Using Cluster Analysis to Define Geographical Rating Territories

2008 CAS Spring Meeting

Discussion Paper Program

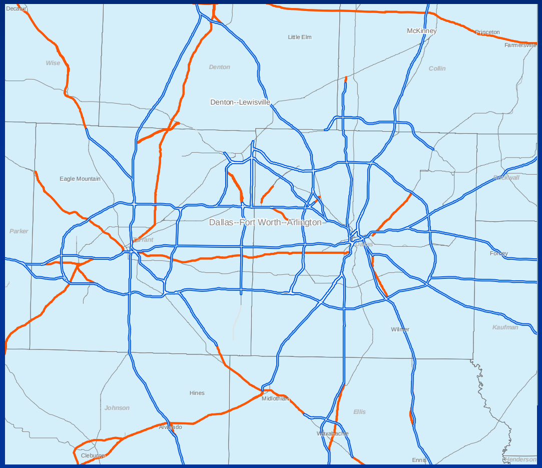
Philip J. Jennings, FCAS, MAAA

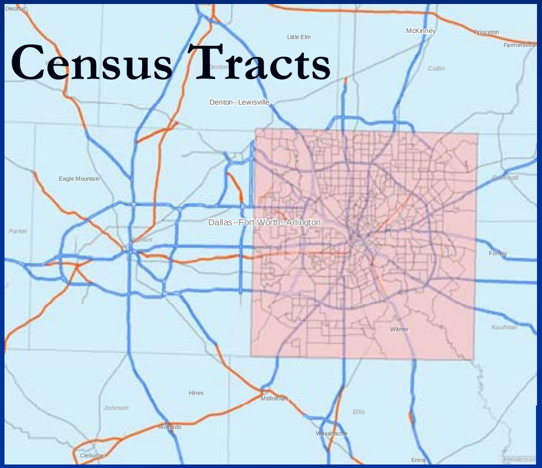
Outline of Presentation

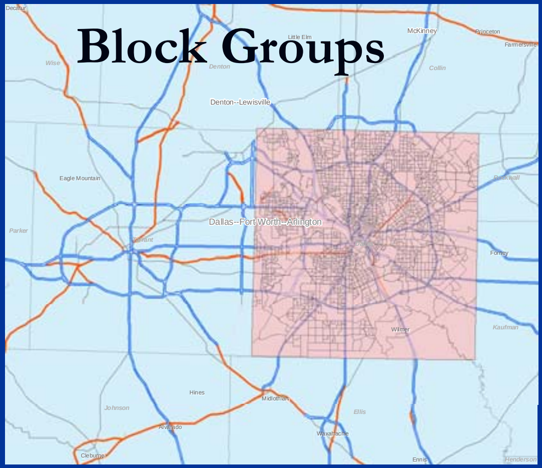
- Building Blocks
- Data
- Variables to Cluster On
- Credibility and it's Complement
- Clustering Method
- Implementation Issues
- Final Results

