Handling High Dimensional Variables

Roosevelt C. Mosley, Jr., FCAS, MAAA
Pinnacle Actuarial Resources, Inc.
March 16, 2010

PINNACLE
ACTUARIAL RESOURCES, INC.
The Firm of Choice

Discussion Topic

- The problem
- Techniques for handling high dimensions
- Comparisons of different techniques
- Conclusions

PINNACLE
ACTUARIAL RESOURCES, INC.
The Firm of Choice

The Problem

- High dimensional variables: variable with many units or levels
- Examples
  - ZIP code
  - Vehicle classification
  - Workers compensation classes
- Complication
  - Credibility at individual levels
  - Determining proper groupings

PINNACLE ACTUARIAL RESOURCES, INC. The Firm of Choice

Complications of High Dimensional Variables

- Credibility at individual levels
  - Models convergence errors
  - Results that do not make sense
- Examples
  - Thousands of ZIP codes
  - Thousands of model year/make/model/series combinations

PINNACLE ACTUARIAL RESOURCES, INC. The Firm of Choice

Techniques for Handling High Dimensional Variables

PINNACLE ACTUARIAL RESOURCES, INC. The Firm of Choice

Techniques for Handling High Dimensional Variables

- Groupings
  - Simple
  - May not be immediately obvious
  - May not be practical
- Curve fits
  - Great for continuous variables
- **Clustering using target variable** *
- **Clustering/Segmentation analysis based on inputs**
- **Use characteristics of variable level is model**
- **Principal Components**

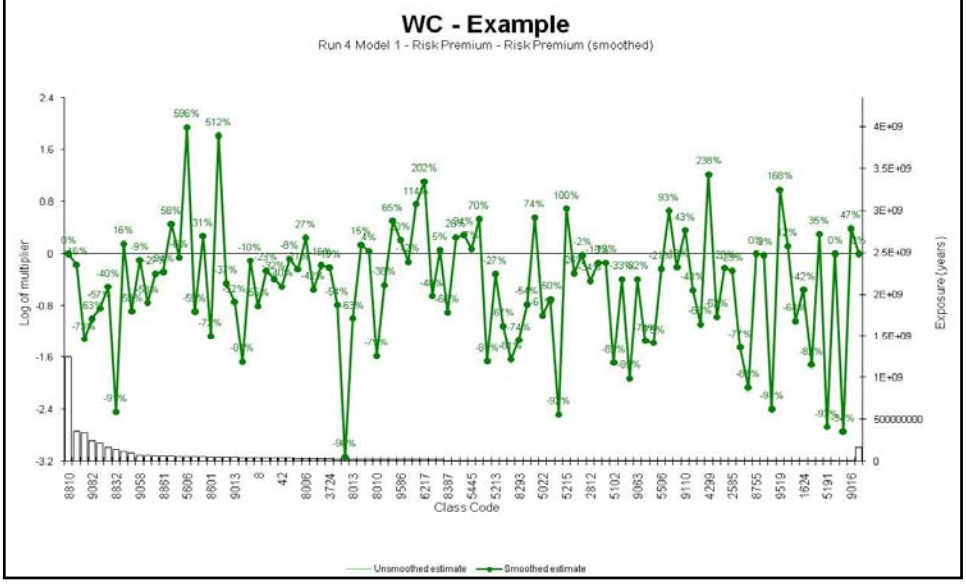
PINNACLE ACTUARIAL RESOURCES, INC. The Firm of Choice

Clustering Using Target Variable

- Combine levels of high dimensional variable by similarity in target variable
- For levels without credible experience, combine with credibility complement
  - Proximity - ZIP
  - Similar type – Make/model
  - NCCI loss costs – Worker's Compensation Class
- Put cluster back into predictive model

