Model Efficiency Through Data Compression

Actuaries' Club of the Southwest

Houston, TX

November 15, 2012

Trevor Howes, FCIA, FSA, MAAA

VP & Actuary, GGY AXIS

Agenda

- The need for model efficiency
- Types of model efficiency
- Data compression by Clustering
  - How it works
  - Possible refinements to basic method
  - Advantages and disadvantages
  - Comparison to traditional grouping

Model Efficiency – the challenge

- Run time vs. resource cost challenge is faced by many applications, especially:
  - Existing valuation and required capital calculations for VA's (AG 43, C3 Phase II, FAS 133, SOP 03-1)
  - Pending valuation and required capital calculations for Life products under PBA
  - Economic Capital, MCEV, and future Solvency II
  - Hedging (Greek Calculation) and simulated hedging within actuarial valuation and projections
    - On projected inforce to simulate hedging strategy
    - Double nested projections for current valuation/capital
    - Triple nested projections for projected valuation/capital
  - Any large scale, time consuming modeling application

An IT Manager's View of an Actuarial Model

- As a result, Actuarial Models are rapidly growing in size, complexity and demand for IT resources …

Inspiration:
Audrey II from
"Little Shop of Horrors

Building a Larger Grid

- Adding more cores to your farm is not always the best option
  - Expensive – servers, memory, network connections, operating software, power, cooling, security, IT support
  - Hardware has limited lifetime – performance, reliability
  - Scalability may be an issue with some software
  - Consolidating results becomes more difficult as Grids increase
    - More nodes feeding results simultaneously into single DB
    - Stretches capacities and speed of disk drives
  - Disaster recovery implies duplicate facility built, tested, then kept idle for guaranteed availability within a day

How to Find Greater System Efficiency?

1. Improved performance hardware, new technology
   - May require redesigning software, recoding in new language
   - May place restrictions on size or design
2. Optimize software implementation for greater efficiency
   - Should have no impact on results
3. Model Efficiency Techniques (find ways to do less work and achieve acceptable model accuracy)

Note: Model Efficiency Work Group (MEWG) created in 2007 by the AAA SVL2 Committee to support PBA project

- Main MEWG web page on AAA website:
  www.actuary.org/content/academy%E2%80%99s-model-efficiency-work-group

Preferred Model Efficiency Techniques

Attractive characteristics:

- Greatest efficiency for least amount of “error” added
- Both mathematical and intuitive support
- Impact should be fully understood, readily quantifiable
- Easily incorporated in regular production routines without manual intervention
- Flexibility in application
  - option to specify reduction factor and turn on/off
- Useful throughout projections with nested stochastics and not just for current date calculations
- Useful for multiple applications and purposes

How to Find Greater System Efficiency?

What types of Model Efficiency Techniques are used?

- Model simplifications and approximations
  - Less frequent time steps
  - Ignoring or simplifying assumptions
- Scenario generation and selecting reduced scenarios
- Compressing asset and liability inforce models
- Hybrid techniques (data and scenario compression)
- All the above in combination

Model Compression Techniques

- Many companies use traditional grouped data approach:
  - Groups of similar or identical policies treated as one model point
  - Groupings formed by predefined criteria such as specific values or ranges of values; for example:
    - average issue age reflecting 5 or 10 year issue age range
    - average issue date within 3 months or full calendar year
    - similar plan codes represented by most common plan code
  - Model point is a set of identical policies reflecting the average or most common characteristics of the grouping
  - Scaling factor based typically on policy count to reflect the group
- Traditional grouping approach may still be simplest and best method in some cases!

Model Compression Techniques

- "Clustering" is an advanced compression technique
  - Some attractive features vs. traditional grouping
  - Reported to have potential for significant efficiencies in some lines of business, and some applications
  - Included in 2011 SOA funded research project
    - See report by Ernst & Young available from: www.soa.org/Research/Research-Projects/Life-Insurance/research-2011-11-model-eff.aspx
  - Also reported in various Section newsletter articles in recent years; e.g. *The Financial Reporter* June 2010 issue:

