EMB America

Multivariate Analysis in Territorial Ratemaking

Presented by:
Serhat Guven, FCAS, MAAA

CAS 2004 Fall Meeting
November 15, 2004

Agenda

- Background
- Traditional Methods
- Multivariate Methods
- Concerns and Issues
- Results

Territorial Ratemaking

Background

Principles of Territory Ratemaking

- Three Primary Purposes of Risk Classification
  - Protection of Program's Financial Soundness
  - Multidimensional Solutions
  - Economic Incentives
- Statistical Considerations of Risk Classification:
  - Homogeneity
  - Incorporating the Location Dimension
  - Predictive Stability
- Operational Consideration of Territory Ratemaking
  - Avoidance of Extreme Discontinuities
  - Related to Principle of Locality

Territorial Ratemaking

- Background
- Traditional Methods
- Multivariate Methods
- Concerns and Issues
- Results

Traditional Methods

- Indications in the Loss Ratio environment
  - Indication $_{(CZP)}$ = f (Losses$_{(CZP)}$/Premium$_{(CZP)}$)
  - f() is a spatially smoothing technique based on Latitude, Longitude coordinate of the countyzip centroid
- Concerns
  - Possible Distributional Biases
  - Volatile Results
  - Allocates unsystematic risk to the territory variable

Multivariate Methods: Background

Response Variable = Systematic Component + Random Component

- Frequency
- Severity

Signal: Function of the Rating Factors/Predictors

Noise: Reflect stochastic process
Loss Ratios ignore the random piece

= Function of Independent Predictors + Random Component

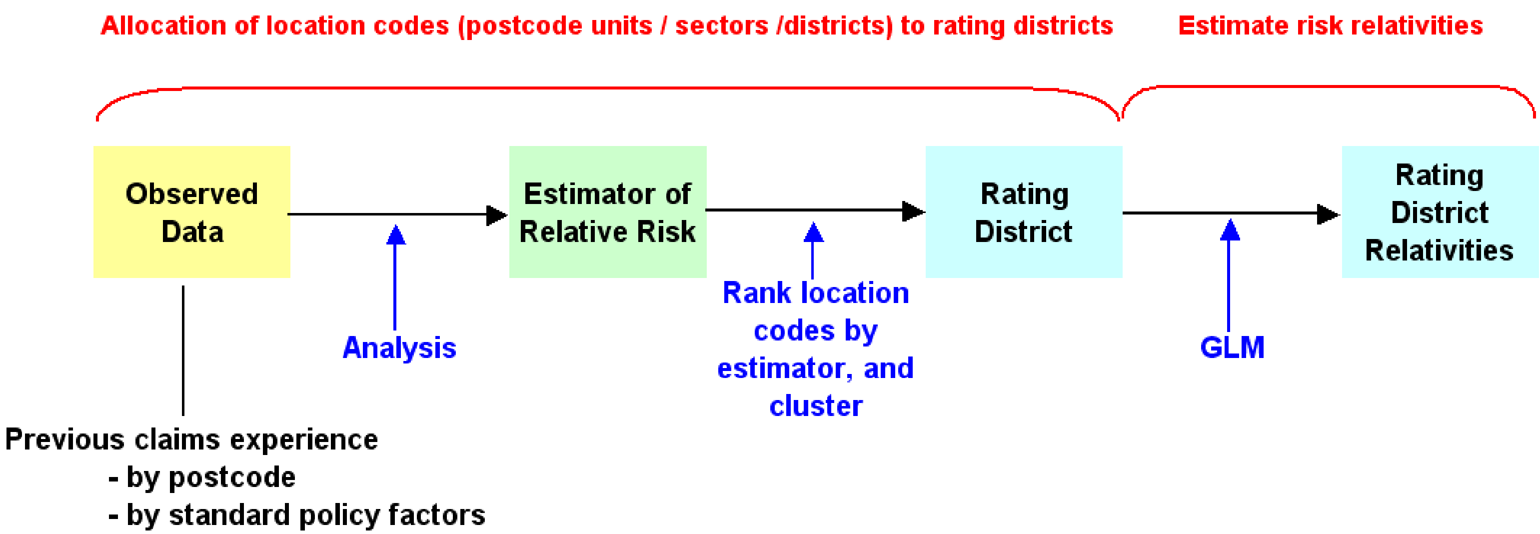
- Car Type
- Gender
- Protective Device
- Age of Home
- AOI
- Insured Age

Classification of Independent Predictors

- Categorical Variables
  - Variables in which the levels take on distinct values
  - Car Type, Gender, Protective Device
- Continuous Variables
  - Variables in which the levels can be quantitatively compared to each other
  - Age of Home, AOI, Insured Age

Territory Rating - Overview

- Aim is to accurately estimate the underlying risk associated with geographic location in order to predict expected claims experience for a given location
- Normally a two stage process:



Multivariate Methods: Techniques

- How do we obtain the best estimator of relative risk for the allocation process when our data is very **sparse** at zipcode level?
  - Spatial Curves: Representing geographical location as x-y co-ordinates and build as part of GLM
  - Residual Modeling: Spatial smoothing of residuals outside GLM
  - Normalization Modeling: External factors (e.g. socio-demographic) as part of GLM