PINNACLE
ACTUARIAL RESOURCES, INC.
The Firm of Choice

Worker's Compensation Class

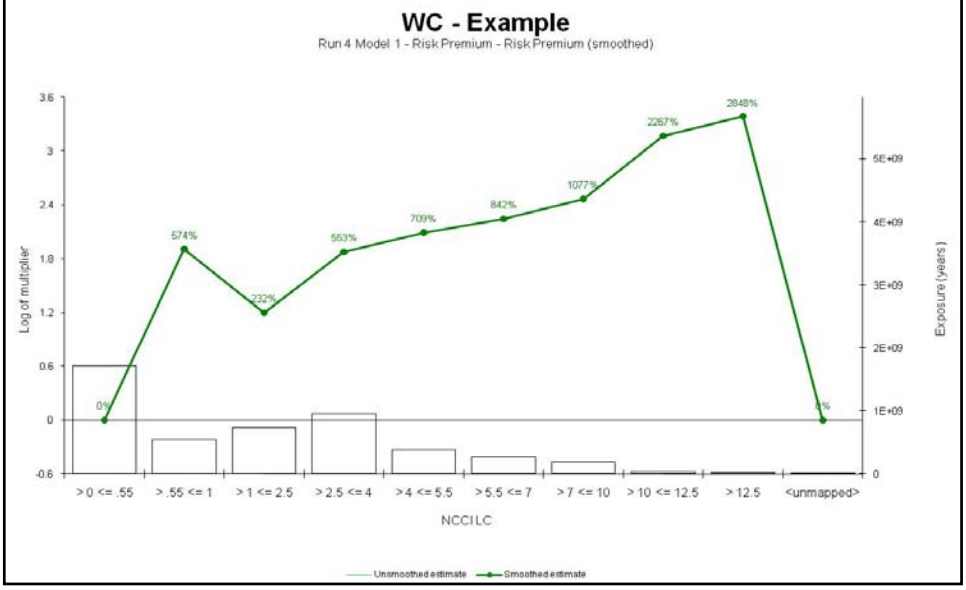


WC Class Analysis

NCCI Class Code	Class Description	Revised Company Exposure ($00)	2004 NCCI Loss Cost	Proposed Loss Cost	Change from Current
7228	Trucking: Local Hauling Only	66,891	9.81	6.68	-32%
7380	Commercial Drivers	95,217	5.48	7.39	35%
7382	Bus Company	12,658	5.41	6.04	12%
8006	Gasoline Station	52,068	3.78	5.39	42%
8385	Bus Company	10,496	3.47	4.32	25%
8387	Service Stations	32,628	3.20	1.83	-43%
8391	Automobile Body Repair Shop	118,611	4.01	2.64	-34%
8393	Automobile Body Repair	33,165	2.99	0.98	-67%
8748	Automobile Salespersons	96,971	1.75	2.36	35%
42	Landscape Gardening & Drivers	70,943	8.45	4.46	-47%
5190	Electrical Wiring - Within Buildings	206,541	3.60	2.51	-30%
5437	Carpentry - Interior	111,300	6.87	4.83	-30%
5474	Painting Or Paperhanging	113,721	6.65	4.36	-34%
5606	Contractor - Executive Supervisor	113,489	2.09	14.49	593%
9052	Hotel: All Other Employees	863,123	3.22	2.53	-21%
9058	Hotel: Restaurant Employees	167,039	2.15	2.34	9%
9082	Restaurant NOC	709,574	2.61	1.83	-30%
9083	Restaurant: Fast Food	139,396	2.45	3.36	37%
8017	Store: Retail NOC	280,356	2.40	2.74	14%
8033	Store: Meat, Grocery & Provision Stores	458,603	2.85	2.14	-25%
8601	Architect Or Engineer - Consulting	135,932	0.99	1.25	26%
8742	Salespersons, Collectors or Messengers	672,184	0.95	1.20	26%
8810	Clerical Office Employees NOC	2,628,543	0.55	0.74	34%
8820	Attorney - All Employees & Clerical	148,121	0.44	1.48	237%
8832	Physician & Clerical	342,051	0.63	0.39	-38%

The Firm of Choice

NCCI Loss Cost Indication



Clustering/Segmentation

- Unsupervised classification technique
- Groups data into set of discrete clusters or contiguous groups of cases
- Performs disjoint cluster analysis on the basis of Euclidean distances computed from one or more quantitative input variables and cluster seeds
- Objects in each cluster tend to be similar, objects in different clusters tend to be dissimilar
- Can be used as a dimension reduction technique

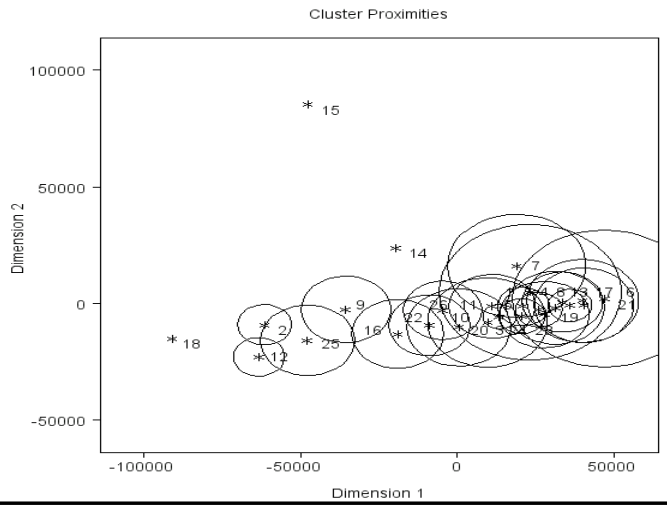
PINNACLE ACTUARIAL RESOURCES, INC. The Firm of Choice

