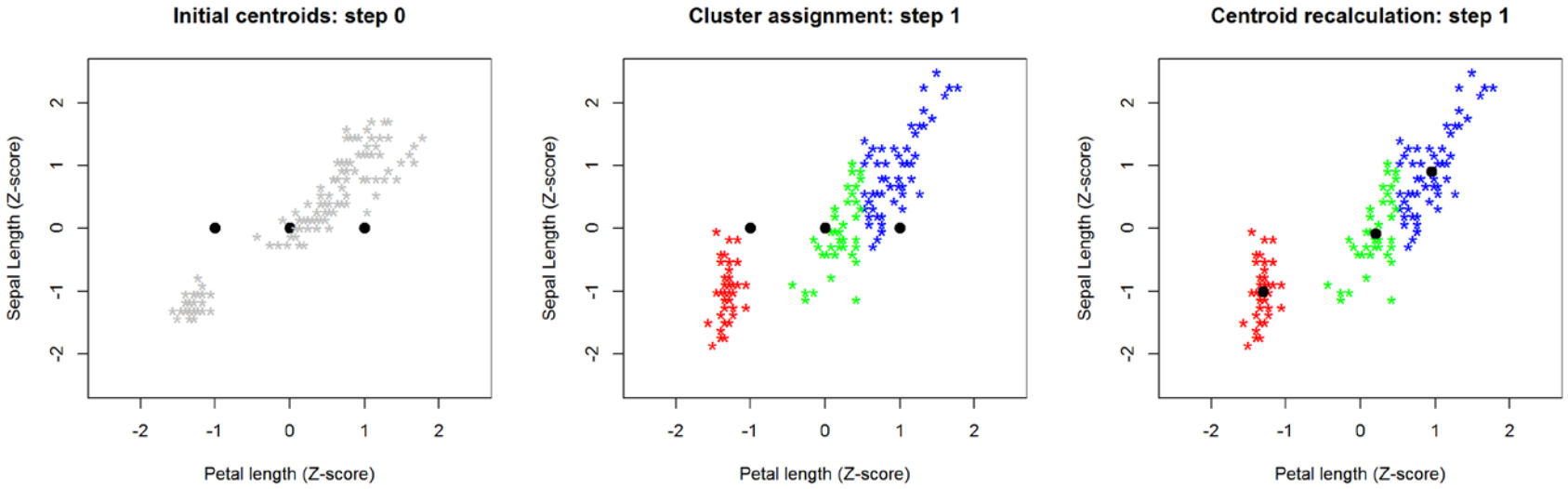
https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/images/tree1a.png

Basic k-means algorithm

- Select ***k*** cluster "centroids" in the data space
- Assign each of ***n*** observations to nearest cluster by shortest Euclidean distance to centroid
  - Standardize variables first
  - No categorical variables allowed
- Re-calculate new centroids as means
- Repeat until convergence



Visualizing K-means



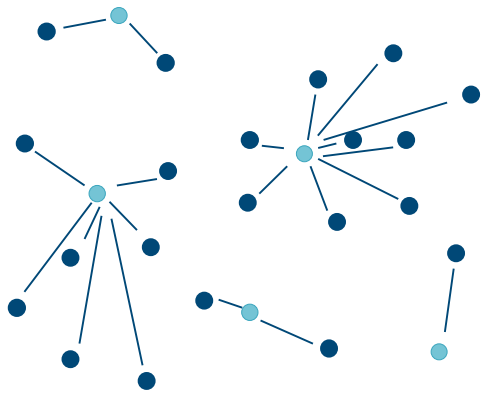
Documentation can be found at:
https://github.com/milliman/SOA_PAS_KmeansClustering

K-means with missing values

SOCIETY OF ACTUARIES

What to do with missing values

- Calculate centroid means:
  - Ignore missing values for each dimension's mean
  - Weight observations using proportion of known values
- Observation assignment:
  - Calculate distance in a reduced dimension (ignore missing values)



Example: Calculating cluster centroid

Cluster A	Credit score	Delinquencies	Population density	Home value	Appreciation
Obs A1	1.0	1	-0.5	NA	NA
Obs A2	2.0	3	-0.4	-0.5	-0.7
Obs A3	NA	NA	-0.6	-0.5	-0.3
Unweighted	1.5	2	-0.5	-0.5	-0.5
Weighted*	1.625	2.25	-0.48	-0.5	-0.55