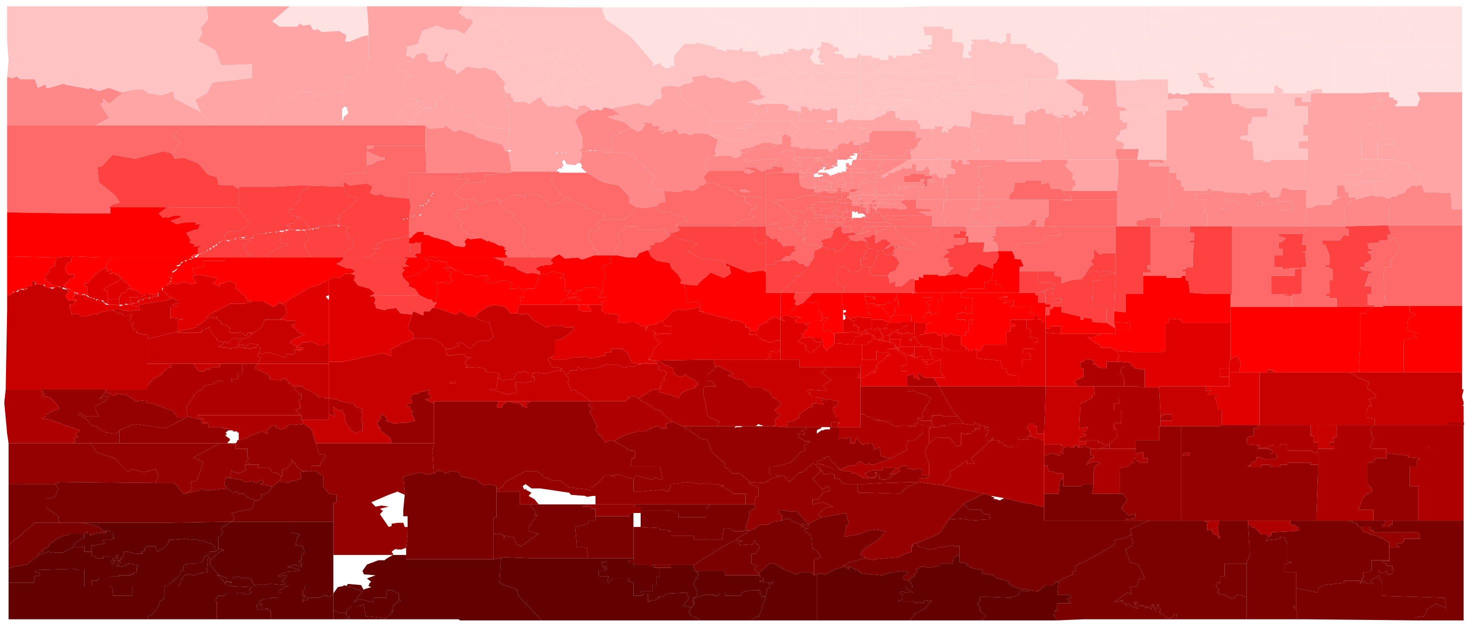
Multivariate Methods: Spatial Curves

- Indication (CZP) = GLM (Rating Variables & f(x,y))
  - Treat the countyzip xy coordinates as a continuous rating variable
  - $f(x,y) = ax + bx^2 + cx^3 + \ldots + dy + ey^2 + fy^3 + \ldots$
    - Coefficients a,b,c … are developed reflecting the multidimensional nature of the rating algorithm with the countyzip concept
    - The x y coordinates of the countyzip combined with the coefficients produces the indicated relativity
- Concerns: Sensitivity
  - Increase the degree of the polynomial
  - Utilize splines to change the curvature while maintaining continuity
  - Nonlinear terms

Multivariate Methods: Spatial Curves

Indication (CZP) = GLM (Rating Variables & f (Lat, Long)):

f (Lat, Long)= a Lat + b Long



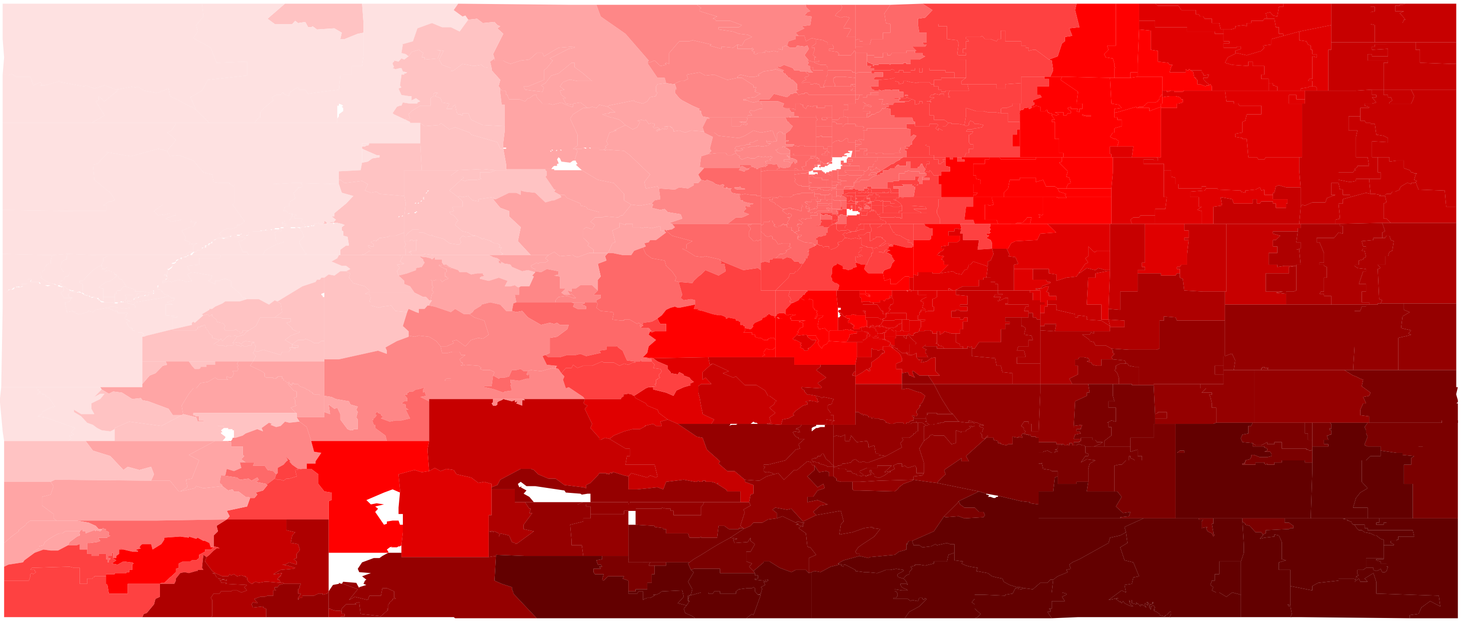
Territorial Ratemaking

- Background
- Traditional Methods
- Multivariate Methods
- Concerns and Issues
- Results

