Territory Analysis with Mixed Models and Clustering

PRESENTED BY ERIC J. WEIBEL

2008 CAS SPRING MEETING

QUEBEC CITY, QUEBEC
JUNE 15 - 18, 2008

Agenda

- Introduction
- Risk Classification Challenges to Territory Analysis
- Homogeneity vs. Credibility
- Mixed Model Approach
- Mixed Model Component 2: Arithmetic Model
- Mixed Model Component 3: Proximity Complement
- Mixed Model Results
- Constrained Cluster Analysis
- Final Results
- Future Research
- Conclusions

Agenda

- **Introduction**
- Risk Classification Challenges to Territory Analysis
- Homogeneity vs. Credibility
- Mixed Model Approach
- Mixed Model Component 2: Arithmetic Model
- Mixed Model Component 3: Proximity Complement
- Mixed Model Results
- Constrained Cluster Analysis
- Final Results
- Future Research
- Conclusions

Introduction

- We apply an objective two-staged territory analysis to California Proposition 103 frequency and severity data at the zip code level
- Stage 1: We apply a mixed model consisting of three components
  - Indication for zip code
  - Predicted value from model or causal geographical variables
  - Complementary indication from proximate zip codes
- Stage 2: Constrained cluster analysis of stage 1 results to assign zip codes to frequency and severity bands

Agenda

- Introduction
- **Risk Classification Challenges to Territory Analysis**
- Homogeneity vs. Credibility
- Mixed Model Approach
- Mixed Model Component 2: Arithmetic Model
- Mixed Model Component 3: Proximity Complement
- Mixed Model Results
- Constrained Cluster Analysis
- Final Results
- Future Research
- Conclusions

Risk Classification Challenges

- Homogeneity vs. Credibility
- Causality
- Controllability
- Loss Control/Incentive Value
- Objectivity
- Integration
- Affordability

Agenda

- Introduction
- Risk Classification Challenges to Territory Analysis
- **Homogeneity vs. Credibility**
- Mixed Model Approach
- Mixed Model Component 2: Arithmetic Model
- Mixed Model Component 3: Proximity Complement
- Mixed Model Results
- Constrained Cluster Analysis
- Final Results
- Future Research
- Conclusions

Homogeneity vs. Credibility

- Loss Cost Gradient (LCG) Dominance
  - Occurs in other variables
  - Solutions with other variables
  - Why no simple solution in territory analysis?
    - Answer: Lack of causality
- Resolution in Territory Analysis
  - Resolution without Auxiliary Data
    - McDonald Approach
    - Proximity Complement Approach
    - Spline & Graduation Approaches
  - Subjective Resolution with Auxiliary Data
  - Objective Resolution with Auxiliary Data
    - Riegel's Approach
    - Arithmetic Model of Causal Geographical Variables
  - Our Approach: Mixed Model of Zip Code Indication, Arithmetic Model, and Proximity Complement

Agenda

- Introduction
- Risk Classification Challenges to Territory Analysis
- Homogeneity vs. Credibility
- **Mixed Model Approach**
- Mixed Model Component 2: Arithmetic Model
- Mixed Model Component 3: Proximity Complement
- Mixed Model Results
- Constrained Cluster Analysis
- Final Results
- Future Research
- Conclusions

Mixed Model Approach

- Introduction by Bishop, Fienberg and Holland (1975) [25]
- Discussed actuarially in general terms of combination of cellular and arithmetic model indications in:
  - Chang & Fairley (1978) [27]
  - Venter (1990) [36]
  - Mildenhall (1999) [33]
- Our proposal is in this very simple general sense of combining an arithmetic model result with dichotomous cellular indications.
- Three Components:
  - Indication for Zip Code
  - Predicted Value from Arithmetic Model of Causal Geographical Variables
  - Proximity Complement
- A more formal mixed model approach might improve the results. See:
  - Searle, Casella and McCulloch, *Variance Components*, 1992
  - Rao, *Variance Components Estimation*, 1997
  - Actuarial Discussions of Hierarchical Generalized Linear Models (HGLM) or discussions of Generalized Linear Mixed Models (GLMM)