* Our algorithm uses the weighted centroid means

Example: Calculating distance

Metric	Credit score	Delinquencies	Population density	Home value	Appreciation
Centroid A	1.625	2.25	-0.48	-0.5	-0.55
Centroid B	0.5	0.5	0.0	0.25	0.25
Obs A1	1.0	1	-0.5	NA	NA

$$Distance\ to\ A: \sqrt{0.625^2 + 1.250^2 + 0.020^2} = 1.40$$

$$Distance\ to\ B: \sqrt{0.500^2 + 0.500^2 + 0.500^2} = 0.87$$

Considerations

- How do you choose initial cluster centroids?
- Why Euclidean distance?
- Binary variables: could we handle them?
- Could weight some fields differently

Variable Annuity Example

SOCIETY OF ACTUARIES

K-means data summary

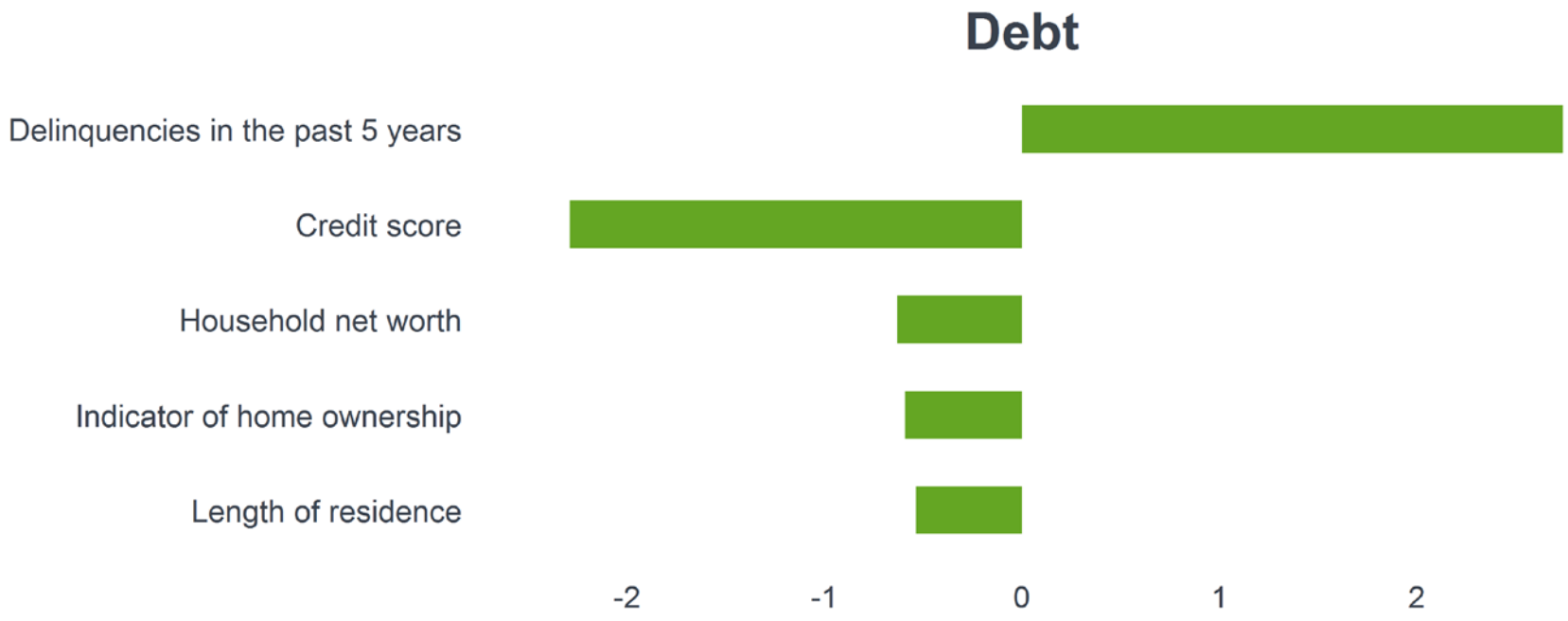
- 220,000 policyholders
- 20 third-party data fields
  - +Attained age, account value

Observations Missing	Proportion
0	11.2%
1	2.8%
2	13.3%
3	7.1%
4	20.4%
5	19.2%
6+	26.0%