Cluster/Segmentation Steps

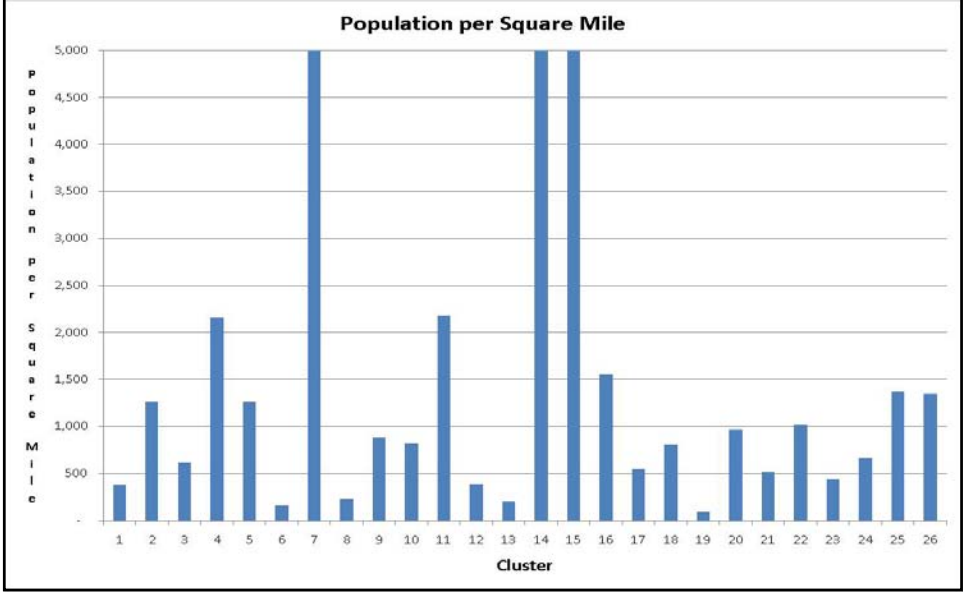
- Begin with descriptive variables related to high dimensional variables
  - ZIP code: geo-demographic data
  - Make/Model: vehicle characteristics
- Levels of high dimensional variables with like descriptive variables will be in like clusters
- Assumes descriptive variables will be related to ultimate target (frequency, severity)