Agenda

- Introduction
- Risk Classification Challenges to Territory Analysis
- Homogeneity vs. Credibility
- Mixed Model Approach
- **Mixed Model Component 2: Arithmetic Model**
- Mixed Model Component 3: Proximity Complement
- Mixed Model Results
- Constrained Cluster Analysis
- Final Results
- Future Research
- Conclusions

Mixed Model Component 2: Arithmetic Model

- Model Form
  - Multiple Linear Regression Elected for Simplicity
  - Alternative: Spatial Autocorrelation Model with Similar Covariates
- Selecting Causal Geographical Variables
  - Spatial Interaction
  - Causal Geographical Variables

Regression Models

- Objectives:
  - Our overriding objective is prediction; we want to produce the best credibility complement.
  - Our secondary objective is to provide groundwork for further research into the introduction of causal geographical variables.
    - This involves favoring quantitative variables over categorical ones.
    - Involves selecting variables that are likely to be deemed acceptable as rating variables.
    - Also involves structuring the model in a way that is easier to understand.
      - Our models ended up involving a lot of variables, and definitely had to sacrifice ease of explanation for accuracy
      - Any project to directly introduce causal geographical variables for the first time might need to use simpler models whose coefficients are easy to explain.

Regression Models

- Spatial Interaction: “the movement of people, materials, capital and information between geographic locations.” Miller and Han (2001) [48].
    - The fact that vehicles are not driven in a single zip code creates spatial interaction.
    - Causal variable measurements should account for spatial interaction in automobile insurance.
    - Our general approach was to compute values for our variables within the zip code itself, and for zip codes within three mutually exclusive radii, of 10, 25 and 50 miles.
        - We did vary this approach at times in response to the data

Review of Causal Geographical Variables

- Review of variables posited as being causal in the geographical LGP
- We discuss the most immediately promising variables and sources of data
- We elected to include three of these promising variables as candidates for our models

We Model

- Traffic Density
- Legal Climate
- Population Density

We Discuss

- Nature of Population
- Enforcement
- Weather

Others

- Topography
- Roads
- Regulation
- Education
- Medical Costs
- Repair Costs

Traffic Density

- Available at the zip code level from the decennial census:
  - Population
  - Number of Vehicles
  - Time Spent on the Road to Work
- We elected to focus on the number of minutes spent commuting one-way by each commuter.
  - Derived from a 1990 decennial census variable
- Miles of road lane were not available below the county level
- land area and populated land area used as spatial denominator
  - Basic Land area taken directly from decennial census
  - Populated land area – only include populated census block area

Legal Climate

- Difficult to Measure Variables:
  - History and Current Philosophy of Local Court Jurisdiction
  - Friendliness of Potential Juror Pool to Claimants
  - Nature and Level of Activity of Local Bar
  - Existence of Networks of Physicians and Lawyers who Cooperate
- Easy to Measure: Lawyer Density
  - We used the number of people employed in legal offices for each zip code as our numerator. This was taken from the 2005 survey of economic conditions.
  - Land Area and Population are both plausible denominators
    - We elected population, since many of our other variables employ land area as a denominator.
    - Mismatch between 1990 decennial census data and 2005 survey data.

Population Density

- Numerator simply the population for each zip code from the 1990 decennial census.
- Denominator
  - Total Land Area (see slide 17)
  - Populated Land Area (see slide 17)
  - Block Level
- Block Weighted Density, surprisingly, did not perform well at all
  - Population weighted average block level density is the measure
- Populated Land Area and regular Land Area performed about the same.
- Hence we elected basic population density measure using basic land area as denominator.