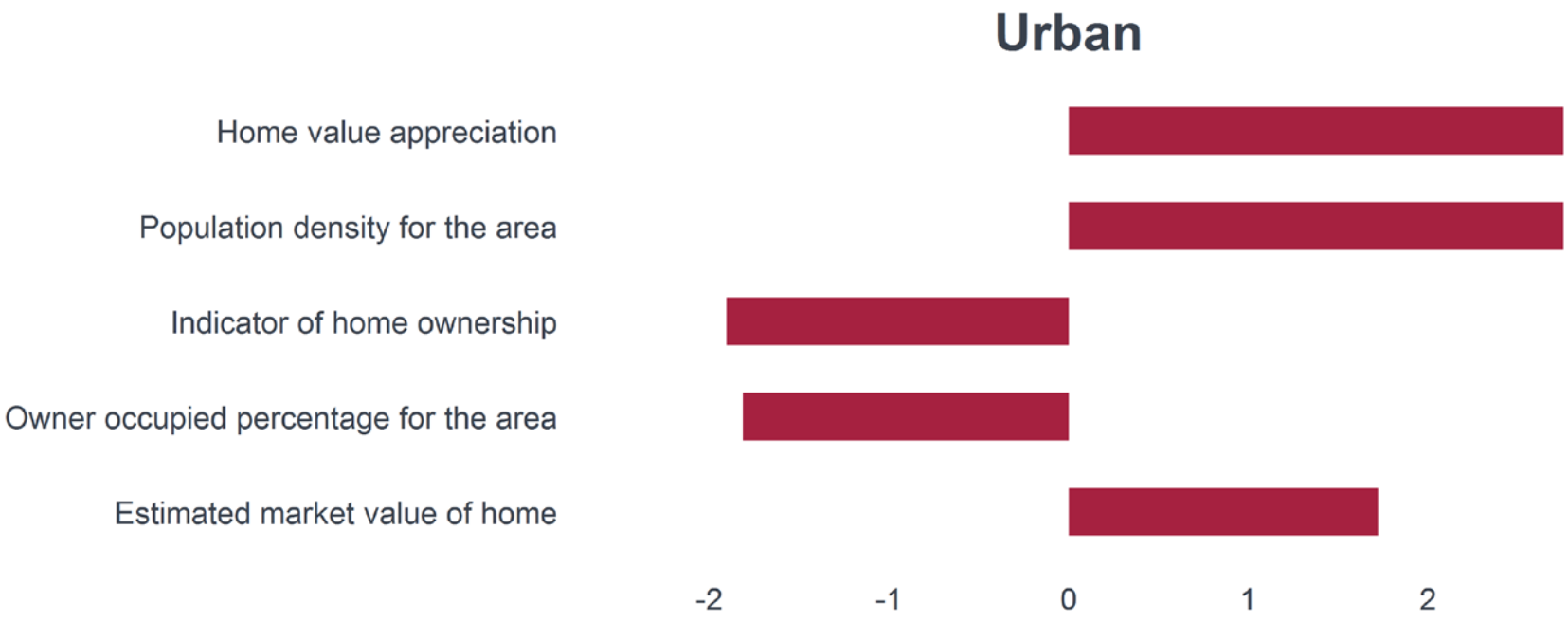
How to define clusters

- Unsupervised
  - No preconceived notion of group labels
- How different is each cluster across each dimension?
  - Calculate average Z-scores of each dimension
- Identify dimensions that make each cluster unique

The Debt cluster



The Urban Renters cluster



Customer Segments

In **Debt**:
Low credit scores, high counts of credit delinquencies in the last five years

Lower Income:
Lower than average education levels, home values, and income levels

Middle Income:
Slightly higher than average education levels, home values, and income levels