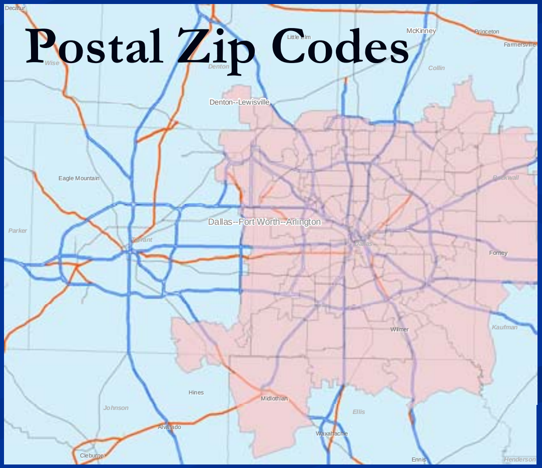
Building Blocks – County Level

Building Blocks

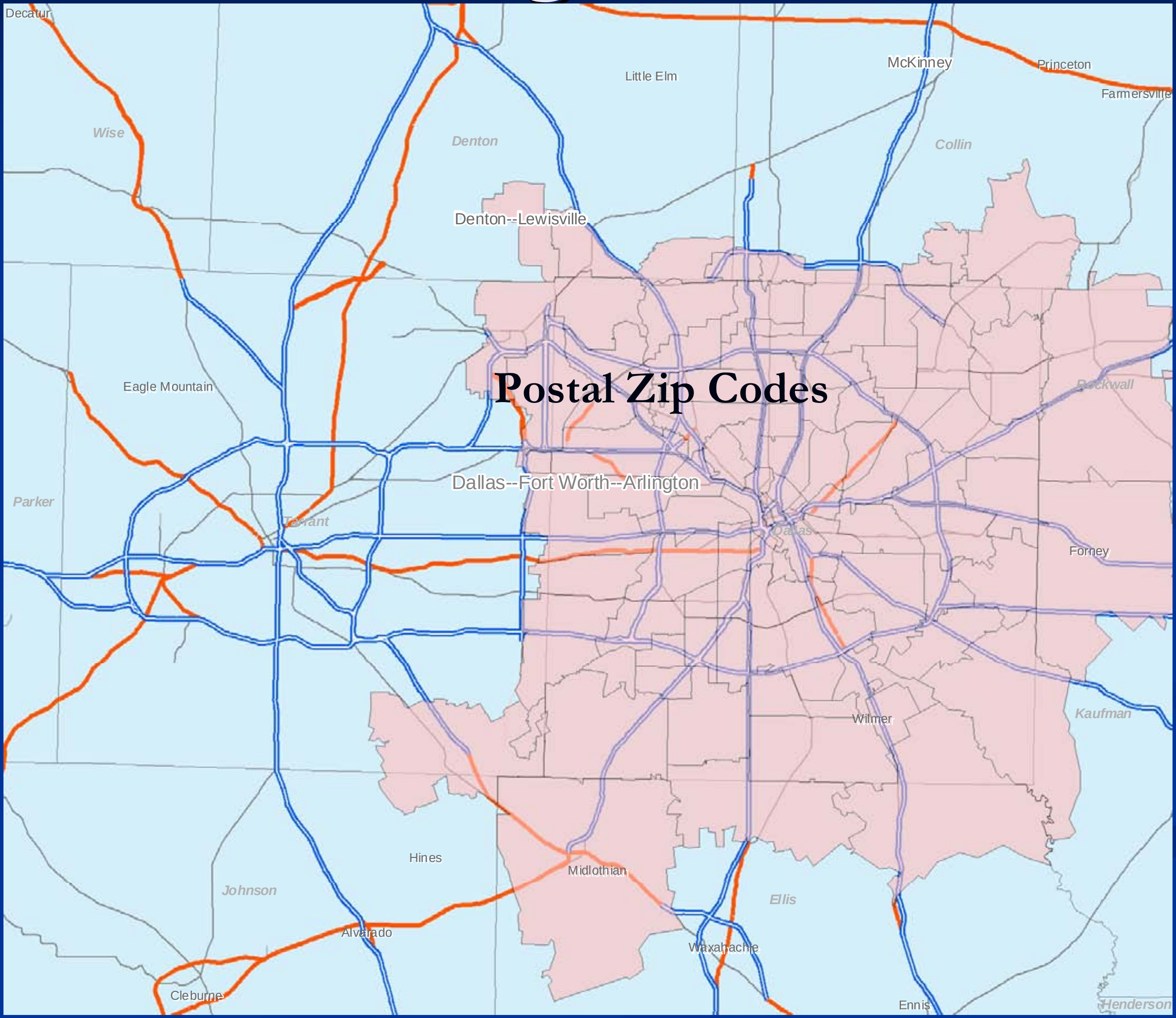








Building Blocks



Building Blocks Should Be…

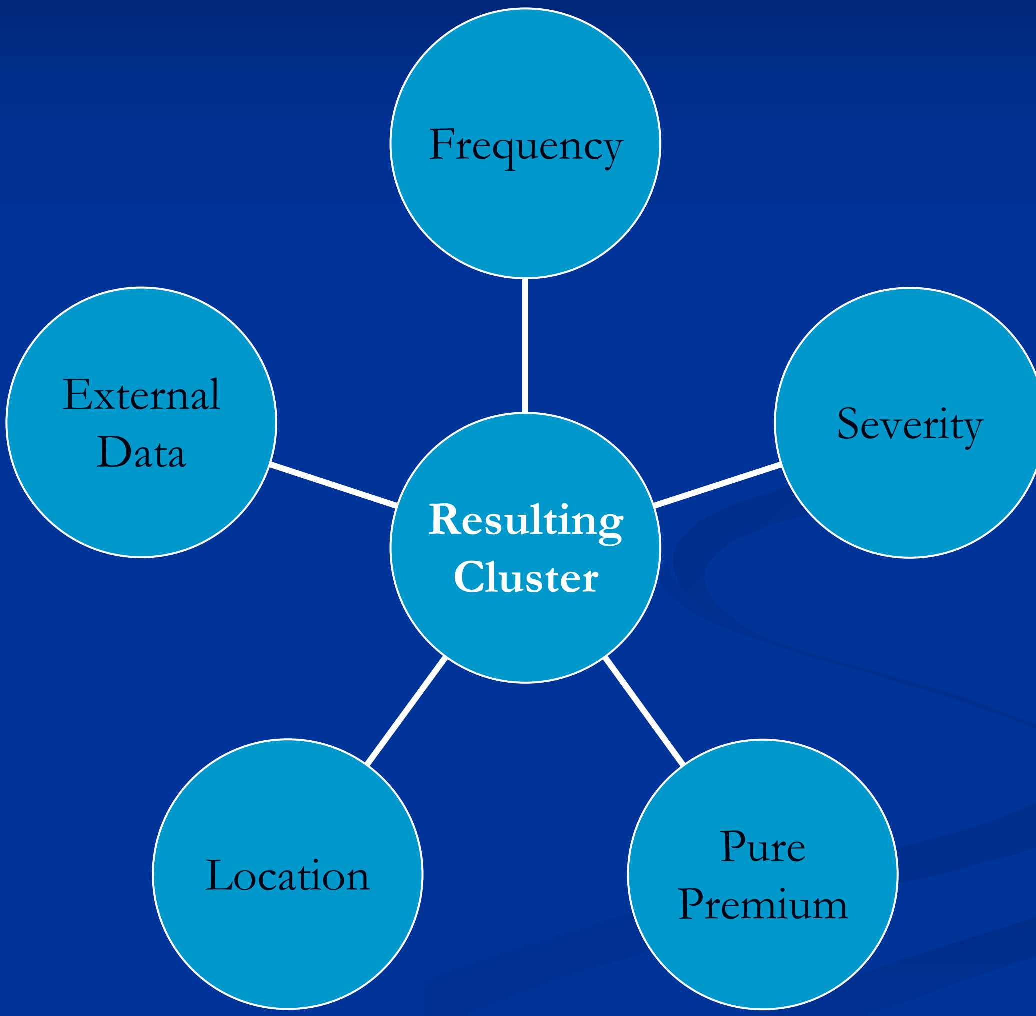
- Small enough to be homogeneous with respect to geographic risk.
- Large enough to produce credible results.
- Collected loss and premium data should be easily assigned.
- Competitive and/or external data can be easily mapped to the geographical unit.
- Easy for the insured and company personnel to understand.
- Politically acceptable.
- Verifiable.
- Stable over time.

Werner, Geoffrey, FCAS, "The United States Postal Service's New Role: Territorial Ratemaking", *Casualty Actuarial Society Forum*, 1999, Winter, 287-308

Data

- Internal Company Data
  - Exposures, Premium, Losses, Claim Counts
  - Losses developed and trended to the average settlement date
  - Liability losses capped at a predetermined amount
  - May need to clean up messy data
- External Data
  - Anything that can be geo-referenced to your building block level

Variables to Cluster On



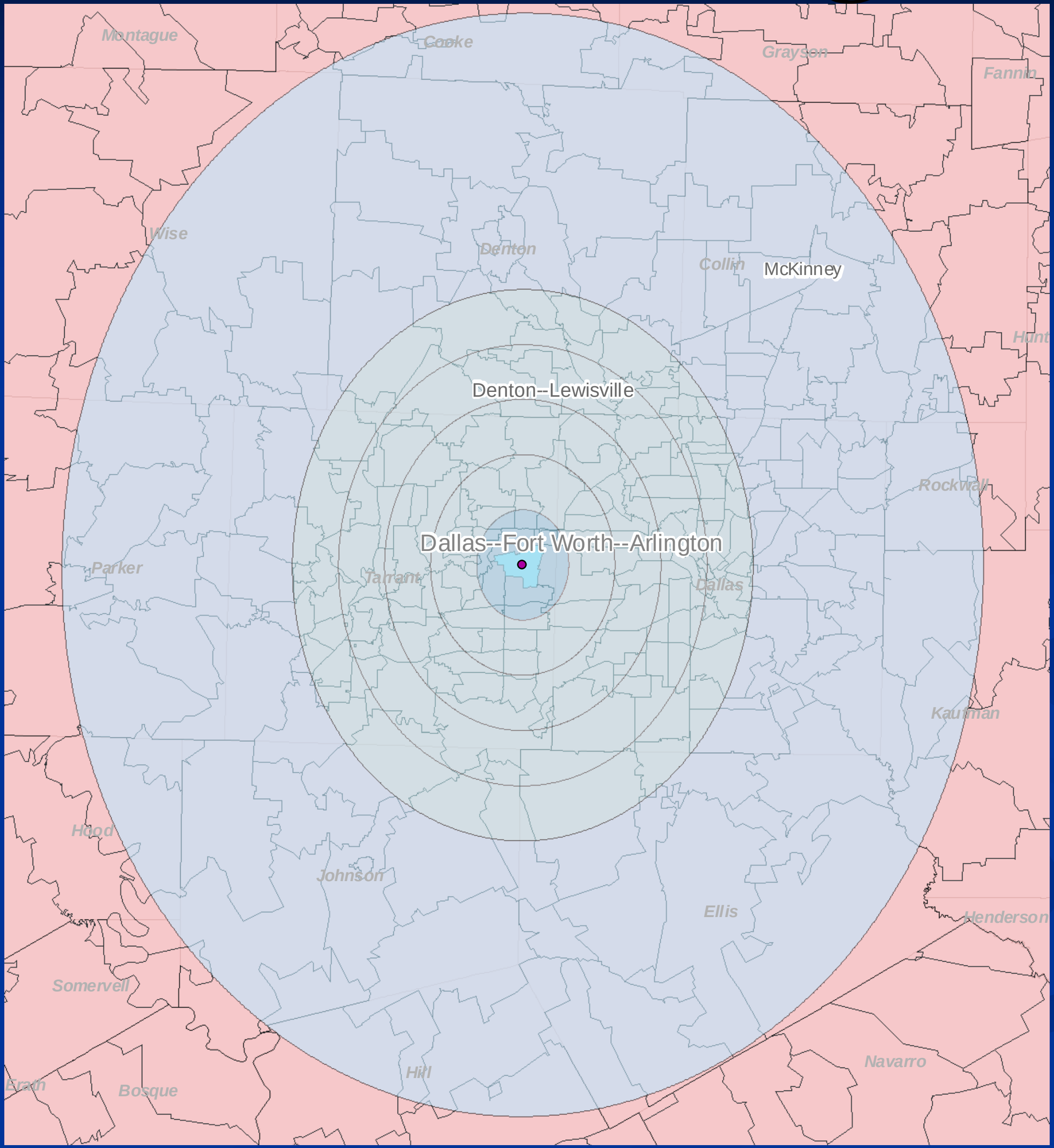
Deriving the Clustering Variables

- For each building block (zip code) calculate
  - Pure Premium = Incurred Losses / Earned Exposures
  - Frequency = Incurred Claims / Earned Exposures

Deriving the Clustering Variables

- For each building block (zip code) create concentric rings around zip centroid
  - 5, 10, 15, 20, 25, and 50 mile rings to get groupings of local zip codes
  - Aggregate Premium, Losses, Claims, and Exposures for each grouping
  - Calculate the Pure Premium and Frequency for each grouping