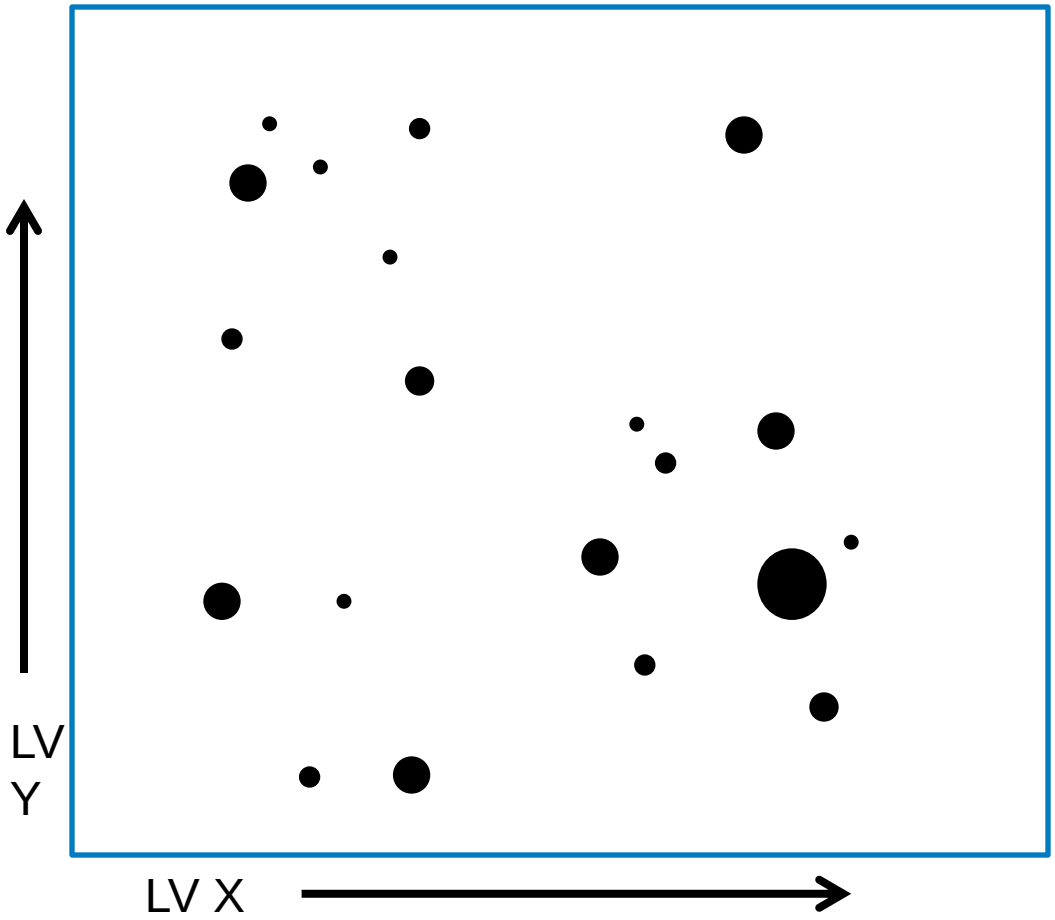
    www.soa.org/library/newsletters/financial-reporter/2010/june/frn-2010-iss81.pdf

Clustering Compression Technique

- "Clustering" is a generic technique used in data analysis
- For Model Compression useful for actuarial applications, typically "agglomerative" clustering algorithm is used:
  - Each policy starts as its own "cluster"
  - Distance (similarity) between every pair is calculated using chosen location values (characteristics) for each policy
  - Least important cluster by size and distance from nearest cluster is merged into nearest cluster
  - Original policy in cluster is representative policy of cluster; scaling factor calculated to reflect total volume of cluster
  - Clustering process continues until target ratio (# of clusters) met
  - Better representative policies may be found when finished