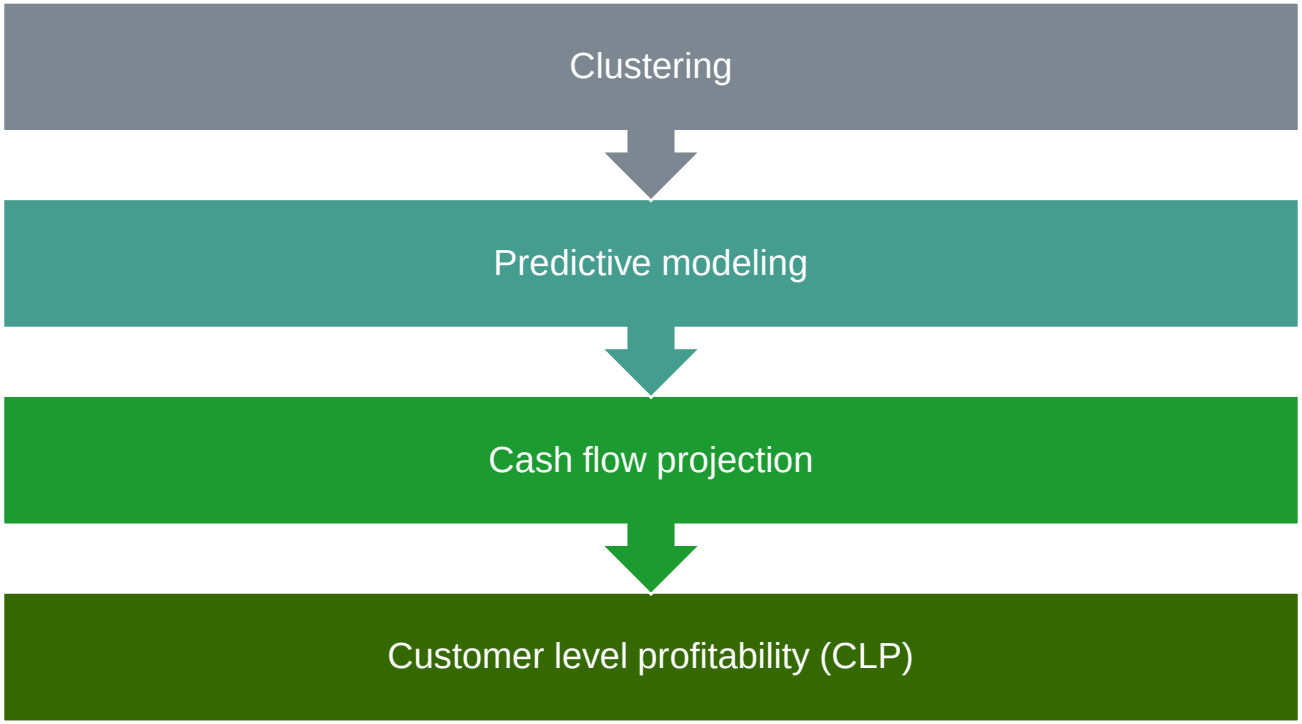
High Income:
Highest education levels, home values, and income levels

Urban Renters:
Live in high population density areas, with low proportion of homeowners