Concentric Rings



Assigning Credibility

- For the pure premiums I used

$$Z = P / (P + K)$$

Where P = Earned Premium and K=2,500,000

- For the frequencies I used

$$Z = \sqrt{(n / n_f)}$$

Where n = incurred claim count and n_f = 1,082

Assigning Credibility

- Credibility for the concentric ring groupings of zip codes
  - $Z_5 = Z_{5Actual} - Z_{zip}$
  - $Z_{10} = Z_{10Actual} - Z_5 - Z_{zip}$
  - $Z_{15} = Z_{15Actual} - Z_{10} - Z_5 - Z_{zip}$
  - Similar calculations for Z_{20}, Z_{25}, and Z_{50}

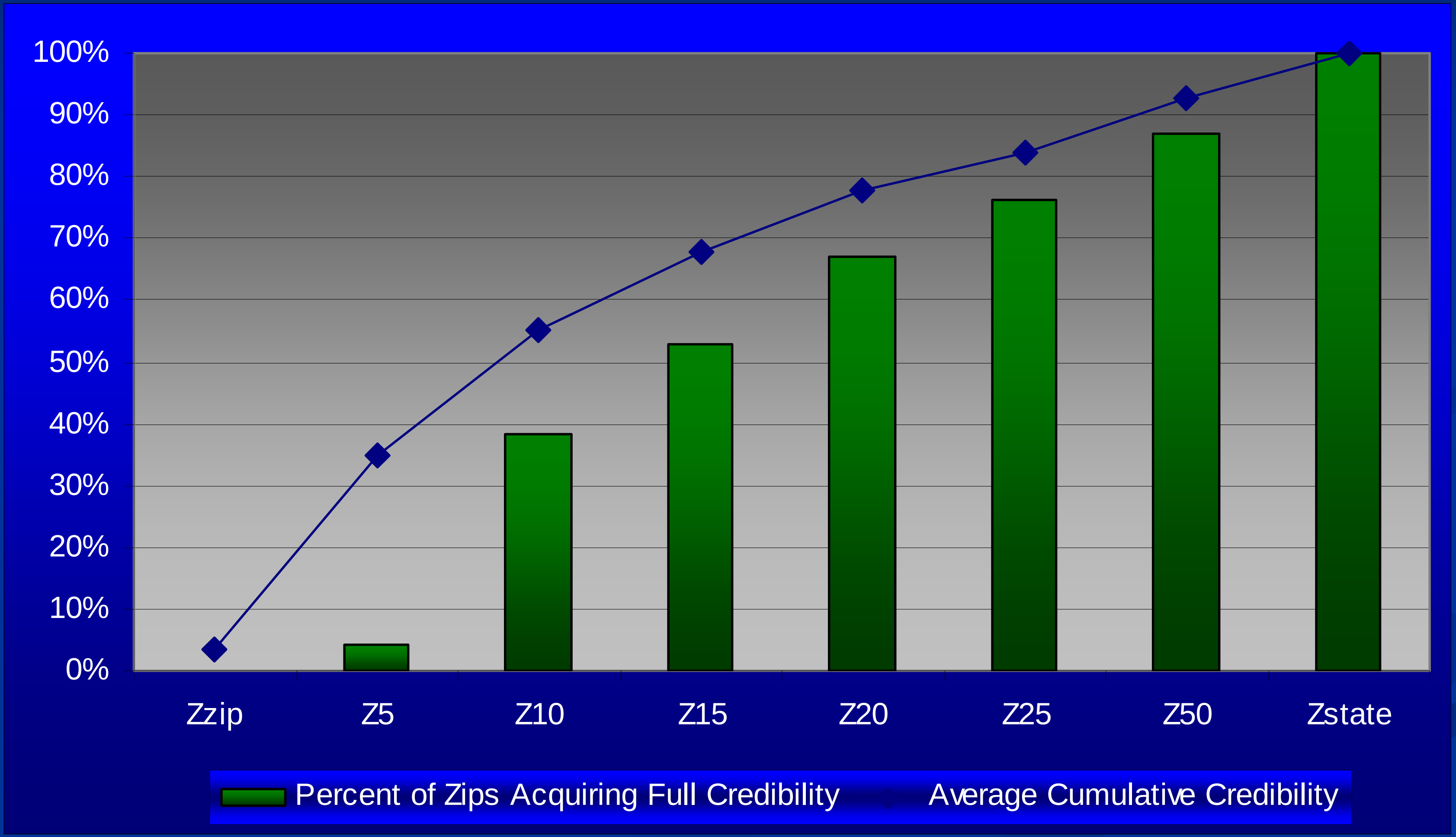
Deriving the Clustering Variables

- For each zip code calculate a credibility weighted pure premium and frequency

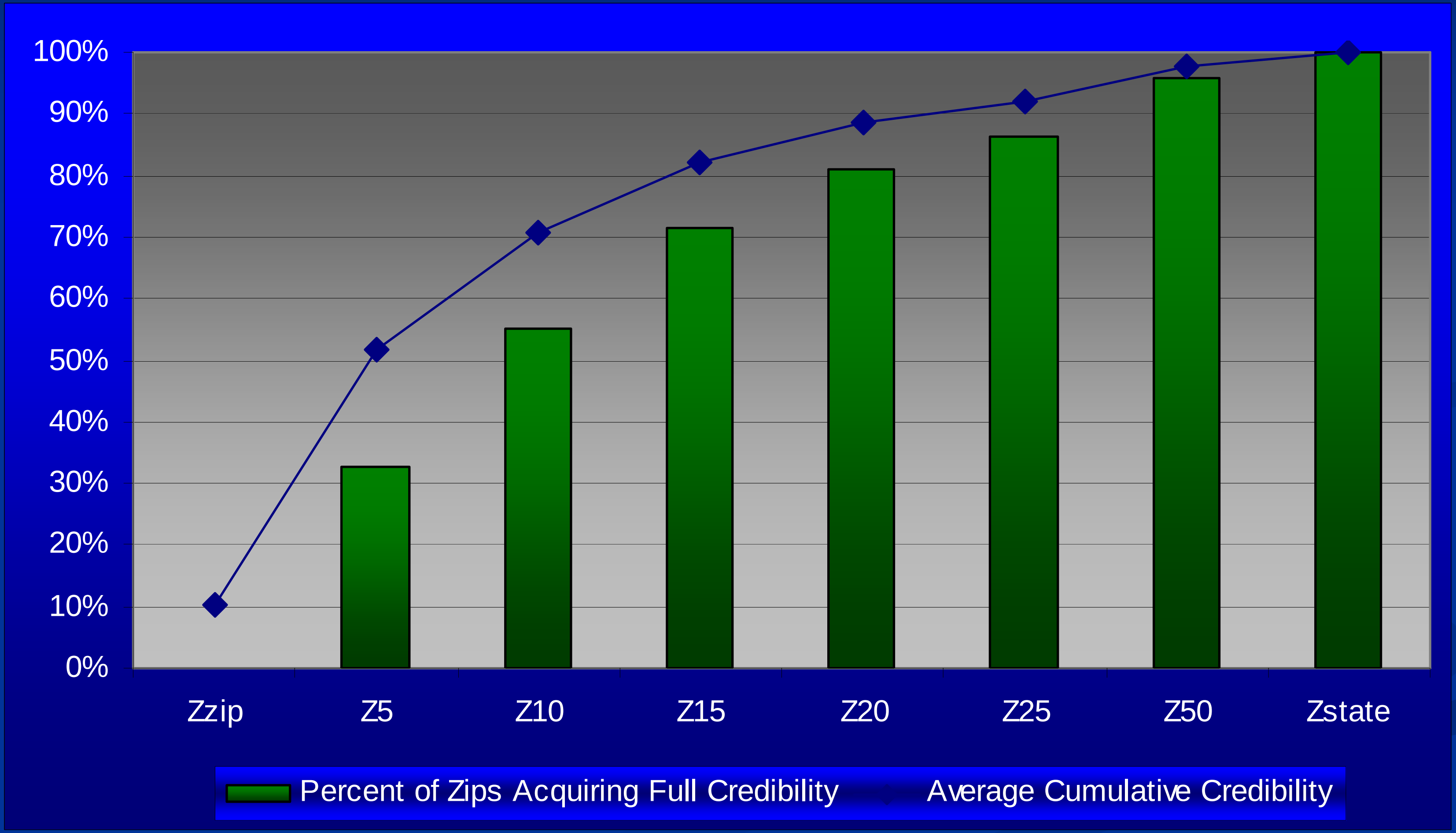
- $CWPP = PP_{zip} * Z_{zip} + PP_5 * Z_5 + PP_{10} * Z_{10} + PP_{15} * Z_{15} + PP_{20} * Z_{20} + PP_{25} * Z_{25} + PP_{50} * Z_{50} + (1 - Z_5 - Z_{10} - Z_{15} - Z_{20} - Z_{25} - Z_{50}) * PP_{State}$