Multivariate Methods: Spatial Curves

Indication (CZP) = GLM (Rating Variables & f (Lat, Long)):

f (Lat, Long)= aLat + bLat2 + cLat3 + dLong + eLong2 + fLong3



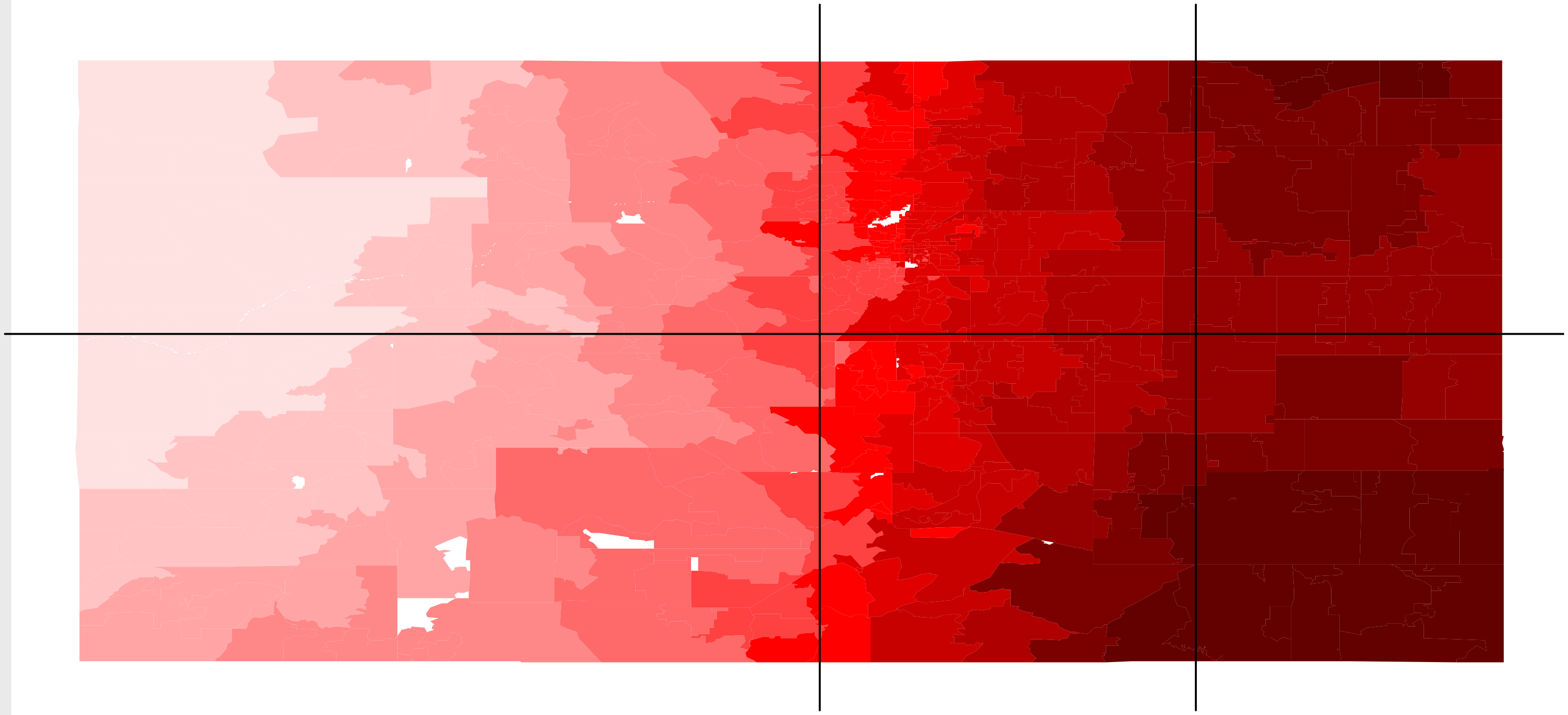
Territorial Ratemaking

- Background
- Traditional Methods
- Multivariate Methods
- Concerns and Issues
- Results

Multivariate Methods: Spatial Curves

Indication (CZP) = GLM (Rating Variables & f (Lat, Long)):

f (Lat, Long)= Spline(Lat) + Spline(Long)