Clustering Process Illustration

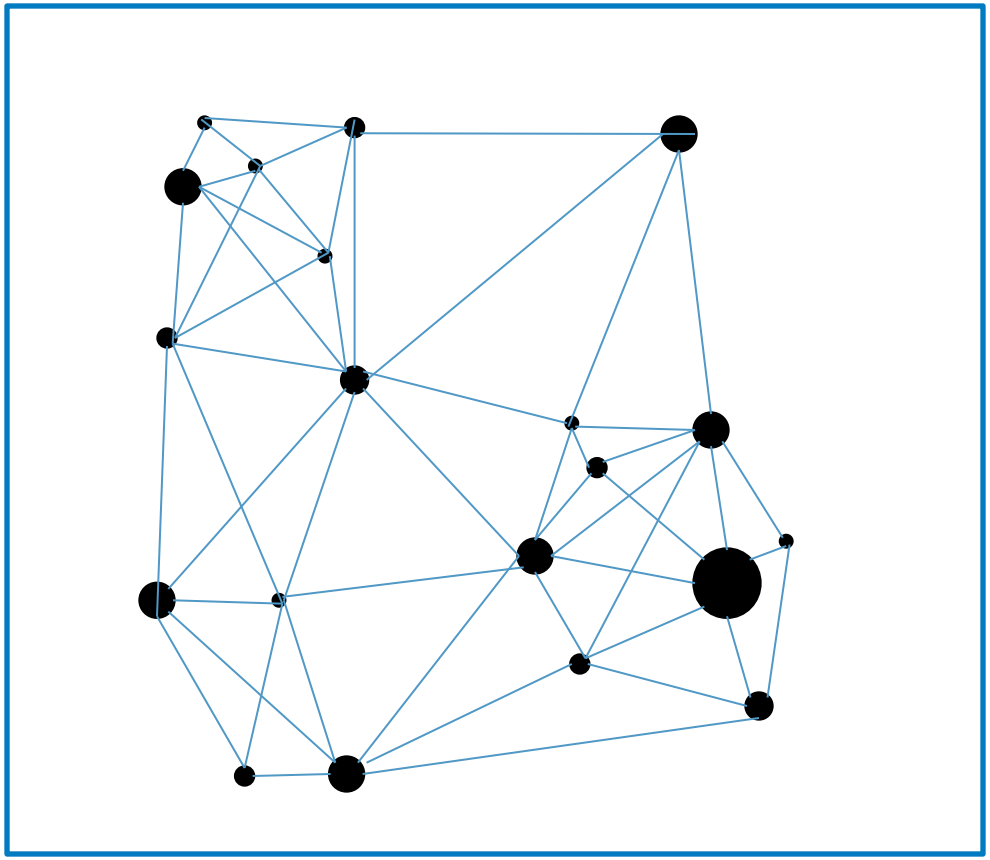
- Step 1: Seriatim data assigned location variables and size



- 20 seriatim policies of varying sizes and characteristics
- Target compression ratio of 20% (5:1)

Clustering Process Illustration

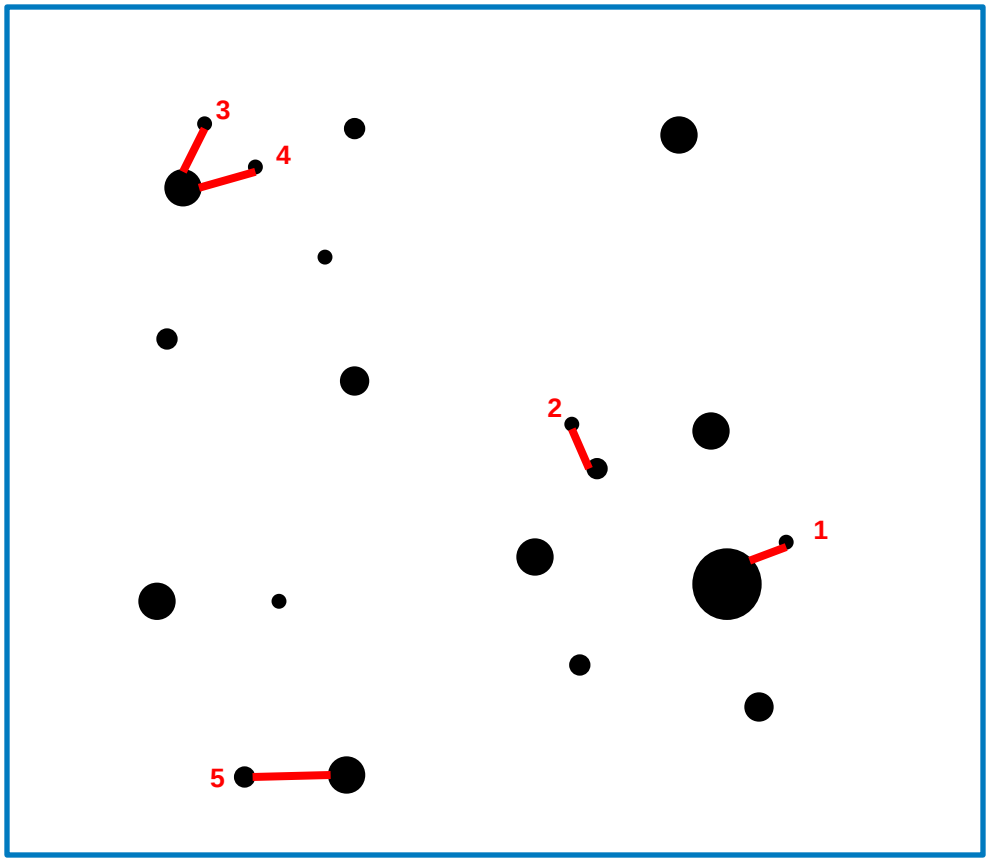
- Step 2: Distances between all pairs calculated