- Similar Calculation for Frequencies

Credibility – Pure Premiums



Credibility – Frequency



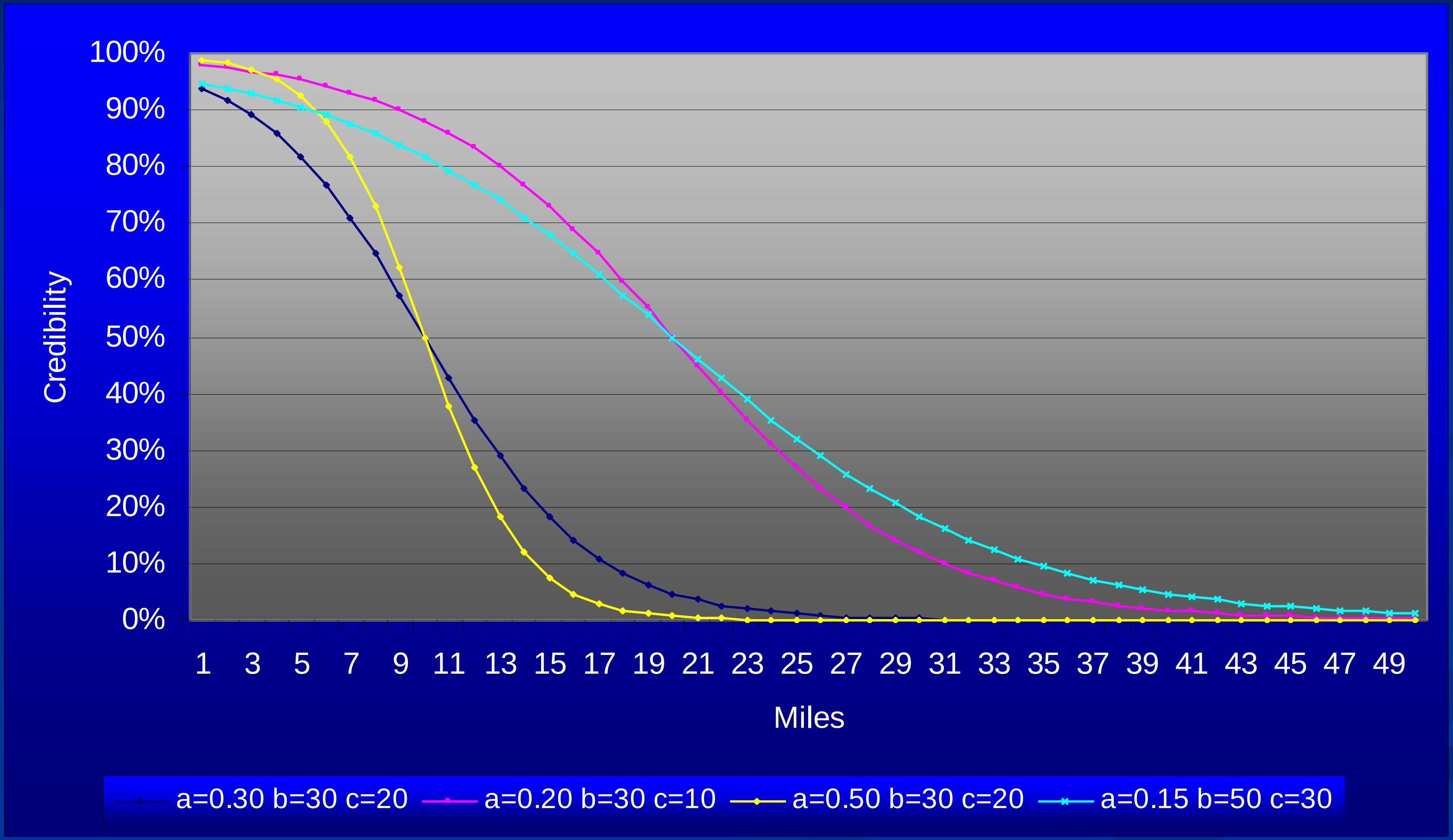
Alternative Choices for the Complement of Credibility

- Population Density Groups
- Vehicle Density
- Accidents Per Registered Vehicle
- Injuries Per Accident
- Thefts Per Vehicle
- Medical Cost Index

Additional Considerations

- Concentric Rings
  - Zip Codes 50 Miles away may not represent the same geographical risk
  - Zip Codes along a state's border or coastline
- Analysis By Coverage or All Coverages Combined
  - Should your complement of credibility and/or variable selection vary by coverage

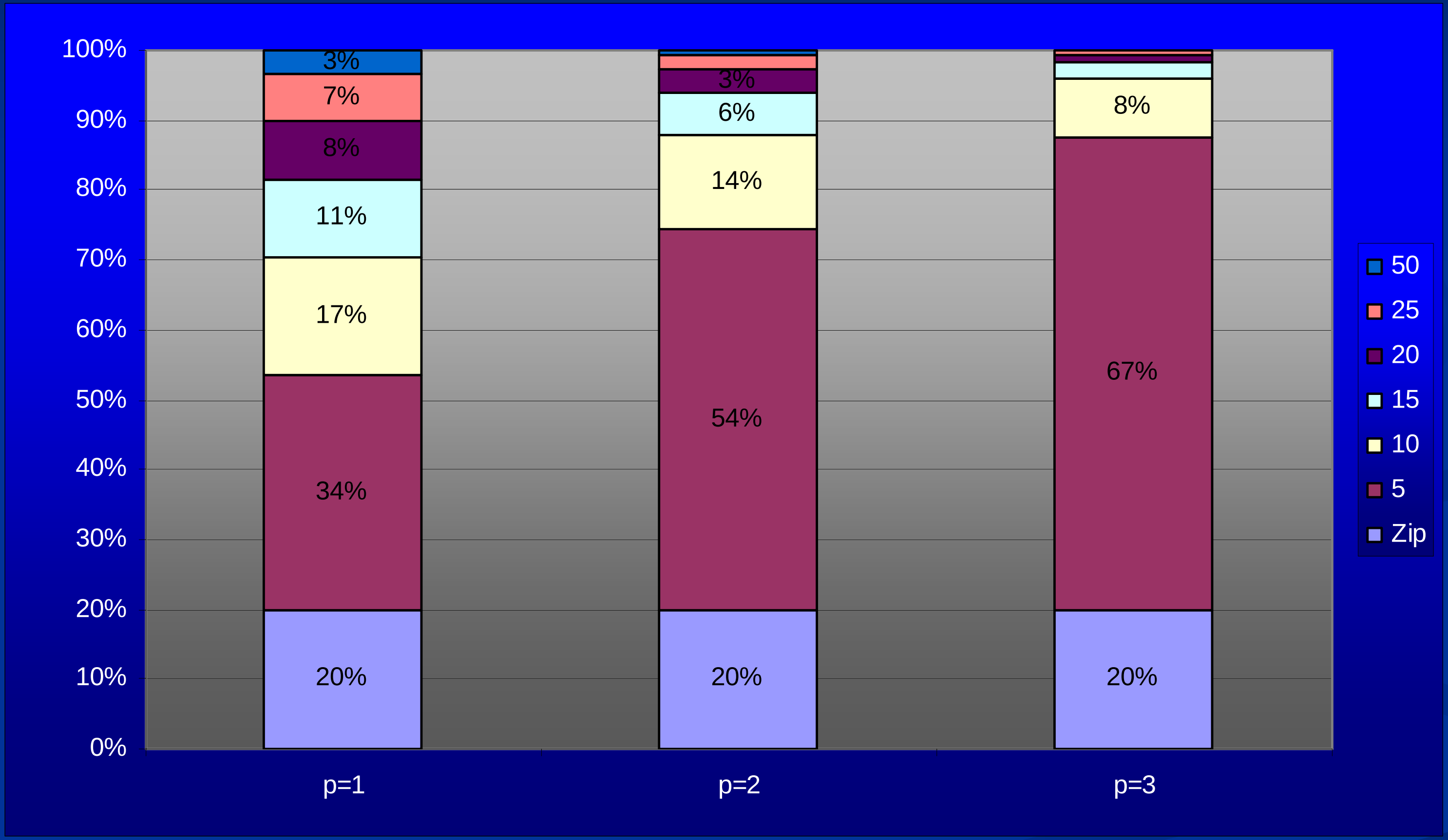
Inverse Distance Weighting - Alternatives