Geographical Binary Variables

- We only resorted to these variables when no other variable combinations could come close to the level of fit.
- We only introduced very basic, large variables based on our *a priori* expectations: with variables for San Francisco, Los Angeles, and Remainder of State.
  - In the modeling process, we discovered that central Los Angeles and the remainder of Los Angeles behaved somewhat differently, so we split Los Angeles into two variables. So we ended up with three binary variables:
    - Los Angeles = Central Los Angeles = 90001 to 90077
    - Los Angels Area = Remainder of Los Angeles County
    - San Francisco = City of San Francisco
- To a large extent, these variables probably reflect differences in the legal environment for BI coverage and perhaps for PD severity. But other effects may be picked up as well.

Other Causal Variables Discussed

- Population Characteristics
  - Class Plan Off-Balance Effects
  - Externality Effects from Variables Reflected in Class Plan
  - Externality Effects from Variables not Reflected in Class Plan

- Implementation and Enforcement
  - Traffic Enforcement
  - Enforcement Ratio

- Weather

Variable Definitions

CT_i Time spent commuting to work, one-way.

$TD10_i$ Commute time one-way / Land Area

$TD25_i$ $TD50_i$

$LD25_i$ Law office employees / Population

$LD50_i$

PD_i Population / Total Land Area

$PD10_i$ $PD25_i$ $PD50_i$

LA_i = 1 if zip 90001 to 90077, else 0

LAC_i =1 if remainder of LA County, else 0

SF_i =1 if city of San Francisco, else 0

Variables with a numerical suffix: This refers to the mile radius. Note that these radii were mutually exclusive and exhaustive. For instance LD25 includes all lawyers and land area for zip codes within 25 miles, including the zip code being modeled. If no numerical suffix for a quantitative variable, then only includes zip code being modeled. Binary geographical variables have no need for numerical suffix.

Final Models

$$BIFQ_i = \hat{\alpha} + \hat{\beta}(CT_i) + \hat{\gamma}(TD10_i) + \hat{\delta}\ (TD25_i) + \hat{\varepsilon}(TD50_i) + \hat{\epsilon}(LD25_i) + \hat{\theta}(LD50_i) + \hat{\vartheta}(LA_i) + \hat{\pi}(LAC_i) + \hat{\rho}(SF_i) + \hat{\tau}(CT_i TD25_i) + \hat{\phi}(CT_i LA_i) + \hat{\omega}(LD25_i LAC_i) + \hat{\varsigma}(LD50_i LAC_i) + \hat{\xi}(CT_i LD25_i)$$

$$PDFQ_i = \hat{\alpha} + \hat{\beta}(CT_i) + \hat{\gamma}(TD10_i)^{0.5} + \hat{\delta}\ (TD25_i)^{0.5} + \hat{\vartheta}(LA_i) + \hat{\pi}(PD_i) + \hat{\rho}(PD10) + \hat{\theta}(PD25_i)^{0.5} + \hat{\epsilon}(CT_i TD10_i) + \hat{\tau}(CT_i TD25_i) + \hat{\xi}(CT_i PD10_i) + \hat{\varphi}(CT_i PD25_i) + \hat{\omega}(CT_i LA_i)$$

$$BISV_i = \hat{\alpha} + \hat{\beta}(CT_i) + (LD25_i) + \hat{\theta}(LD50_i) + \hat{\gamma}(TD10_i) + \hat{\varepsilon}\ (TD50_i) + \hat{\vartheta}(LA_i) + \hat{\pi}(LAC_i) + \hat{\tau}(CT_i LD25_i) + \hat{\phi}(CT_i LD50_i) + \hat{\omega}(LD50_i LA_i)$$

$$PDSV_i = \hat{\alpha} + \hat{\beta}(LD25_i)^{0.5} + \hat{\pi}(PD_i)^{0.5} + \hat{\varsigma}(PD10_i)^{0.5} + \hat{\theta}(PD25_i)^{0.5} + \hat{\gamma}(PD50_i)^{0.5} + \hat{\vartheta}LA_i + \hat{\pi}LAC_i + \hat{\rho}SF_i + \hat{\epsilon}(CT_i * LD25_i)^{0.5}$$