Distances determine order of clustering from closest to furthest

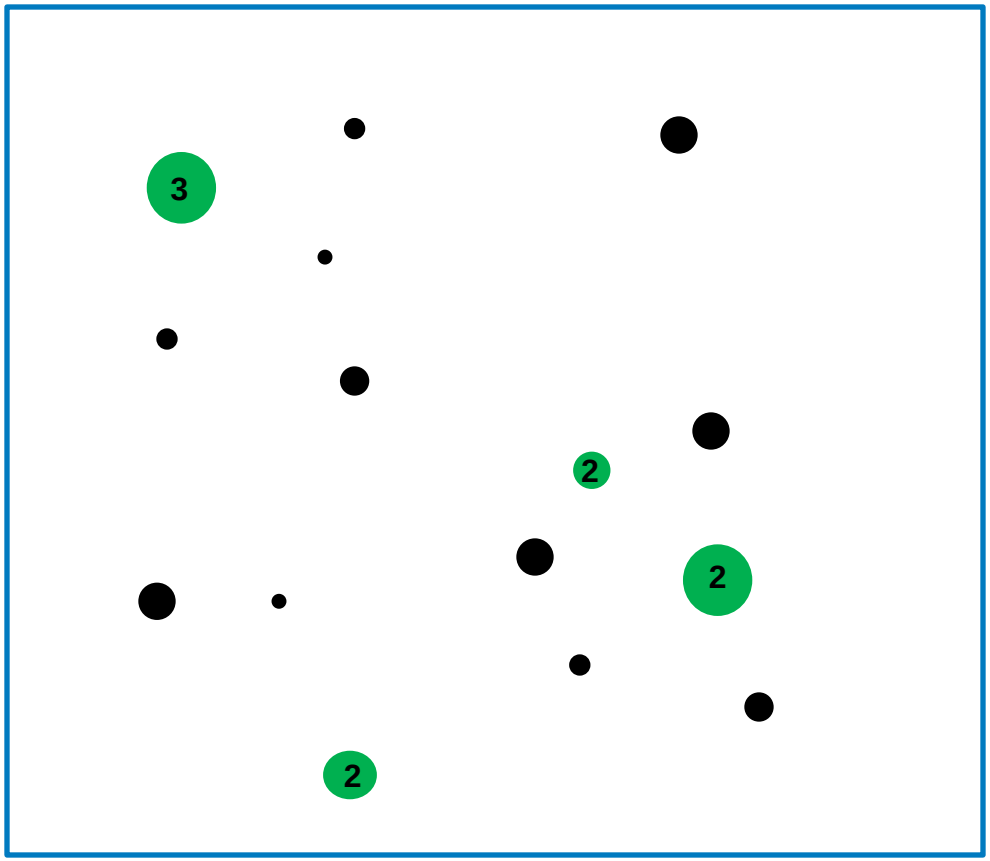
Clustering Process Illustration

- Step 3: Closest pairs identified

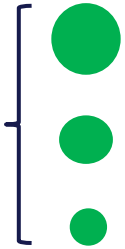


Clustering Process Illustration

- Step 4: Least important clustered with larger adjacent point

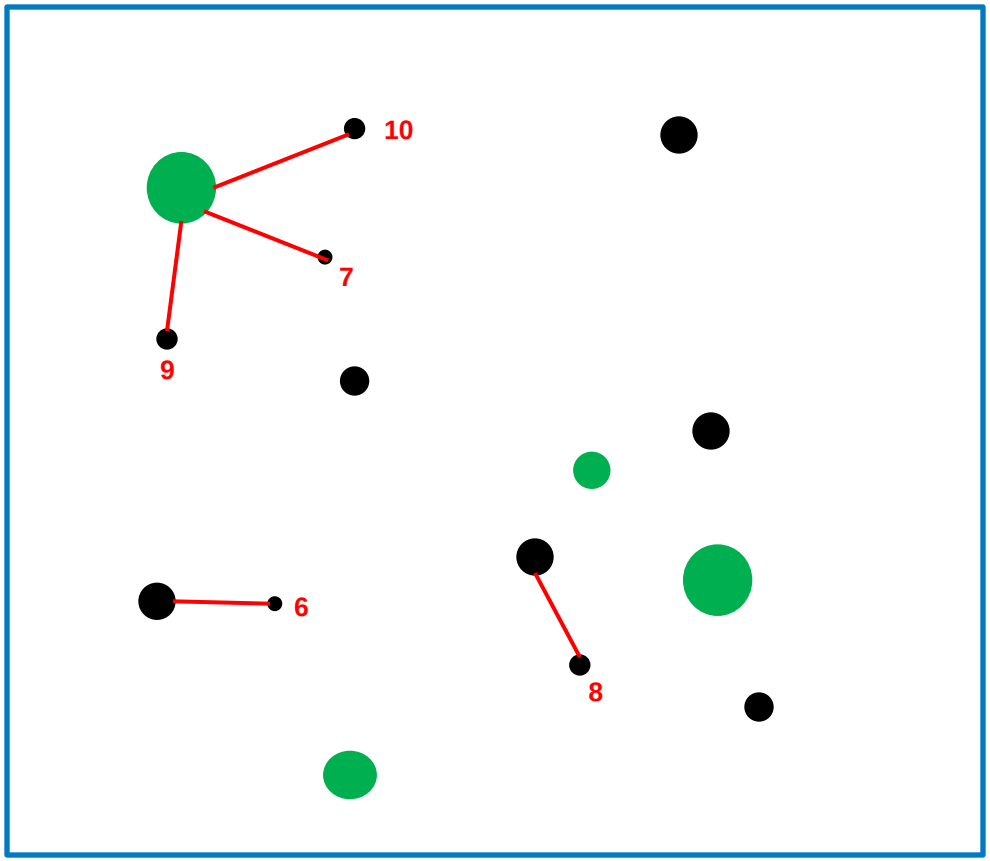


Clustered policies



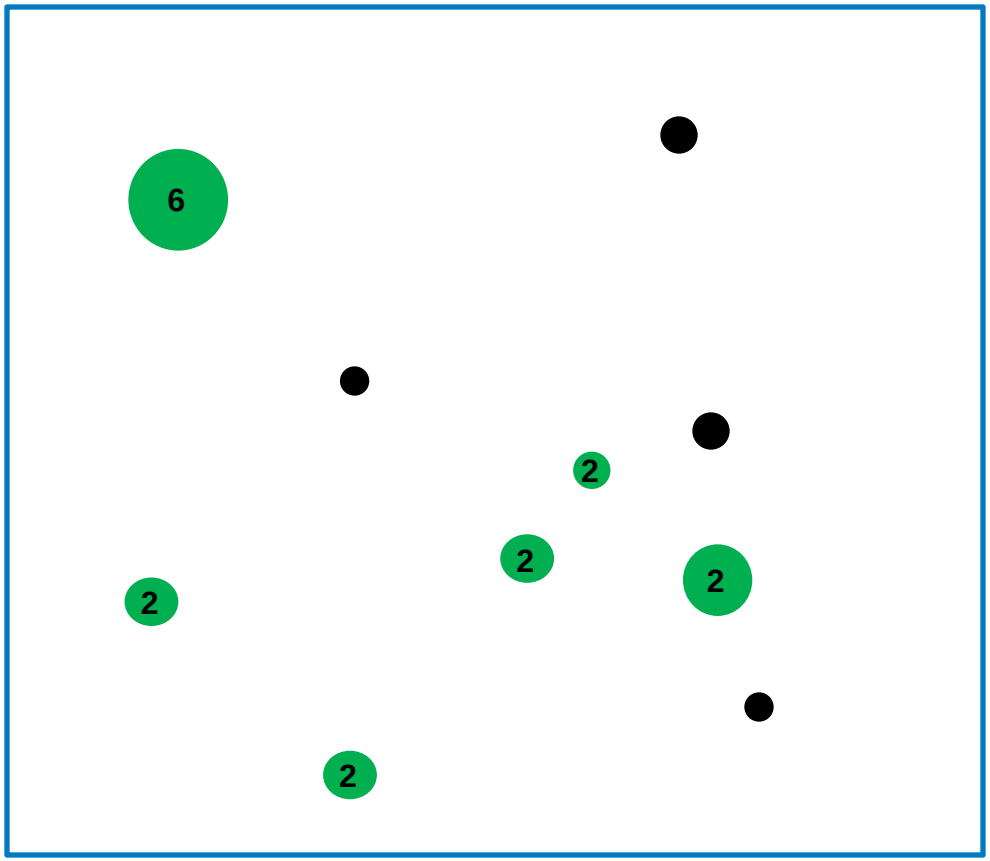
Clustering Process Illustration