$$Y = 1 / (1+ \exp(-a(b-x-c))$$

Sigmoid Curve - Miller

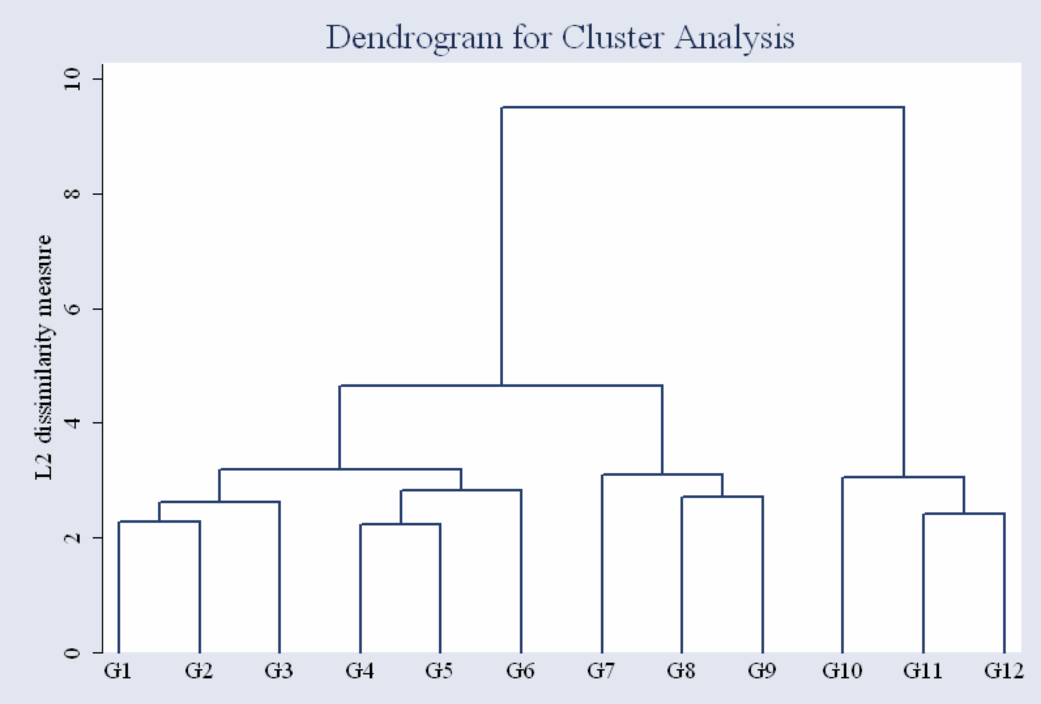
Inverse Distance Weighting - Alternatives



$$CWPP = \Sigma \lambda_i PP_{d_i} \qquad \lambda_i = d_{i0}^{-p} / \Sigma d_{i0}^{-p}$$

Clustering Methods

- Heirarchical – algorithms that find successive clusters using previously established clusters
  - Agglomerative – “bottom-up”
  - Divisive – “top-down”



Clustering Methods

- Partition – algorithms that separate the observations into mutually exclusive groups
  - k-means
    - Begin with k centers or means
    - Each observation is assigned to the group whose mean is closest to that observation's mean.
    - New group means are calculated.
    - Repeat until no observations change groups.
  - k-medians

Distance Measures for Continuous Data

- General Form – L_N Norm

$$\left\{\sum_{m=1}^{p} |X_{mi} - X_{mj}|^N\right\}^{1/N}$$

For observation i and centroid j using p variables

  - When N=1 this is known as Absolute, Cityblock, or Manhattan Distance
  - When N=2 this is Euclidean Distance
  - Linfinity = $\max_{m=1,\ldots,p} |X_{mi} - X_{mj}|$

More Similarity Measures for Continuous Data

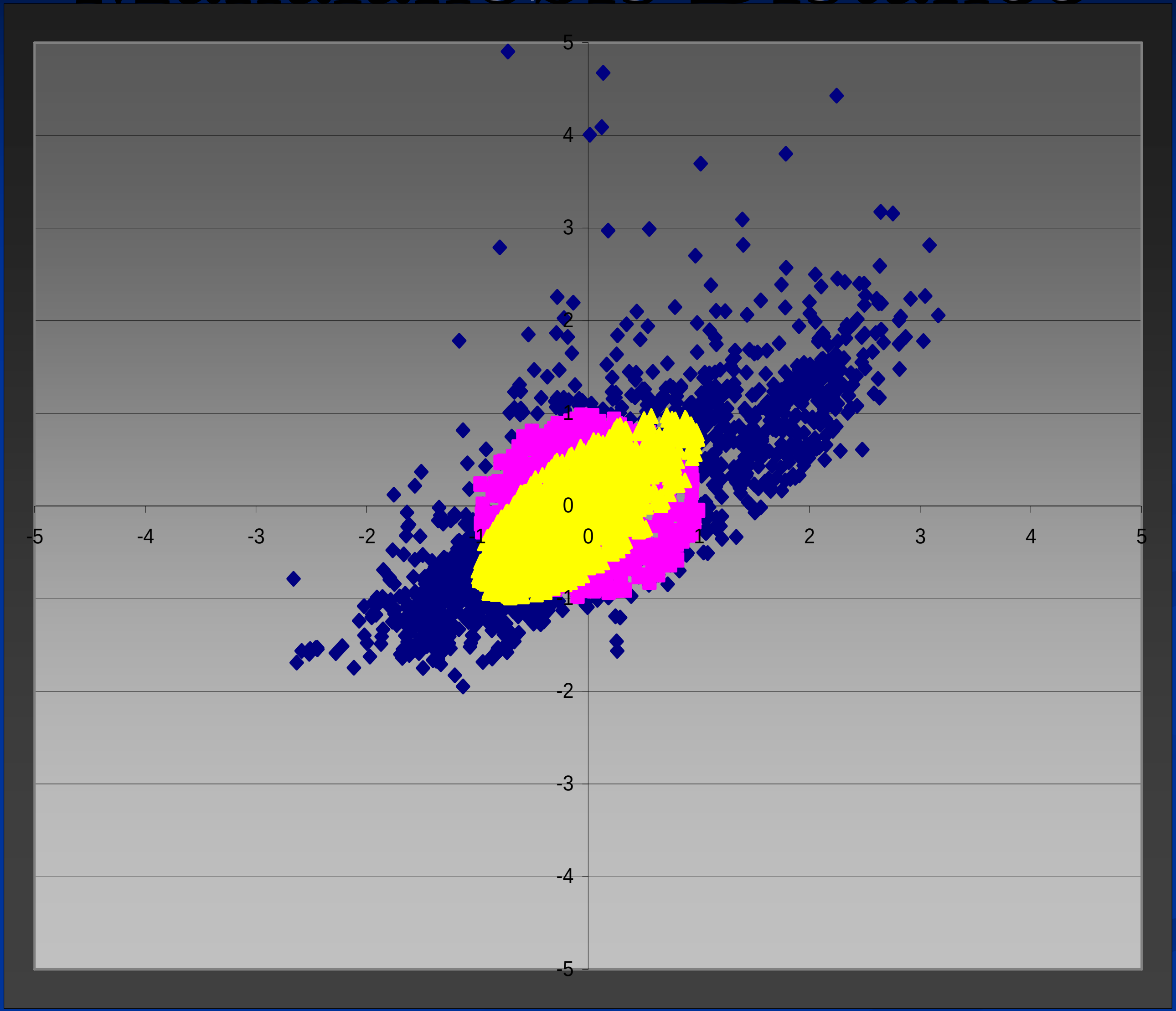
Canberra

$$\sum_{m=1}^{p} \frac{|X_{mi} - X_{mj}|}{|X_{mi}| - |X_{mj}|}$$

Correlation

$$\frac{\Sigma\,(X_{mi} - \overline{X}_{\cdot i})\,(X_{mj} - \overline{X}_{\cdot j})}{\{\Sigma\,(X_{mi} - X_{\cdot i})^2\,\Sigma(X_{nj} - X_{\cdot j})^2\}^{1/2}}$$