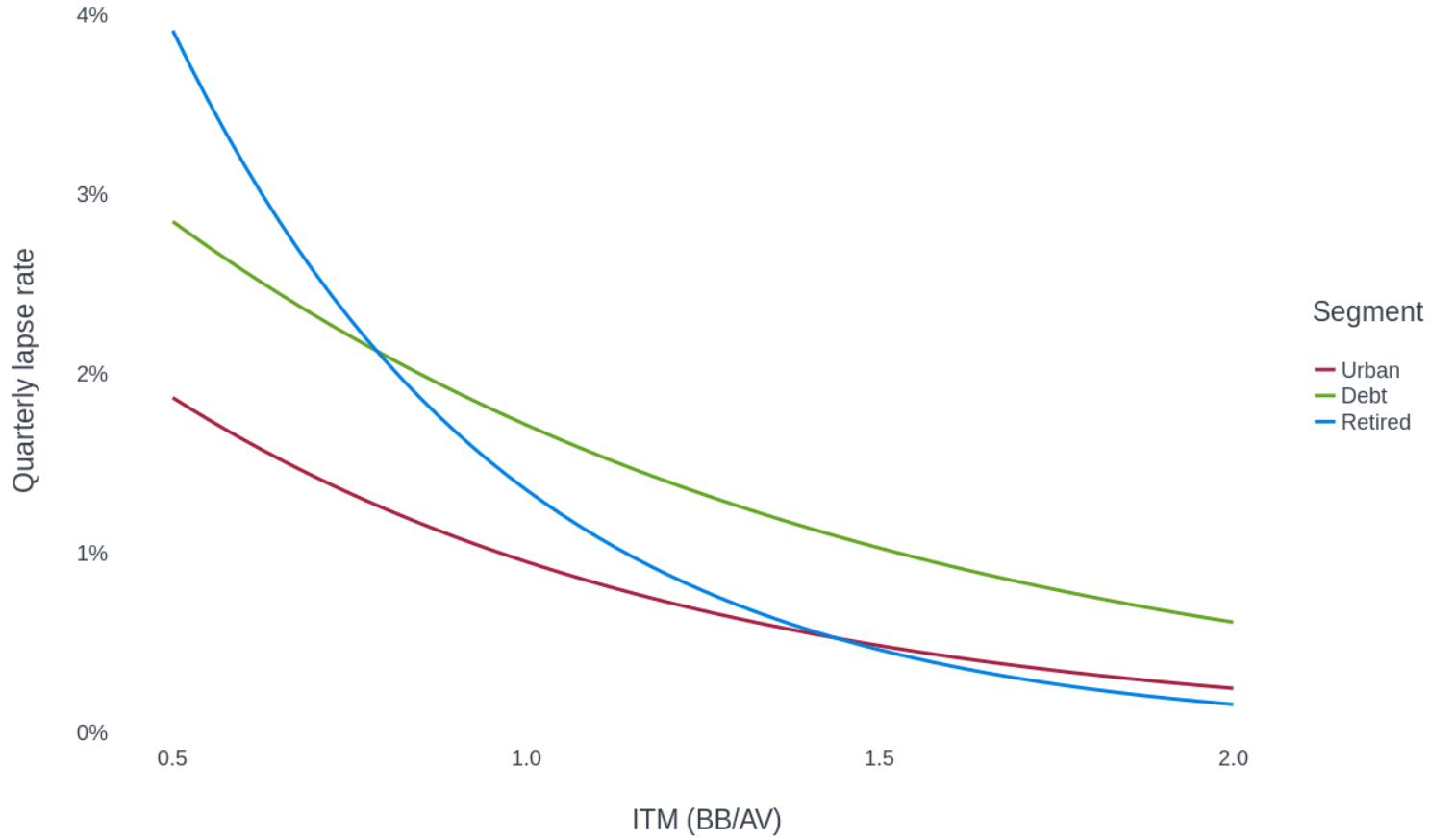
Families:
More likely to have children living at home, younger on average

Retired:
Likely to be older, and live in areas with high proportions of individuals over the age of 65