PINNACLE ACTUARIAL RESOURCES, INC. The Firm of Choice

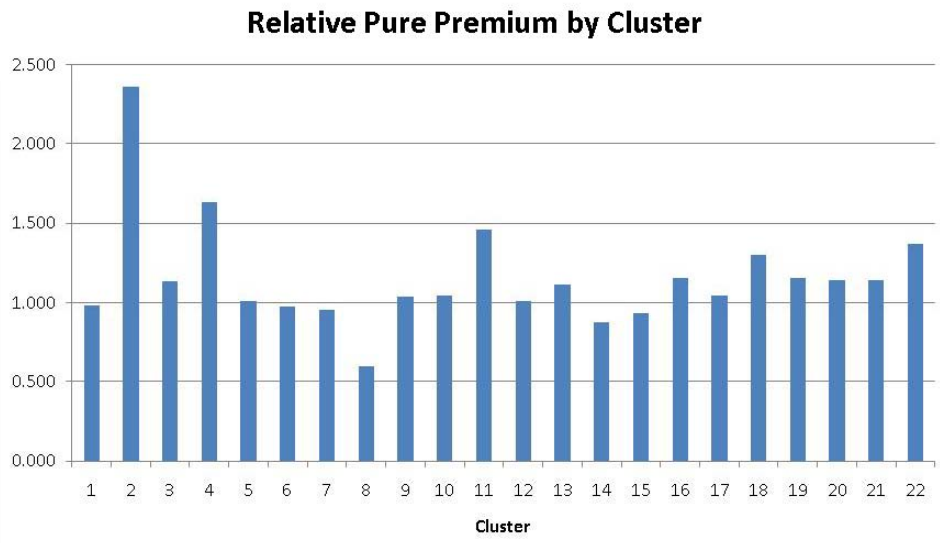
Cluster Proximity Map



Population per Square Mile by Cluster



Relative Pure Premium by Cluster



Use Variable Characteristics Directly

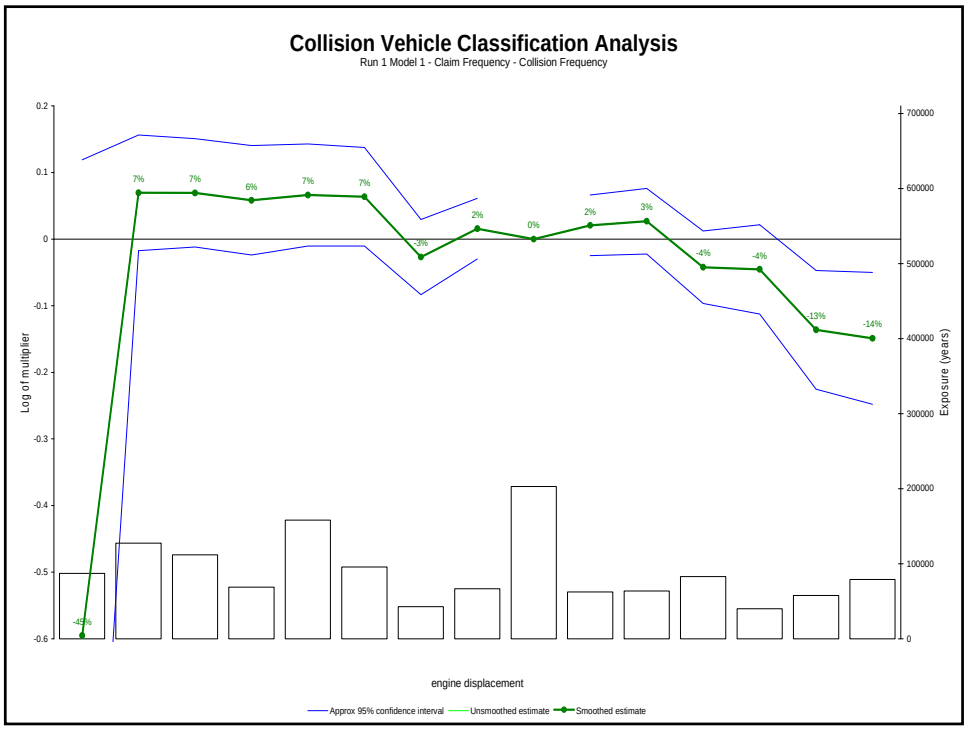
- Begin with descriptive variables related to high dimensional variables
  - ZIP code: geo-demographic data
  - Make/Model: vehicle characteristics
- Use descriptive variables in model development
- The resulting parameter estimates of descriptive variables can then be used in rating or can be grouped into new symbol/territory

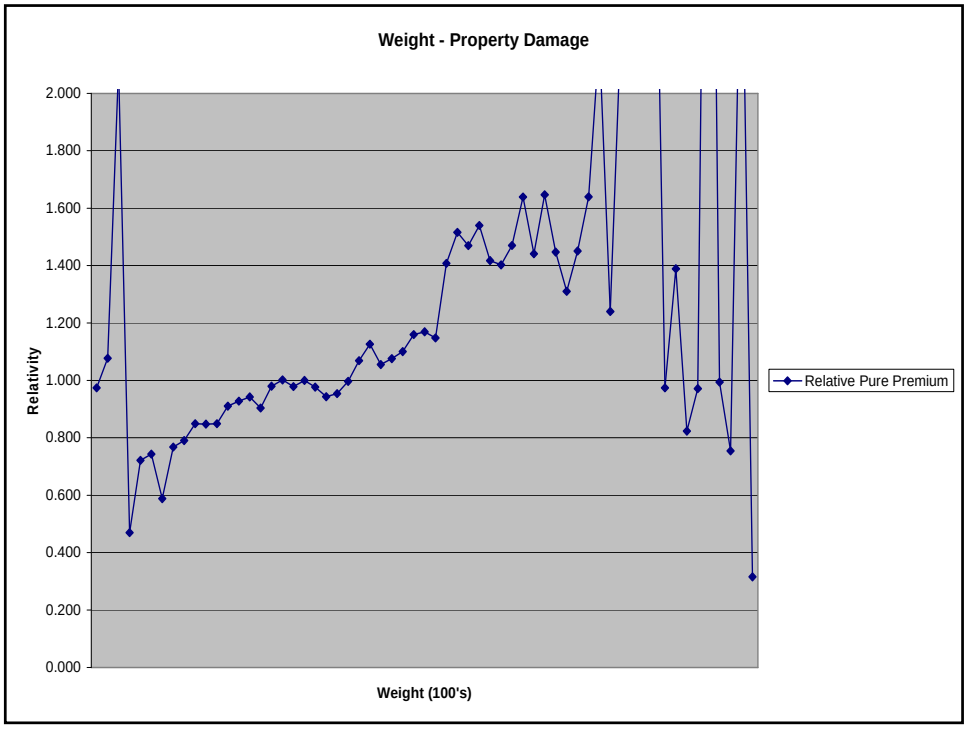
PINNACLE ACTUARIAL RESOURCES, INC.
The Firm of Choice