Mahalanobis Distance



$$D^2 = (\mathbf{x} - \mathbf{m})^T C^{-1} (\mathbf{x} - \mathbf{m})$$

Variable Standardization

- The distance metric is summed over the p variables
- If a variable has a significantly wider range it will dominate the cluster

Standardize or transform the variables

Some distance measures require non-negative input

- May wish to leverage the influence of certain variables

Variable Standardization
Some Alternatives

	Standardization	Mean	Standard Deviation
1.	$\frac{X - \bar{X}}{s}$	0	1
2.	$\frac{X}{s}$	$\frac{\bar{X}}{s}$	1
3.	$\frac{X}{Max(X)}$	$\frac{\bar{X}}{Max(X)}$	$\frac{s}{Max(X)}$
4.	$\frac{X}{Max(X)-Min(X)}$	$\frac{\bar{X}}{Max(X)-Min(X)}$	$\frac{s}{Max(X)-Min(X)}$
5.	$\frac{X - Min(X)}{Max(X)-Min(X)}$	$\frac{\bar{X} - Min(X)}{Max(X)-Min(X)}$	$\frac{s}{Max(X)-Min(X)}$
6.	$\frac{X}{\Sigma X}$	$\frac{1}{n}$	$\left\{ \frac{1}{n} \left[\frac{\Sigma X^2}{(\Sigma X)^2} - \frac{1}{n} \right] \right\}^{1/2}$
7.	Rank(X)	$\frac{n+1}{2}$	$\left\{ (n+1) \left[\frac{(2n+1)}{6} - \frac{(n+1)}{4} \right] \right\}^{1/2}$

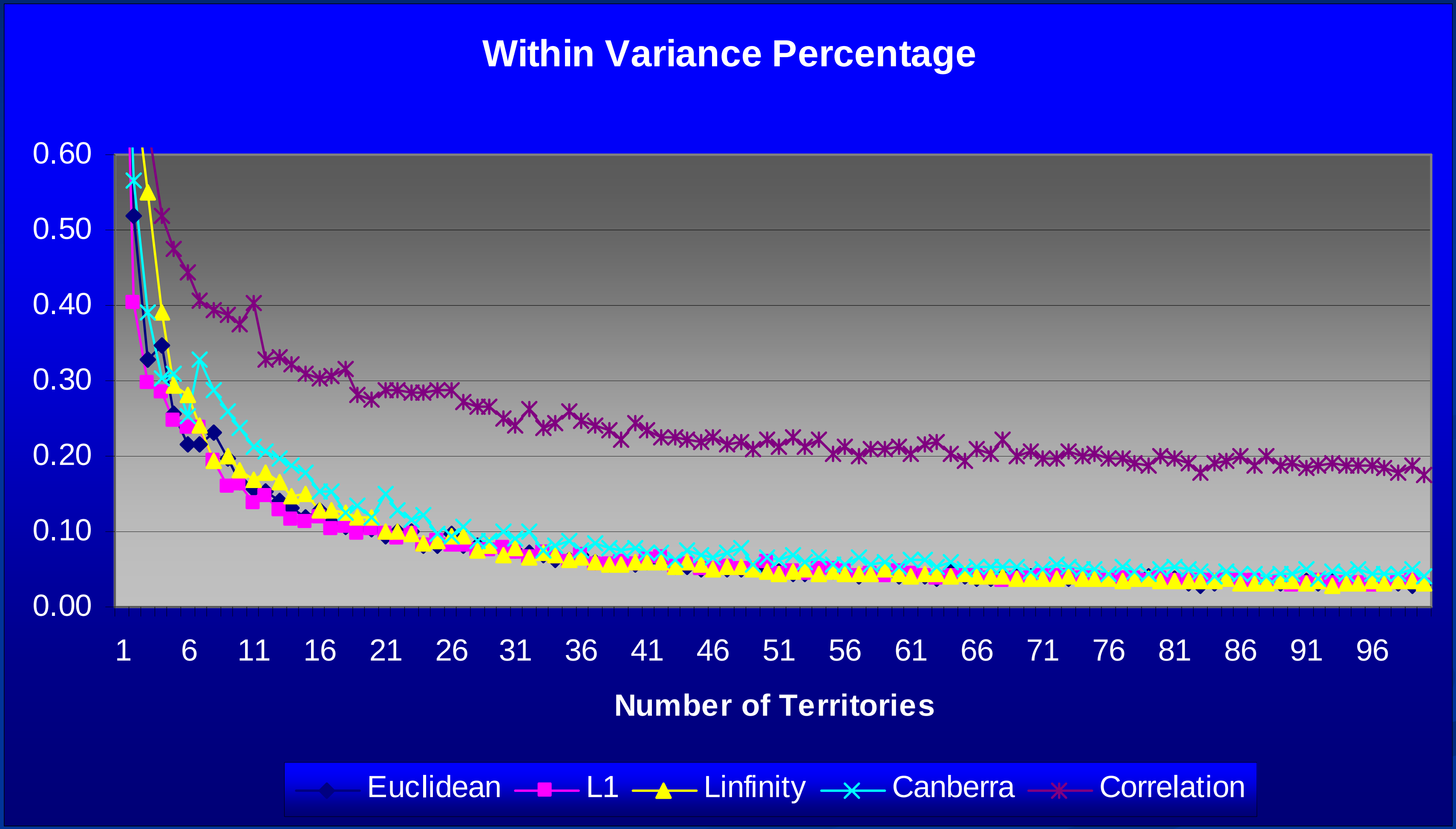
Starting Values

- The starting values, or initial centers, can affect the resulting clusters.
- Choices for starting values
  - Random (with optional seed)
  - First k or last k observations
  - Means of k random partitions
  - Assign observation 1, 1+k, 1+2k… to group 1 etc..
  - Group on a variable to form k groups and use these means
  - First N/k obs for first group, second N/k for second group etc.. Use the means of these groups as starting values.