- Step 5: Next closest pairs identified

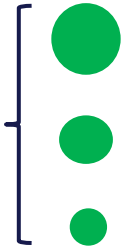


Clustering Process Illustration

- Step 6: Least important clustered with larger adjacent point

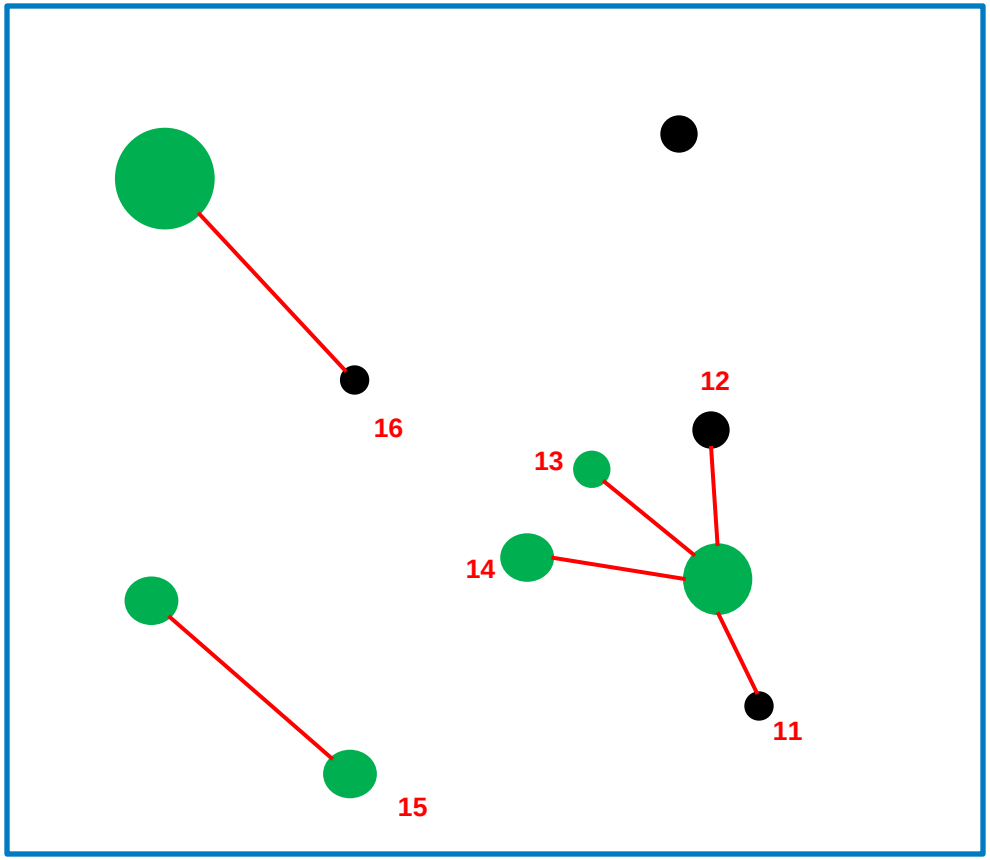


Clustered policies



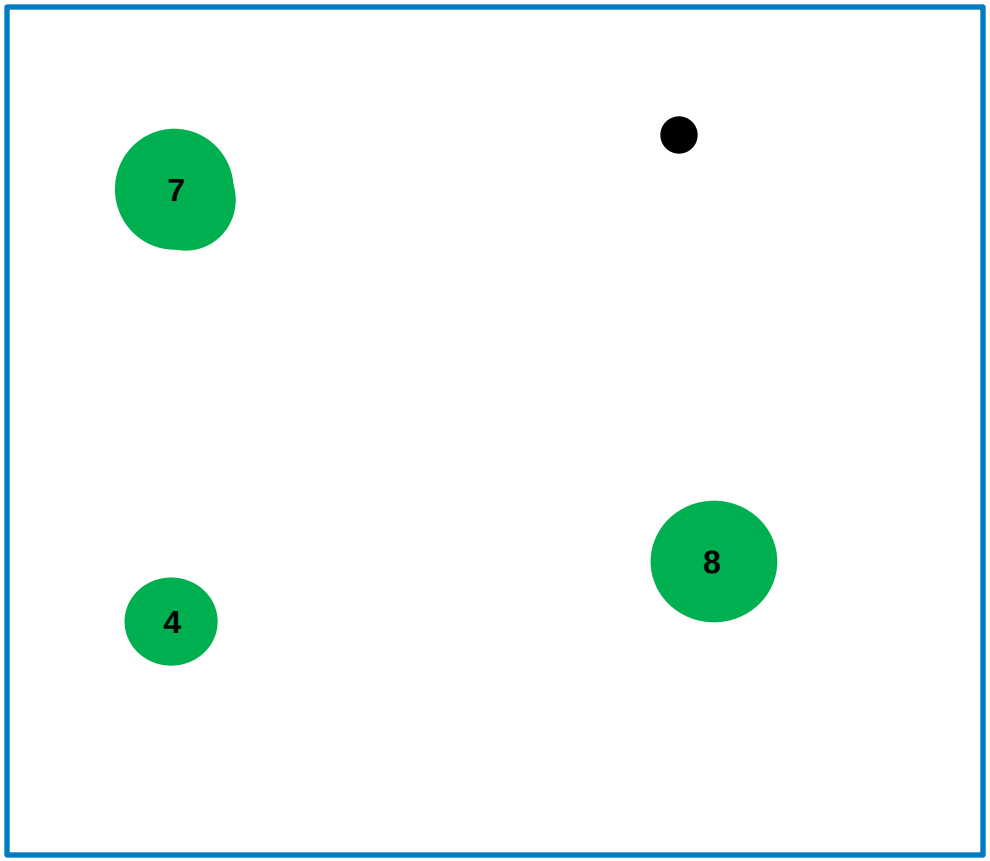
Clustering Process Illustration

- Step 7: Remaining closest pairs identified to reach target



Clustering Process Illustration