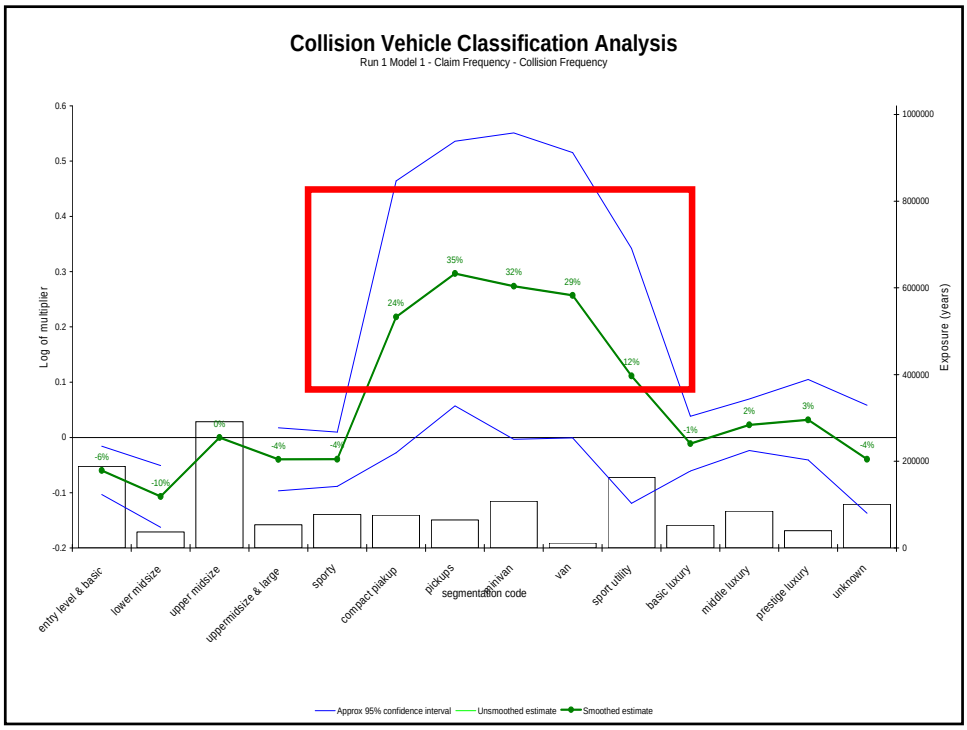
Vehicle Characteristic Data

- **Body**: style, number of doors, weight,
- **Engine**: cylinders, displacement,
- **Performance**: engine displacement, weight, performance, payload capacity
- **Additional**: ABS, anti-theft, base price, ESC
- **Crash test information**

PINNACLE
ACTUARIAL RESOURCES, INC.
The Firm of Choice




Weight - Property Damage
2.000
1.800
1.600
1.400
1.200
1.000
0.800
0.600
0.400
0.200
0.000
Relativity
Weight (100's)
Relative Pure Premium


Collision Vehicle Classification Analysis
Run 1 Model 1 - Claim Frequency - Collision Frequency
Log of multiplier
Exposure (years)
segmentation code
entry level & basic
lower midsize
upper midsize
uppermidsize & large
sporty
compact pickup
pickups
minivan
van
sport utility
basic luxury
middle luxury
prestige luxury
unknown
-6%
-10%
0%
-4%
-4%
24%
35%
32%
29%
12%
-1%
2%
3%
-4%
Approx 95% confidence interval
Unsmoothed estimate
Smoothed estimate

Develop Vehicle Group

- Develop vehicle group
  - Calculate vehicle "relativity" for each risk
  - Group risks of similar loss potential
  - If desired, rename groups for ease of use
  - Append vehicle group to original database
  - Determine final relativities
- Implementation
  - Company specific symbols
  - Enhancement to current symbols
  - Series of additional rating factors

PINNACLE ACTUARIAL RESOURCES, INC. *The Firm of Choice*

Advantages of Using Vehicle Characteristics

- Competitive advantage
- Reflects a company's business
- More efficient use of vehicle data
- Removal of bias due to other characteristics
- Simplifies adjusting for prior years and introduction of new years
- Automatically handles new models
- Works for liability and physical damage