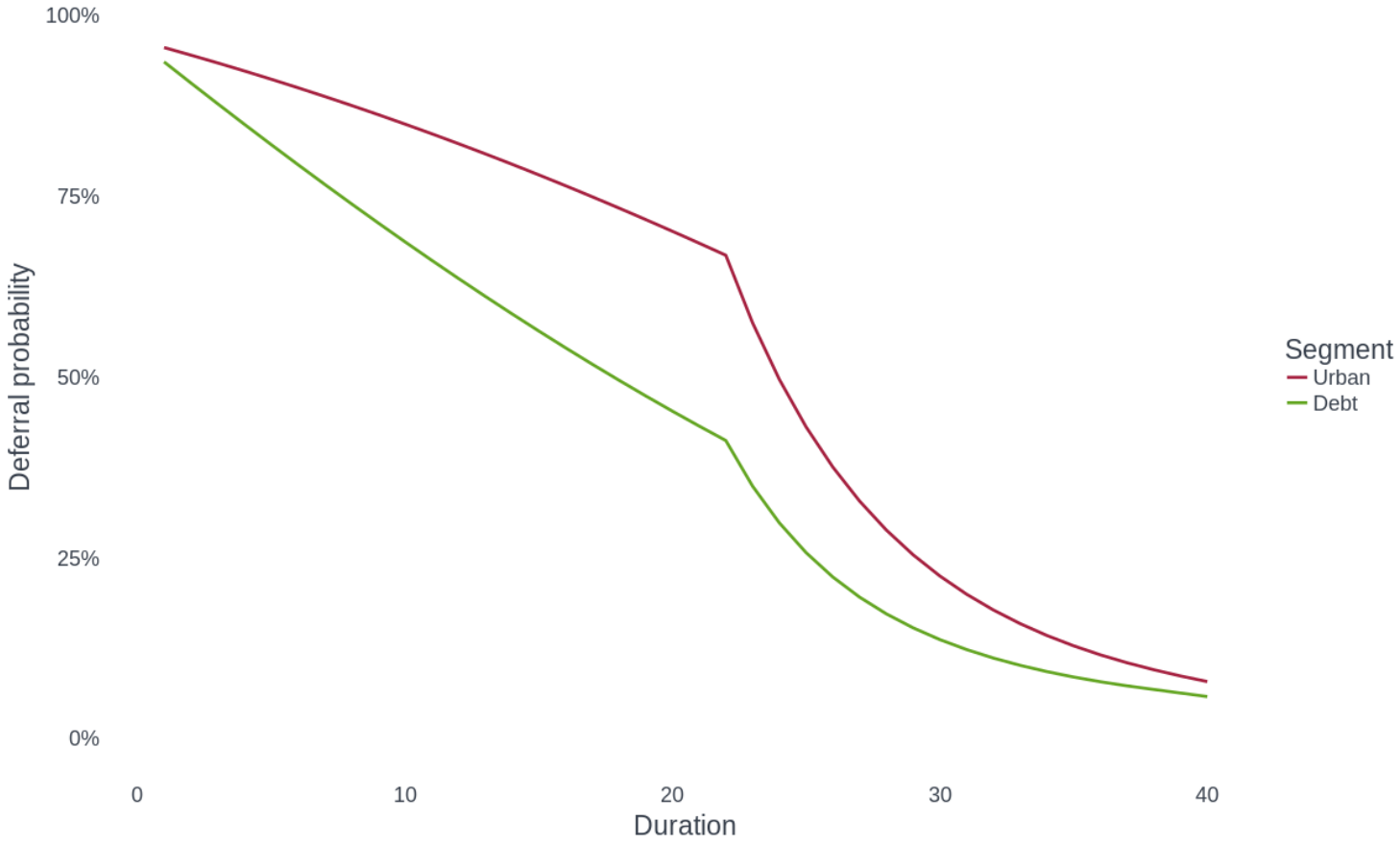
Modeling process



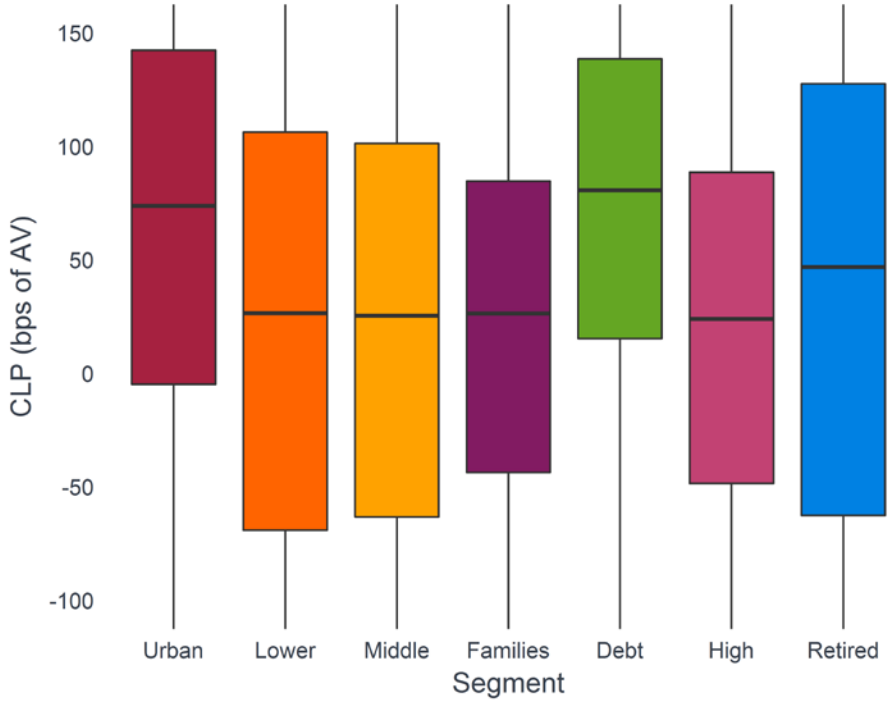
Lapse sensitivity to in-the-moneyness (ITM)



Benefit utilization deferral curves



CLP by cluster



This boxplot shows us that the policies sold to the urban, debt, and retired clusters are the most profitable on average. Also the retired cluster shows the widest range of CLP.