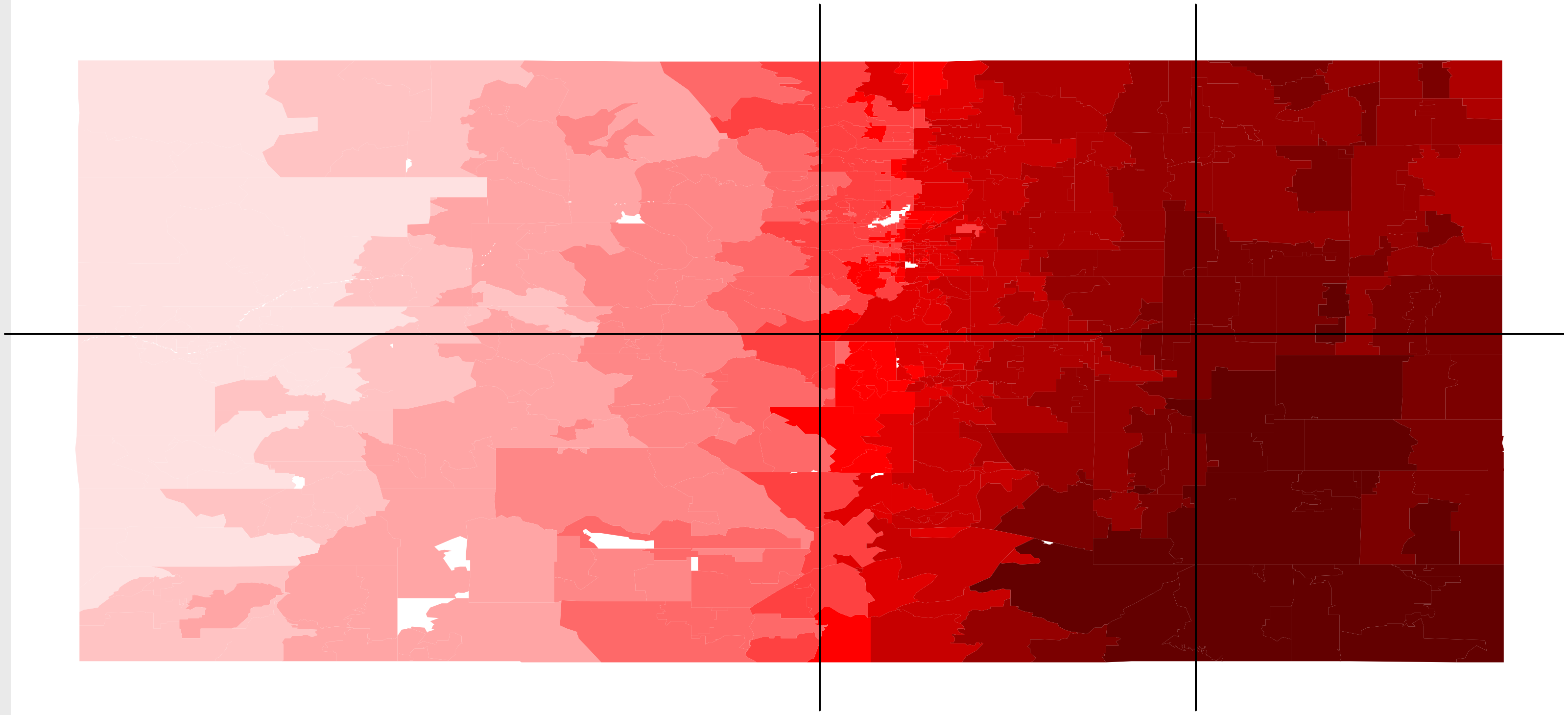
Territorial Ratemaking

- Background
- Traditional Methods
- Multivariate Methods
- Concerns and Issues
- Results

Multivariate Methods: Spatial Curves

Indication (CZP) = GLM (Rating Variables & f (Lat, Long)):

f (Lat, Long) = Spline(Lat) + Spline(Long) + g(Lat x Long)



Multivariate Methods: Residual Modeling

- Indications in the GLM Residual environment
  - $\text{Indication}_{(\text{Boundary})}$ = GLM (Rating Variables & Init Boundary)
  - Init Boundary = Cluster ($\text{SResidual}_{(\text{CZP})}$)
  - $\text{SResidual}_{(\text{CZP})}$ = f($\text{Actual}_{(\text{CZP})}$ – $\text{Modeled}_{(\text{CZP})}$)
  - f() is a spatially smoothing technique based on Latitude, Longitude coordinate of the countyzip centroid
  - $\text{Modeled}_{(\text{CZP})}$ = GLM (Rating Variables excl Territory)
- Concerns
  - Volatile Results
  - To the extent there are distributional biases among other rating variables with location, the residuals do not reflect the entire systematic risk.
  - Allocates unsystematic risk to the territory variable when determining the initial boundaries

Multivariate Methods: Normalization Modeling

- Indications in the GLM Standardization environment
  - Indication $_{\text{(Boundary)}}$ = GLM (Rating Variables & Boundary)
  - Boundary = Cluster (SStandardizedMetric$_{\text{(CZP)}}$)
    - Quantiles
    - Equal Weight
    - Similarity Methods
    - K-means Clustering
  - SStandardizedMetric$_{\text{(CZP)}}$ = f(Resp/(Wt$_{\text{(x,y,z,CZP)}}$ *ModeledRel $_{\text{(x,y,z)}}$))
  - f() is a spatially smoothing technique
    - Distance Based
    - Adjacency Based
  - Modeled $_{\text{(CZP)}}$ = GLM (Rating Variables incl TerritoryProxy)
    - External Data
    - Curves

Territorial Ratemaking

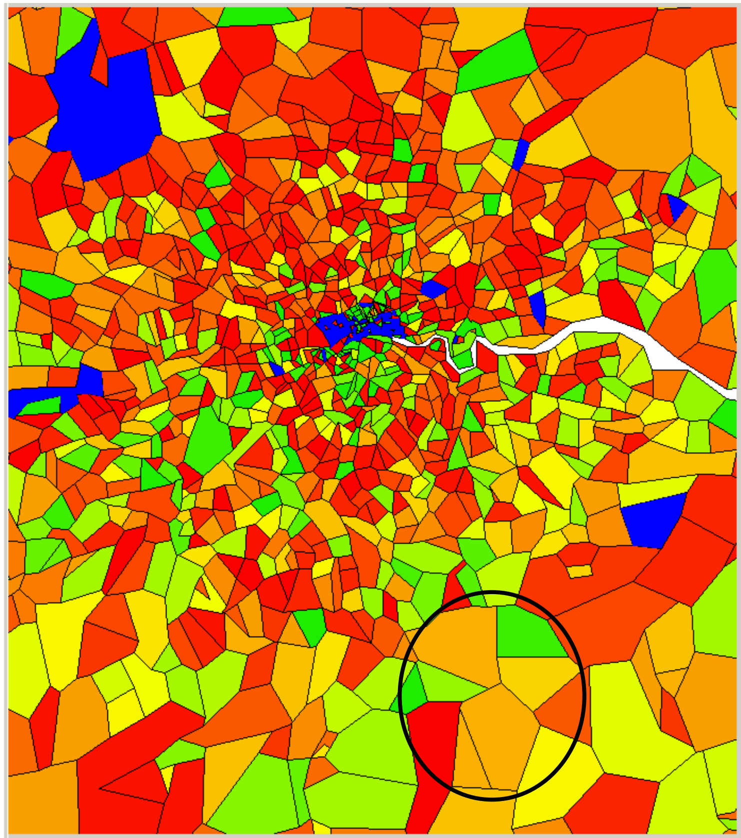
Multivariate Methods: Normalization Modeling

- External data, if available, can be used to enhance the initial territory factor
  - Link in via zipcode and assigns a “score” to each zipcode
- Examples:
  - External Data Provider Scores, e.g. Crime Score based on police crime statistics
  - Lifestyle Factors, e.g. Proportion Married, Educational Attainment
  - Occupation Factors, e.g. Percentage Unemployed
  - Density Factors, e.g. Household Density, Vehicle Density
- Provided the external data is good, should be able to pick out stark borders between neighbours
- However, need to use own data to calibrate the external factors in terms of:
  - Predictiveness
  - Range and steepness of relativities
  - Ability to differentiate between zipcodes (spread of zipcodes between bands)