PINNACLE ACTUARIAL RESOURCES, INC. *The Firm of Choice*

Principal Components

- Mathematical transformation of input variables
- Calculated from the correlation matrix of the input variables
- Transforms a number of correlated variables into a smaller number of uncorrelated variables
- First component accounts for as much of variability as possible, second component accounts for as much of remaining variability as possible, etc.
- Can be used as a dimension reduction technique, creates a summarized version of the inputs to use in predictive models

PINNACLE ACTUARIAL RESOURCES, INC. *The Firm of Choice*

Principal Components

BL5

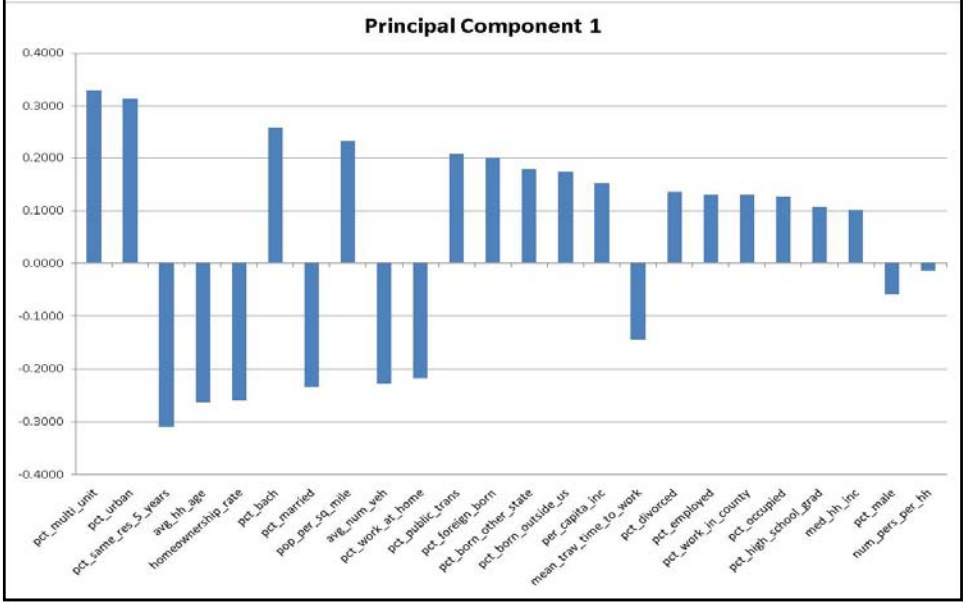
- First j principal components provide least squares solution to Y = XB + E
  - Y = n x p matrix of centered observed variables
  - X = n x j matrix of scores on first j principal components
  - B = j x p matrix of eigenvectors
- Eigenvectors are orthogonal
- Principal component scores are jointly uncorrelated

PINNACLE ACTUARIAL RESOURCES, INC. *The Firm of Choice*

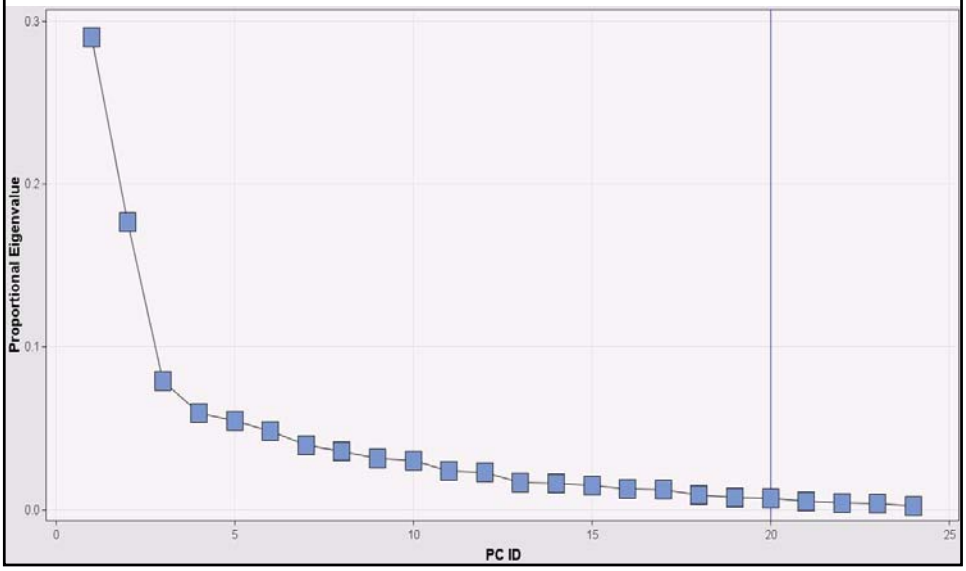
Slide 26

BL5 I'm starting to fall how the sled with the rest of these.
Boison, LeRoy, 3/12/2010