Thank you

SOCIETY OF ACTUARIES

A Numerical Taxonomy Application to Cluster Individuals for Predicting Healthcare Expenditures

Work in Progress

Josh Agterberg, Fanghao Zhong, Richard Crabb, Margie Rosenberg

University of Wisconsin – Madison

We acknowledge the Society of Actuaries CAE Research Grant for their partial support in this work.

Purpose

To present a novel way of clustering individuals to group *similar* individuals for prediction of health care expenditures where covariates are categorical

Secondarily: To define what is a numerical taxonomy system and its relationship to unsupervised clustering

Outline

Area of Application

- Focus on health care data
- Only categorical predictors
- That are not necessarily numbered or ordered
- With no prior clusters defined
- And with no labels as to assignment to cluster

Want to group individuals who are *similar*

Approach: Want to group individuals who are *similar*

- Data need to be definable
- Data need to be measurable
- Need to determine how two individuals are related

Sneath and Sokal (1973)

Examples of Health Care

- ICD classification
- DRG
- HCC

Some Definitions

Taxonomy (Simpson (1961)) : *Theoretical study of classification including its bases, principles, and rules*

Taxon (Sneath and Sokal (1973)) : Abbreviation for *taxonomic group*. Plural is *taxa*

Classification (Simpson (1961)) : *Ordering of entities into groups or sets on the basis of relationships*

Classification could be seen as method, but is sometimes used as outcome of process (i.e. result of classification is classification). Taxonomy could be viewed as theory (but sometimes used interchangeably)

Numerical Taxonomy (Sneath and Sokal (1973)) : *Grouping by numerical methods of taxonomic units based on their character states, using methods that are objective, explicit, and repeatable*

Numerical Taxonomic Principles

- More information available for classification, the better the classification
- Every variable has equal weight (or not?)
- Overall similarity between 2 individuals is function of individual-level similarity in each of the variables
- Distinct groups recognized due to correlation of variables within each group
- Classification based on similarity
- Inferences valid from developed clusters (originally based on biology)

Note: of course, data need to be definable and measurable

Sneath and Sokal (1973) (Note: Traced back to 1700s)

Taxonomic Principles Applied (ICD, DRG, HCC, and biology)

- Approach Defensible
- Reproducible process
- Individuals classified into reliable groups by different coders
- Set of operational protocol with standardized names
- Interpretable clusters with characteristics generally constant
- Manageable number of clusters
- Predictive power of cluster

Evans, Pope, Kautter, Ingber, Freeman, Sekar, and Newhart (2011); Feinstein (1967, 1988);

Fetter, Shin, Freeman, Averill, and Thompson (1980); Sneath and Sokal (1973)

Questions when Clustering with Only Categorical Variables

- How many variables to include?
- Which variables to include and how include?
- If ordered categorical variable, how measure differences?
- If multi-level,
  - Convert into multiple indicator variables?
  - Weight rarer levels more?
  - How incorporate (or not) missing values
- How define the center?

Goodall (1964, 1966)

Clustering Operational Protocol with Categorical Variables

To create clusters (labeled as *cluster method*), we need:

1. Definition of cluster center
2. General algorithm of how to create clusters: like k-means, k-medoids
3. Particular function in software to calculate
4. Need a definition of similarity
5. We need a matrix of dis-similarity (per numerical taxonomy literature, like distance, correlation or probability). If distance, then which kind L1, L2 or other?

Note: Above items are inter-related
End result **sensitive** to choice of algorithm

Issues with Clustering Methods

1. Appropriateness of cluster method when all variables are categorical
2. Random choice of starting point produces clusters that could differ (i.e. greedy algorithm may converge to local minimum but may not be global minimum)
3. Choice of similarity measures could change results
4. How justify clusters using taxonomic principles?

Our Cluster Protocols

1. Define center as *medoid* (center represents actual data point)
2. General clustering algorithm: PAM (Partitioning Around Medoids)
3. Cluster function: Use pam() function in cluster package in R
4. Compare two similarity measures: Gower's distance and Goodall's similarity
   - Gower: Define *similar* if attributes are equal; compute simple average
   - Goodall: Define two entities to be more *similar* if attributes are rarer; compute similarity index