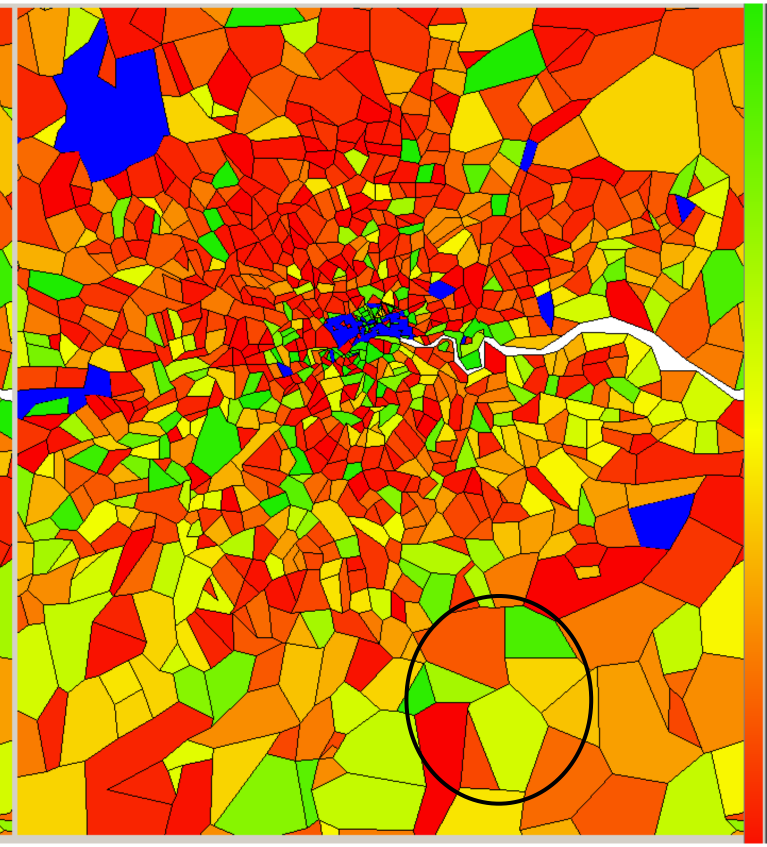
Multivariate Methods: Normalization Modeling

- For each observation, the observed frequency/severity is adjusted to a consistent set of levels for the standard policy factors.
- Use the relativities from the GLM model
- No adjustment is made for any initial grouping of location code
- Can standardize either the observed values or the exposure
  - ***standardizing exposures means no information is lost when a zipcode had no claims***

Multivariate Methods: Normalization Modeling



Observed Private Car AD Frequency



Standardized Observed Private Car AD Frequency

Multivariate Methods: Normalization Modeling

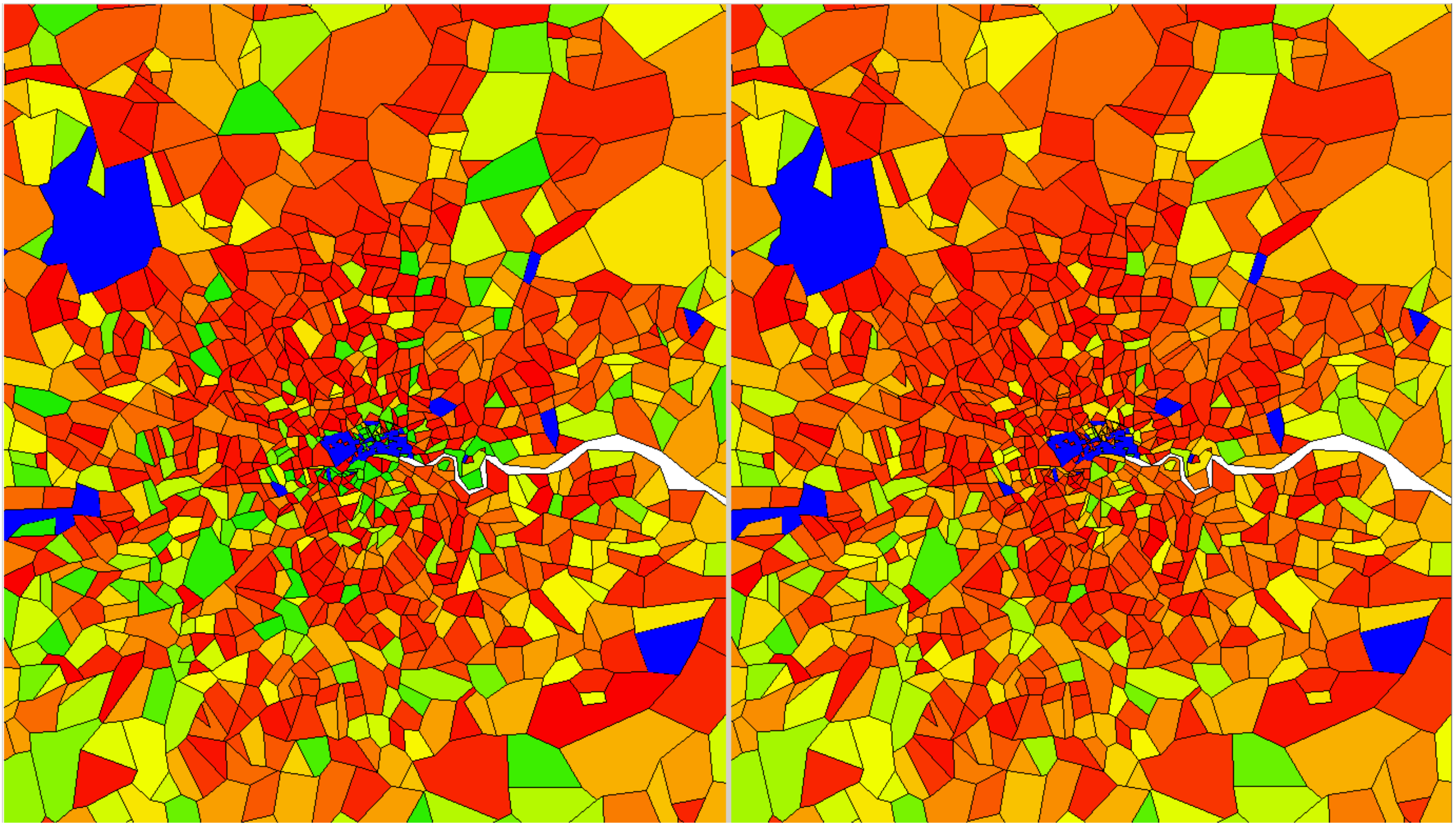
- Distance-Based Spatial Smoothing
  - Advantages
    - Much simpler to implement
  - Disadvantages
    - Takes no account of natural boundaries e.g. rivers, coastline
    - May over smooth urban areas and under smooth rural areas
  - Useful for
    - Windstorm (Homeowners)