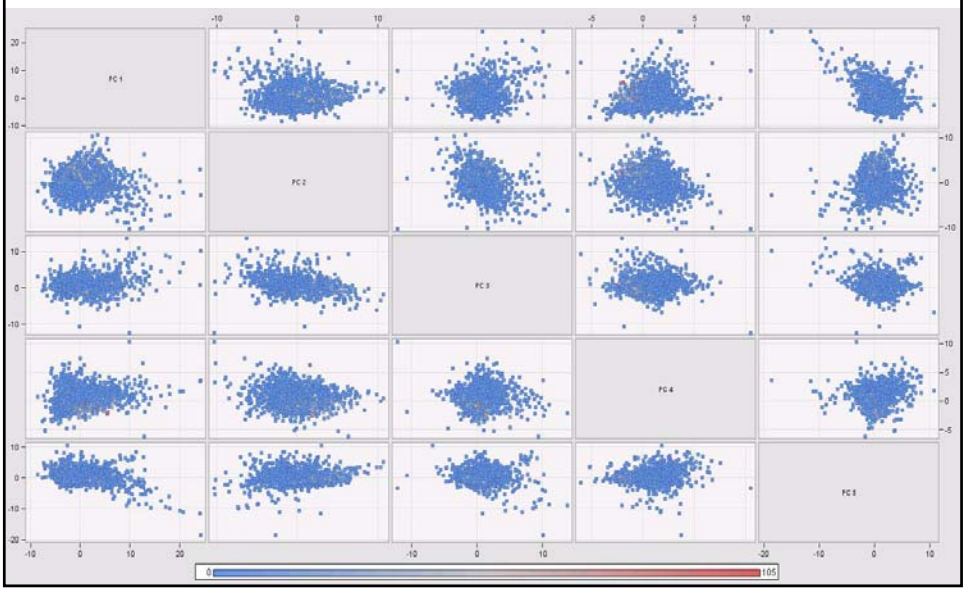
Principal Component 1 Eigenvectors



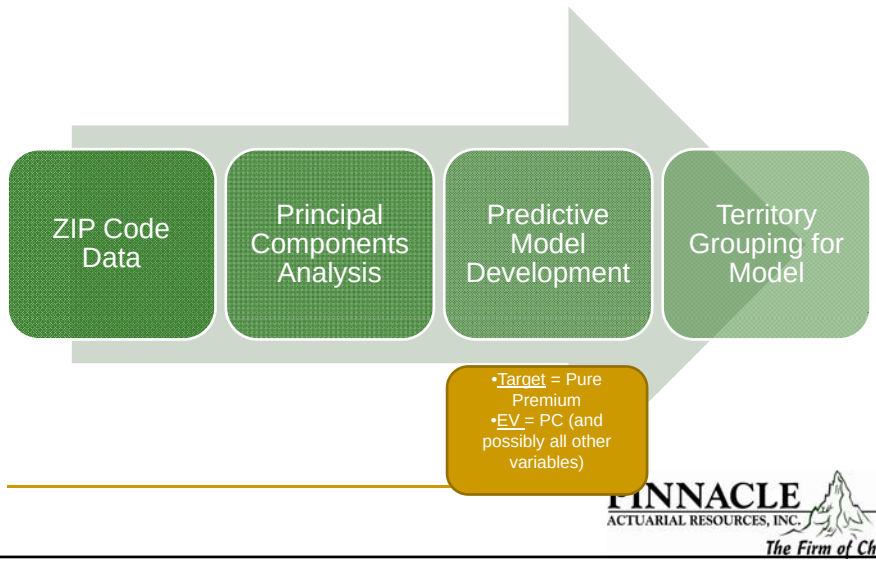
Correlation Matrix Proportional Eigenvalue Plot