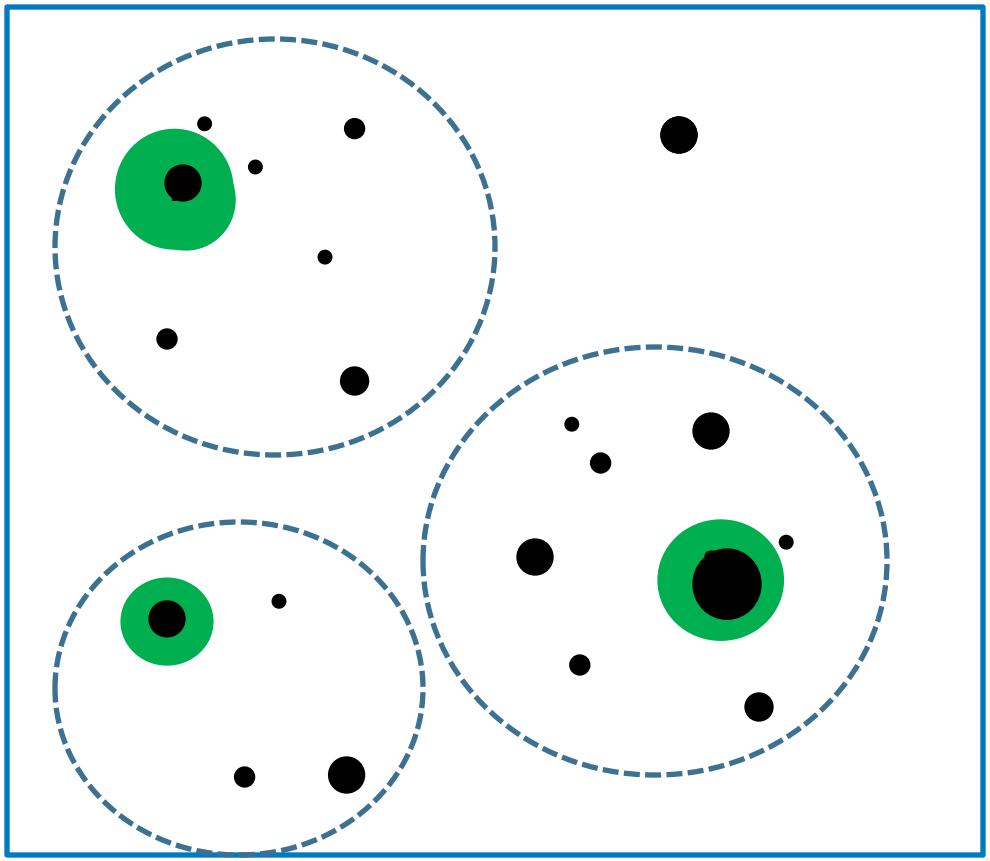
- Step 8: Final clustering performed



Target compression level of 20% reached

Clustering Process Illustration

- Result: Highly compressed model and original model



- 20 original policies compressed into 4 clusters
- Representative policies may need to be reset

Refinements to Basic Clustering

- Total portfolio divided into segments; clustering applied only within each segment
  - Different target compression ratios may be set by segment
  - Segments with zero compression target left as seriatim
  - Segment sizes may impact performance of clustering step since distance table (triangle) size is based on N squared

- Multiple clustering processes with different target ratios may be used to generate compressed models for different purposes, or for research into impact
  - one compression run may generate multiple compressed models