General Methodology

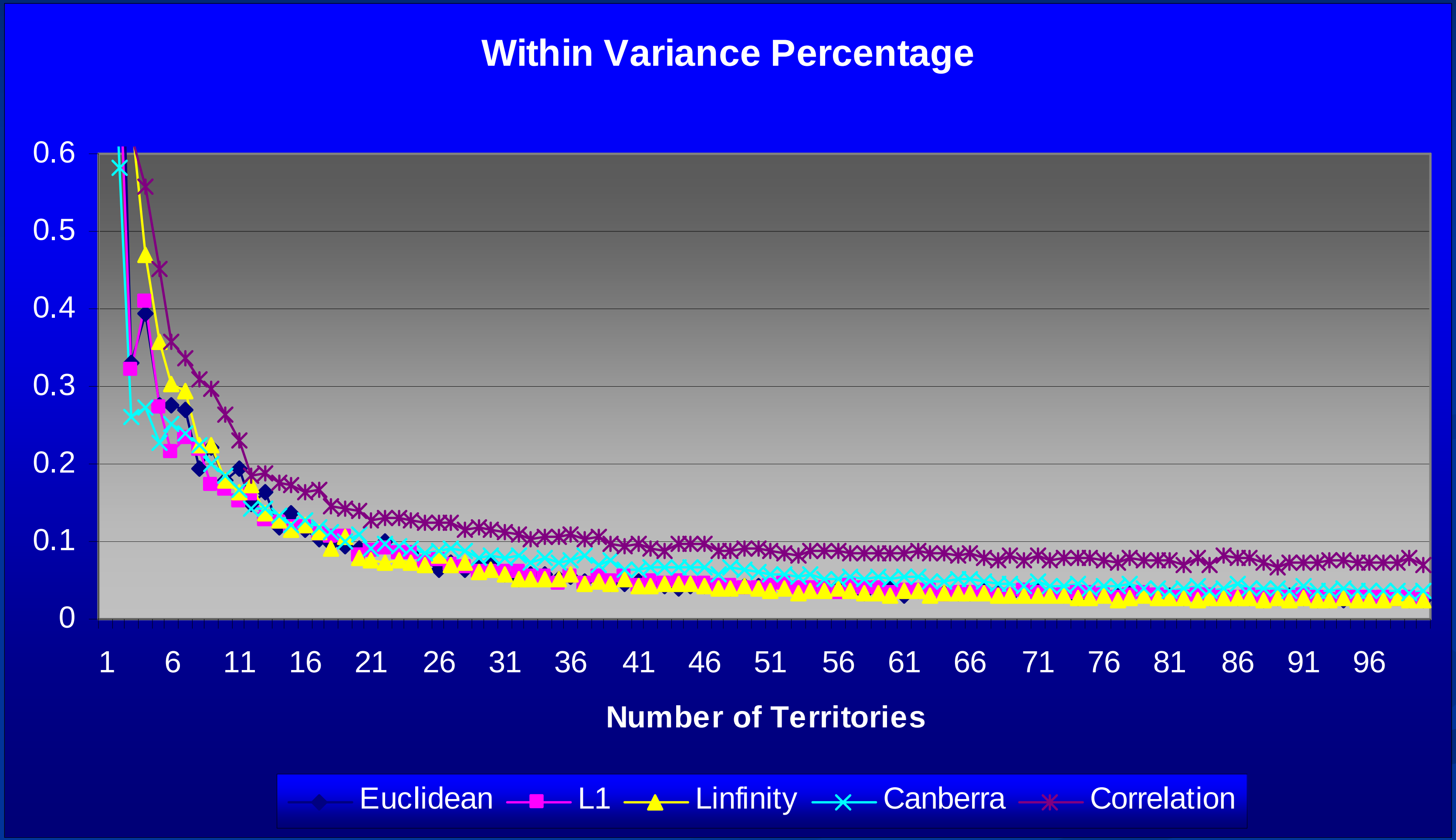
- Standardize Variables
- For k = 2 to 100
  - Create k clusters based on pure premium, frequency, latitude, and longitude
  - Calculate within variance percentage
  - Store cluster assignment and WVP for k
- Next k
- Analyze pattern of WVP and map of clusters

Results From Various Distance Metrics



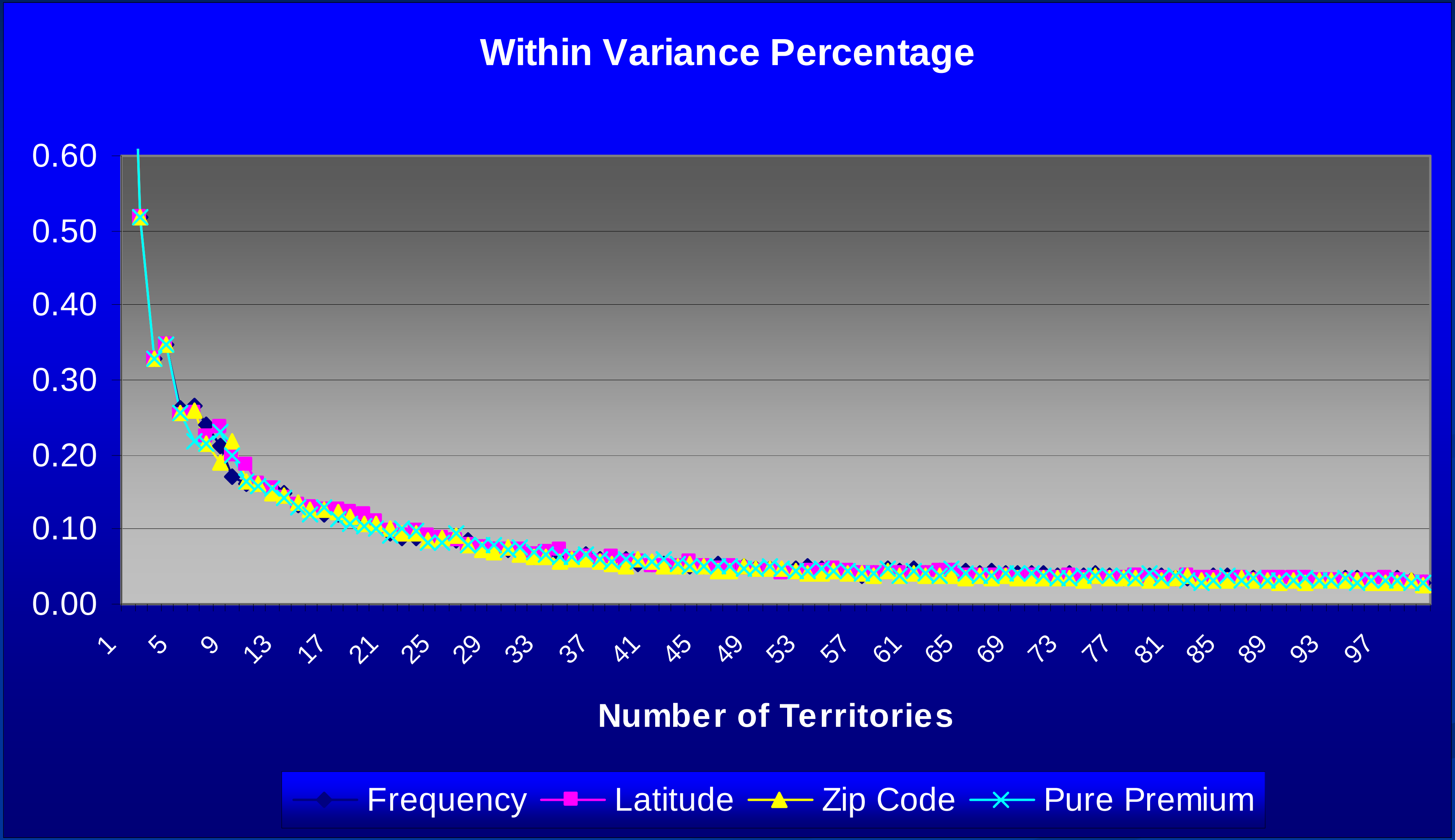
k-means, standardization 1 using pure premium, frequency, latitude, and longitude with k segments as starting values after sorting by pure premium