NHIS/MEPS Data

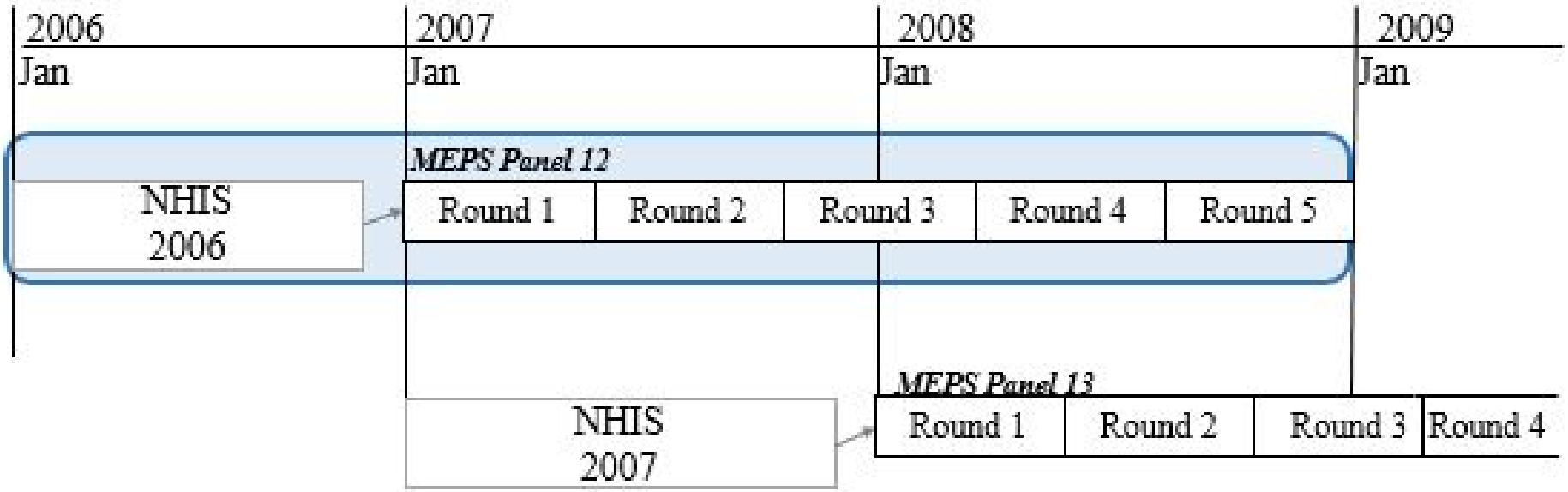
- National Health Interview Survey (NHIS) linked to Medical Expenditure Panel Survey (MEPS)
- NHIS Sample Adult Questionnaire for adult health behavior data
- Complex survey design allowing estimates of US civilian non-institutionalized population

Note: Complex design provides estimates of population, although sample (possibly allow for more reproducible results)

Study Design

- NHIS baseline year 2010: All individual-level characteristics to define initial clusters
- MEPS panel 16 (representing calendar years 2011 to 2012) for validation of clusters using expenditures

MEPS Longitudinal Design



NHIS Variables for Clusters

- Personal demographic information
- Health status information
- Living style or habits
- Working environment and status

Results

- To be filled in

Conclusions

- Numerical taxonomy literature provides history
- Taxonomy principles provide guidance to justify approach
- Operational protocol provide guidance to communicate methods
- Presence of categorical variables can complicate analysis

Bibliography I

Evans, M. A., G. C. Pope, J. Kautter, M. J. Ingber, S. Freeman, R. Sekar, and C. Newhart (2011). Evaluation of the CMS-HCC risk adjustment model. *CfMM Services, Editor*.

Feinstein, A. R. (1967). *Clinical Judgment*. Williams and Wilkens Co.

Feinstein, A. R. (1988). ICD, POR, and DRG: Unsolved scientific problems in the nosology of clinical medicine. *Archives of internal medicine 148*(10), 2269–2274.

Fetter, R. B., Y. Shin, J. L. Freeman, R. F. Averill, and J. D. Thompson (1980). Case mix definition by diagnosis-related groups. *Medical care 18*(2), i–53.

Goodall, D. W. (1964). A probabilistic similarity index. *Nature 203*(4949), 1098–1098.

Bibliography II

Goodall, D. W. (1966). A new similarity index based on probability. *Biometrics*, 882–907.

Simpson, G. G. (1961). Principles of animal taxonomy.

Sneath, P. H. and R. R. Sokal (1973). *Numerical taxonomy. The principles and practice of numerical classification.*