Principal Components Matrix



Application of Principal Components

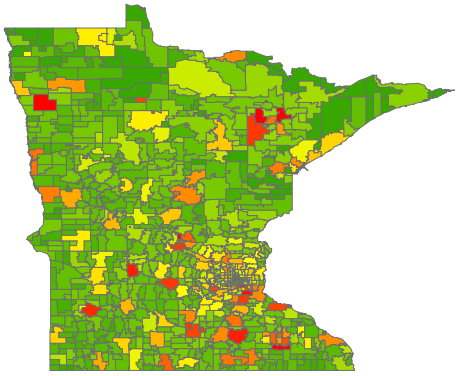


PINNACLE ACTUARIAL RESOURCES, INC.
The Firm of Choice

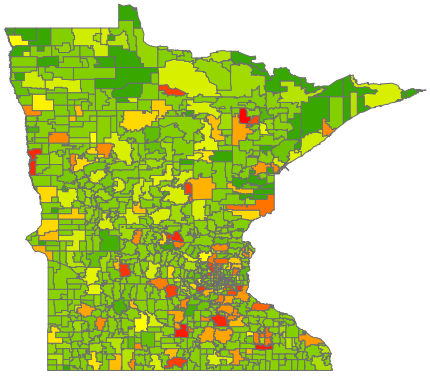
Comparison of Results

PINNACLE ACTUARIAL RESOURCES, INC. The Firm of Choice

GLM



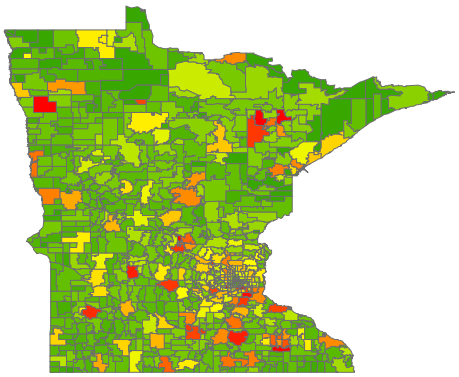
Principal Components

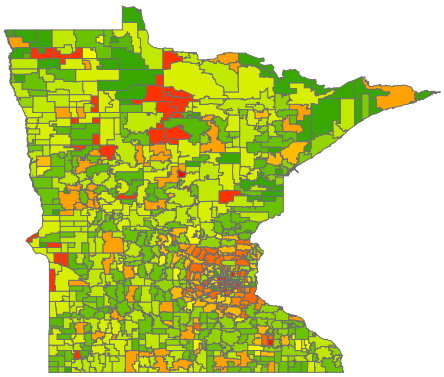


PINNACLE ACTUARIAL RESOURCES, INC. The Firm of Choice

GLM

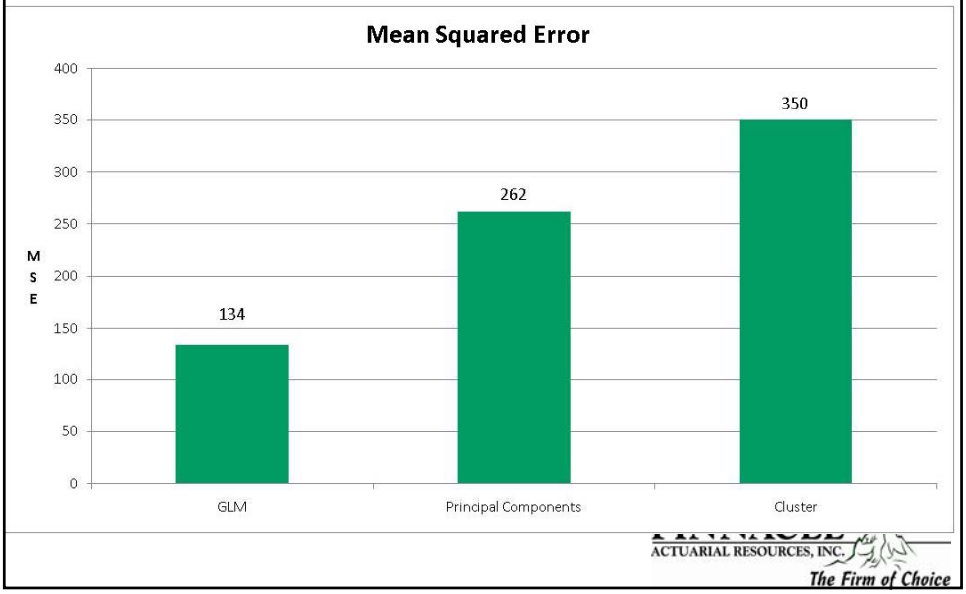
Clustering





PINNACLE
ACTUARIAL RESOURCES, INC.
The Firm of Choice

Mean Squared Error Comparison



PINNACLE
ACTUARIAL RESOURCES, INC.
The Firm of Choice

Conclusions

- Using input variable information directly is generally preferable when building predictive models
- There are many cases when this is not feasible
  - Unknown target
  - Input variable with too many levels
  - Too many input variables
- Techniques for handling high dimensional variables still result in models that produce predictive results

PINNACLE
ACTUARIAL RESOURCES, INC.
The Firm of Choice