Refinements to Basic Clustering

- Choices of Distance measure – measure of similarity that determines order of clustering
  - Distance calculations based on comparing location variables
  - Choices include:
    - Euclidean distance (Square root of sum of squares)
    - Square of Euclidean distance
    - Sum of absolute values of differences
  - Different measures may isolate or combine outliers more effectively
  - Will need to normalize each location variable
    - may also want to apply weightings

Refinements to Basic Clustering

- Choosing Measure of Size and Importance
  - A volume or size “Measure” must be selected be used to accumulate the total size of each cluster and the scaling factor to apply to the representative policy
    - Usually Face Amount or Fund Value, but others possible
    - Only the measure used for scaling can exactly replicate the portfolio; other measures will be approximate
  - A variation would be to allow each Representative Policy values to be adjusted to the cluster average
    - Greater effort, increased risk (need to assure that the adjusted policy is internally consistent, realistic, etc.)

Refinements to Basic Clustering

- Using Different Linkage Rules (How to define distance to a cluster?)
  - Which policy to use?
    - Original representative policy when first merged?
    - Nearest policy, or largest policy?
    - Policy closest to centroid?
  - How often is this representative policy reset for the cluster?
    - Only after target ratio reached?
    - At predefined stages of compression or numbers of steps?
    - Every step?
    - Note: all relevant distances may have to be recalculated for changed clusters unless original distance table retained

Clustering vs. Grouping

- Like traditional grouping, a “central” policy represents the whole cluster, by scaling up the calculated policy results
  - Grouping typically scales by policy count
  - Clustering scales by user selected measure
- Clustering differs from traditional grouping in that:
  - It can work across typical boundaries (sex, UW class, plan)
  - No need to predefine groupings, ranges of ages, issue dates
  - Target compression ratio may be set and compression continues until met
  - Easy to increase or decrease compression ratio to achieve acceptable distortion vs. full seriatim
  - Easy to generate multiple Data Models of different compression levels for different purposes in one process

Clustering Case Study in FR Section Newsletter

- Variable Annuity block with GMAB, GMDB, GMIB, GMWB
  - 100,000 plus policies, $9.5 billion fund value

Figure 3 Impact of Modeling on AG 43 Results ($ millions)			
Compression ratios:	Liability Cell Count	Stochastic Reserve	Ratio to Seriatim
100%	Seriatim	$143.6	100.0%
5.0%	5,000	$144.2	100.4%
2.5%	2,500	$143.9	100.2%
1.0%	1,000	$141.6	98.6%
0.25%	250	$140.6	97.9%
0.05%	50	$136.7	95.2%

Clustering Case Study in FR Section Newsletter

- Variable Annuity block running AG 43
  - 100,000 plus policies, $9.5 billion fund value
  - GMAB, GMDB, GMIB, GMWB
  - Case study used following location variables calculated across 5 representative scenarios:

    - initial GMB face amount for each benefit type and guarantee type,
    - initial account value in-force by fund,
    - present value of net revenue,
    - present value of commission income,
    - present value of revenue sharing,
    - present value of maintenance expenses,
    - present value of M&E fee income, and
    - present value of net benefit costs for each GMB type (benefits paid less associated charges).

    easily obtained from in force files

    requires some seriatim preprocessing

Clustering Case Study in FR Section Newsletter

Figure 1
Analysis of Fit Variables as of Valuation Date
($ millions)

			Ratio to Seriatim for Differing Cell Counts				
	Weights	Seriatim	5,000	2,500	1,000	250	50
Inforce GMB Face Amounts							
GMDB Ratchet	1	$7,733	99.8%	99.8%	99.2%	98.9%	93.6%
GMDB Rollup	1	$4,058	97.6%	96.3%	93.9%	92.4%	94.4%
GMDB ROP	1	$4,515	100.5%	100.9%	103.6%	106.6%	122.5%
GMIB Ratchet	1	$7,545	100.0%	100.0%	99.7%	100.6%	98.2%
GMIB Rollup	1	$8,181	100.4%	100.4%	100.4%	100.6%	99.3%
GMAB ROP	1	$281	99.7%	99.1%	100.0%	94.3%	63.9%

Note the varying impact of compression by benefit type

Clustering Investigation or Implementation

- Clustering algorithms are generally available in software used for data analysis, statistics (e.g. MatLab)
- Available in at least two common modeling platforms
- Preferably it should be easily integrated into your production process and require little manual intervention
- You will need research to discover
  - optimal choices for segments, location variables and target ratios in each product portfolio and application
  - the corresponding distortion in application results

Model Efficiency Through Data Compression

THANK YOU!