- Adjacency-Based Spatial Smoothing
  - Advantages
    - Distribution assumptions allow prior knowledge of claims process to be incorporated
    - Distance can be additionally built in
    - Copes with differences in scale between urban and rural areas
      - rural sectors are larger than urban sectors
  - Disadvantages
    - Artificial boundaries
  - Useful for
    - Auto, Theft (Homeowners)

Territorial Ratemaking

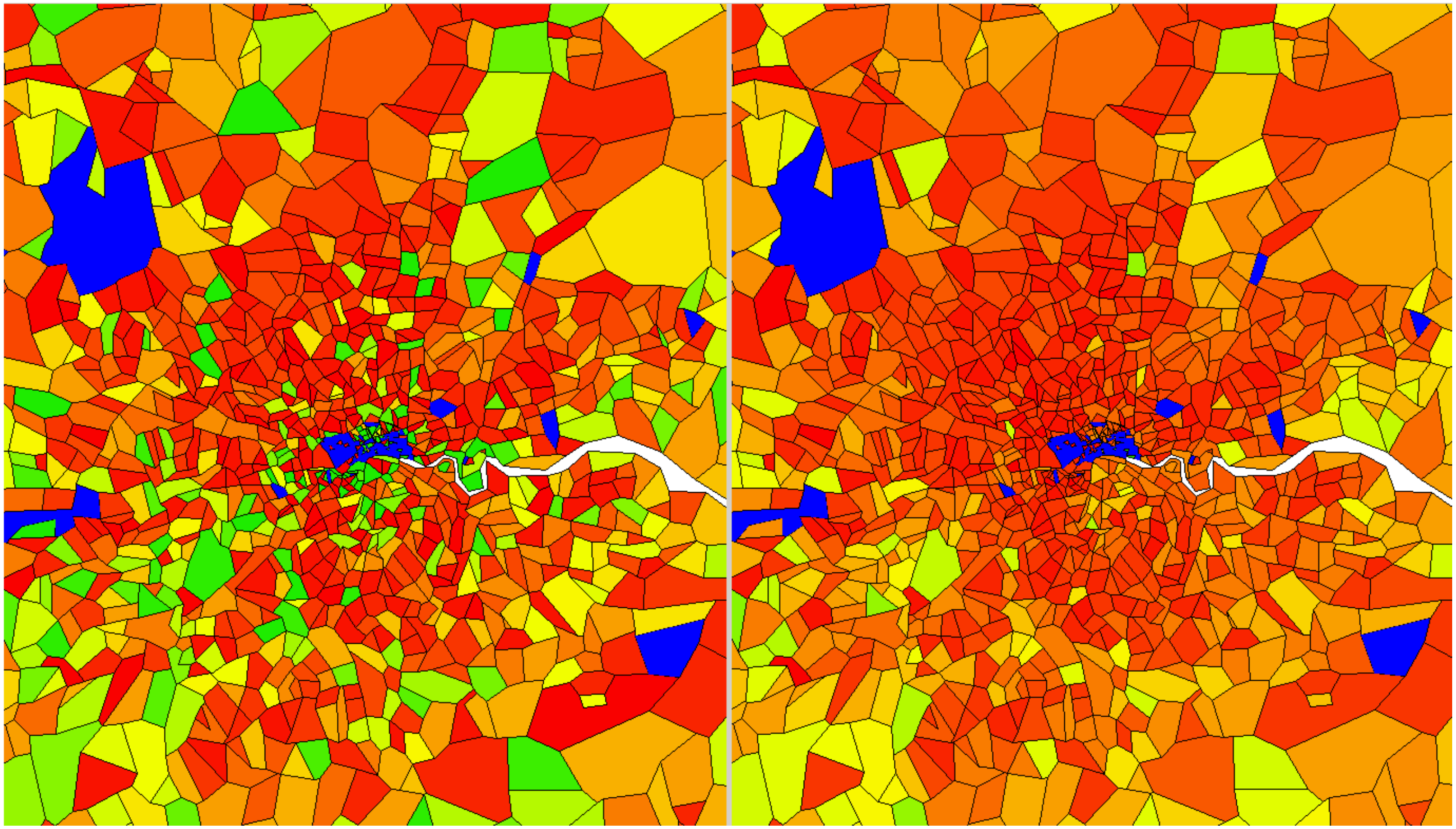
- Background
- Traditional Methods
- Multivariate Methods
- Concerns and Issues
- Results

Multivariate Methods: Normalization Modeling



Standardized Observed Private Car Theft Frequency   Smoothed 20%

Multivariate Methods: Normalization Modeling



Standardized Observed Private Car Theft Frequency　　Smoothed 40%

Multivariate Methods: Normalization Modeling

Clustering

- Goal:
  - Minimize within-group heterogeneity.
  - Maximize cross-group heterogeneity.
  - Produce groupings which are predictive in future.
- Basic Methods
  - Quantiles
  - Equal Weight
  - Similarity Methods
  - K-means Clustering