Results From Various Distance Metrics



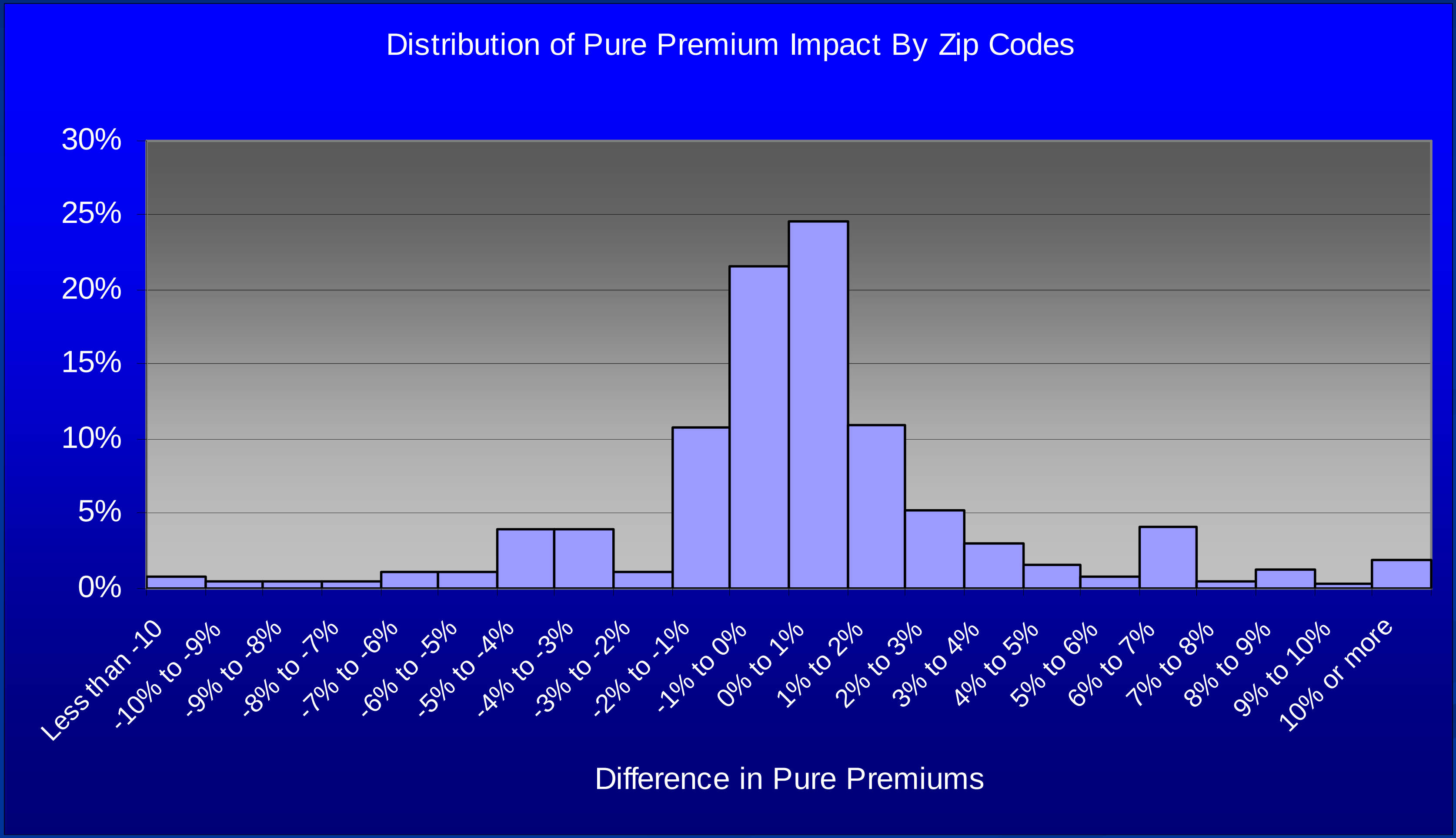
k-means, standardization 5 using pure premium, frequency, latitude, and longitude with k segments as starting values after sorting by pure premium

Sensitivity to Starting Values



Sorted first by given variable to establish different starting values.

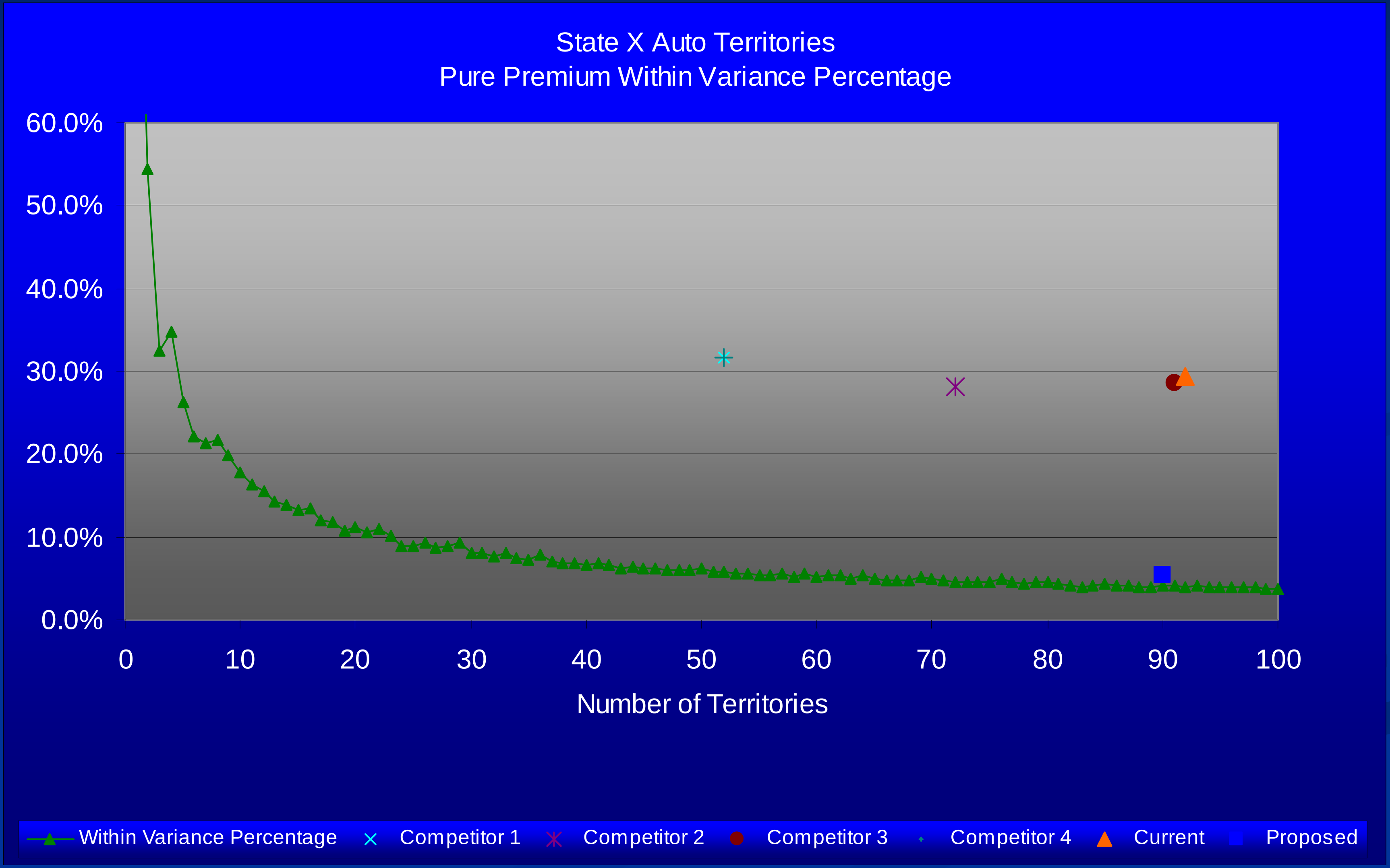
Sensitivity to Starting Values
Pure Premium Impact



Implementation Issues

- Rate disruption
- Sales force acceptance
- Data availablility

Final Results



References

- Brubaker, Randall E., “Geographic Rating of Individual Risk Transfer Costs Without Territorial Boundaries,” *Casualty Actuarial Society Forum*, 1996, Winter, 97-127.
- Christopherson, Steven, and Debra L. Werland, “Using a Geographic Information System to Identify Territory Boundaries,” *Casualty Actuarial Society Forum*, 1996, Winter, 191-211.
- Kaufman, L. and P.J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis* (Hoboken, New Jersey: John Wiley & Sons, 1990).
- Milligan, Glenn W. and Martha C. Cooper, “A Study of Standardization of Variables in Cluster Analysis,” *Journal of Classification*, , 1988, v5, 181-204.
- Miller, Michael J., “Determination of Geographical Territories,” Presented at the 2004 CAS Ratemaking Seminar.
- *Stata 8 Cluster Analysis Reference Manual*, (College Station, TX: StataCorp, 2003), 5. (Parts reprinted by permission of the publisher.)
- Werner, Geoffrey, “The United States Postal Service’s New Role: Territorial Ratemaking,” *Casualty Actuarial Society Forum*, 1999, Winter, 287-308.