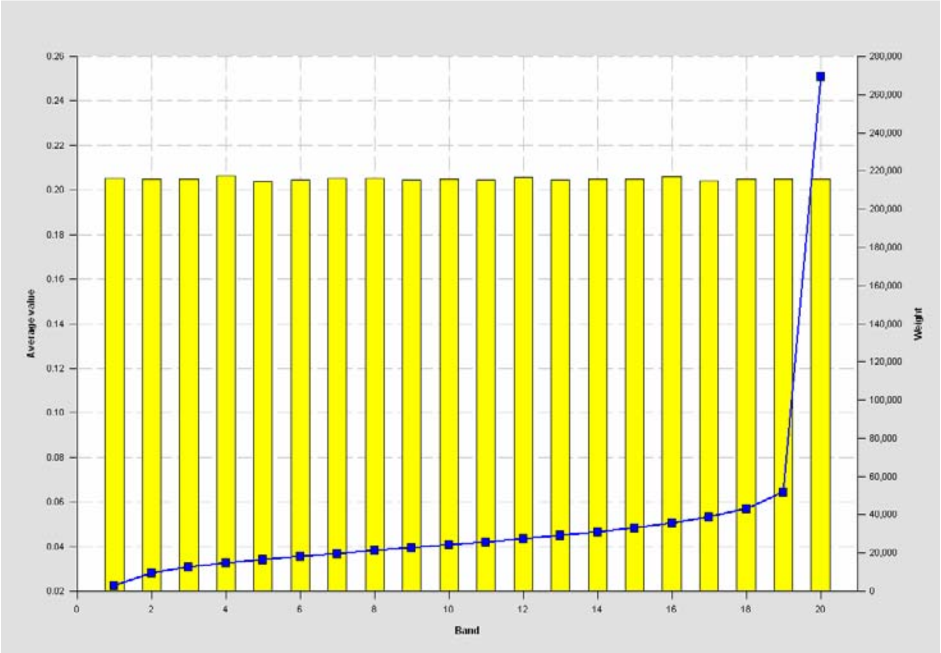
Territorial Ratemaking

- Background
- Traditional Methods
- Multivariate Methods
- Concerns and Issues
- Results

Multivariate Methods: Normalization Modeling

Clustering

- Quantiles
  - Create groups with equal numbers of observations.
- Equal Weight
  - Create groups which have an equal amount of weight.



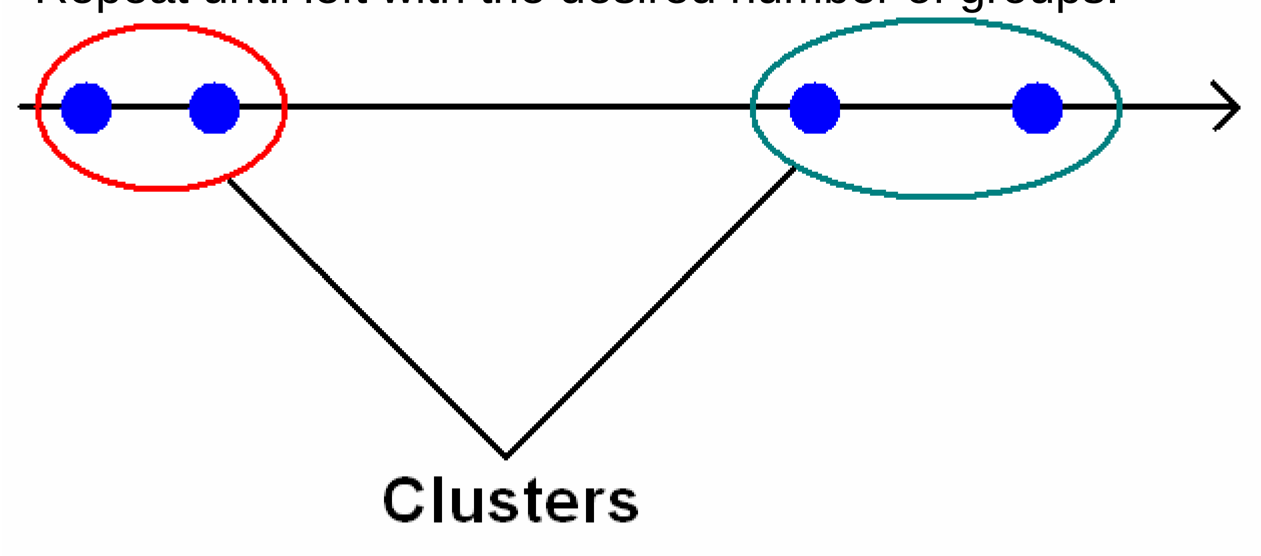
Territorial Ratemaking

Multivariate Methods: Normalization Modeling

Clustering:

Similarity Methods

- General Approach
  - Rank the data set by the statistic you wish to cluster.
  - Decide on which pair of records are the 'most similar.'
  - Group these records.
  - Repeat until left with the desired number of groups.

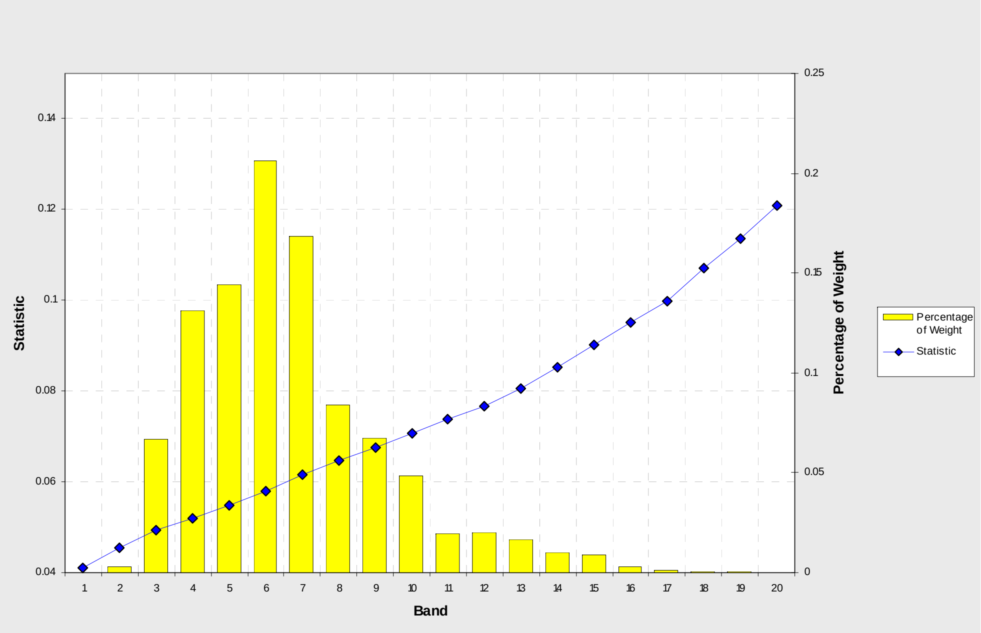


Multivariate Methods: Normalization Modeling

Clustering:

Similarity Methods

- Average Linkage
  - Distance between clusters is the average distance between pairs of observations, one in each cluster.
  - Tends to join clusters with small variances.



Territorial Ratemaking

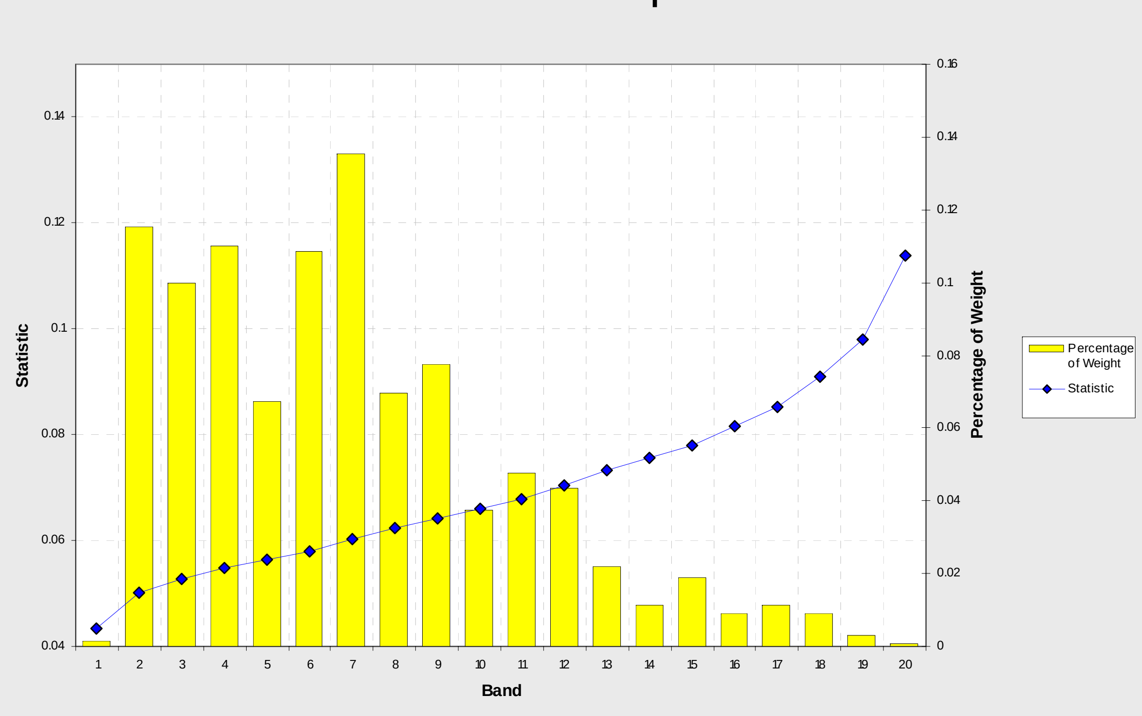
- Background
- Traditional Methods
- Multivariate Methods
- Concerns and Issues
- Results

Multivariate Methods: Normalization Modeling

Clustering:

Similarity Methods

- Centroid
  - Distance between clusters is the difference between the mean values of the clusters squared.



Multivariate Methods: Normalization Modeling

Clustering

- K-means
  - Rank the observations.
  - Split into k groups e.g. using quantile method.
  - Calculate the mean value of each group.
  - Define group start/end-points as being half-way between adjacent mean values.
  - Reallocate each observation.
  - Repeat until group start and end-points converge.

Territorial Ratemaking

- Background
- Traditional Methods
- Multivariate Methods
- Concerns and Issues
- Results

Multivariate Methods: Normalization Modeling

- Concerns
  - Assessing the appropriateness of the external data
    - Levels of significance within the GLM model
    - Hypothesis Testing: Under the null hypothesis zipcodes with the same external factor score have similar underlying experience
  - Assessing the spatial smoothing
    - Hold out samples: difficult due to sparseness of the underlying data in a highly dimensional data space
    - Hypothesis Testing: Under the null hypothesis the p-value should be uniformly spread over [0,1]
  - Assessing the clustering
    - Result boundary should be predictive in the GLM model
- Refinement
  - Internal Constraints
  - External Constraints
  - Local Knowledge

Concerns and Issues

- Defining the Location Predictor:
  - Countyzip units
    - Easy to obtain
    - Definitions unstable over time
    - Responsive?
  - Census tract units
    - Definitions stable over time (change once every ten years)
    - System issues regarding mappings
    - Responsive?
- Defining the continuity
  - Actual centroid of the unit
  - Population weighted centroid of the unit
  - Implement actual Latitude/Longitude coordinates
  - Build company specific coordinate system

Concerns and Issues

- Implementing the results
  - Directly incorporating the metric into a rating algorithm at the unit level
  - Aggregating granular results into a boundary
    - "One Way" Aggregations
    - GLM Aggregation

Territorial Ratemaking

Results

- Isolation of the geographic risk requires the use of multidimensional model.
- Principle of locality allows us to view the geographic risk as a continuous concept.
- Territory Ratemaking Techniques:
  - Use spatial polynomial curves in a GLM to develop indications for the geographic risk.
  - Residuals approaches provide an additional degree of responsiveness to the modeling procedure
  - Normalization techniques adjust for the distributional approach between other rating variables and the location predictor
  - Need to be able to study a variety of spatial smoothing and clustering techniques

Questions?