Agenda

- Introduction
- Risk Classification Challenges to Territory Analysis
- Homogeneity vs. Credibility
- Mixed Model Approach
- Mixed Model Component 2: Arithmetic Model
- **Mixed Model Component 3: Proximity Complement**
- Mixed Model Results
- Constrained Cluster Analysis
- Final Results
- Future Research
- Conclusions

Mixed Model Component 3: Proximity Complement

- Hunstad Method
  - Use Local Assigned Risk Territory Data as Complement
- Tang Method
  - Use immediately contiguous zip codes as 1st complement
  - If necessary use Local Assigned Risk Territory as 2nd Complement
- Hunstad Suggestions
  - Weight each zip code by distance
  - Add individual zip codes until full credibility reached
- Our Approach – 10 mile distance

Agenda

- Introduction
- Risk Classification Challenges to Territory Analysis
- Homogeneity vs. Credibility
- Mixed Model Approach
- Mixed Model Component 2: Arithmetic Model
- Mixed Model Component 3: Proximity Complement
- **Mixed Model Results**
- Constrained Cluster Analysis
- Final Results
- Future Research
- Conclusions

Mixed Model Results

- Credibility Weighting Formula
- Local Mixed Model Component Performance

Credibility Weighting Formula

- Zip code indication credibility, z, determined by the 1,082 claim rule.
- Proximity complement credibility:

$$z_p = \frac{\left(\sqrt{\frac{c}{1082}}\right)(1 - z)}{\left(\sqrt{\frac{c}{1082}} + R^2\right)}$$

- Arithmetic model credibility:

$$z_m = \frac{(R^2)\,(1 - z)}{\left(\sqrt{\frac{c}{1082}} + R^2\right)}$$

- R^2 is the corresponding arithmetic model statistic, c is the number of claims in the proximity complement

Credibility Weighting Formula

- Our goal was to introduce the concept, rather than implement the best possible means of combining mixed model elements.
- As a result we did not devote much effort to arriving at a credibility weighting scheme.
- We leave it to future researchers to arrive at optimal credibility weighting scheme, which ideally would incorporate the relative local fit of the arithmetic model and proximity complement.
- Or, perhaps a more formal mixed model could be arrived at.
- Because of the rudimentary nature of our implementation, we were willing to intervene in the credibility weighting process in the event the local performance of the arithmetic model or proximity complement was too poor.

Local Mixed Model Component Performance

- Plots of actual values, model predicted values and residuals, and proximity complements are presented in Appendix A.
- Bodily Injury Liability Frequency:
  - Los Angeles
    - Central LA exhibits steep LCG and high information density.
    - In this environment, we would expect and do observe poor performance for our proximity complement.
    - The proximity complement radius is static at 10 miles.
    - In central LA 10 miles is too much.
    - We employ two binary geographical variables in the arithmetic model in LA, so the model does not suffer from any significant local bias
      - The high density ensures that credible amounts of data can be obtained with a smaller radius
      - At the same time, the steep LCG means that extending the radius further than necessary will introduce significant heterogeneity.
    - We elected to intervene in the credibility weighting process because of the poor quality of the proximity complement and the good quality of the model. We assigned 0 credibility to the proximity complement for zip codes in and around central LA – 90001 to 91108.

Local Mixed Model Component Performance

- San Francisco
  - While information density is high, there is not a steep LCG.
  - Hence, the proximity complement is not particularly biased, although they are tightly bunched due to information density within the city. The proximate ocean and bay may also contribute to the uniformity. This is worth further study later.
  - The binary geographical rating variable for San Francisco ensures good local performance of the arithmetic model.
- Oakland/Berkeley
  - Modest positive residual bias for arithmetic model
- Suburban Areas
  - Arithmetic Model Underestimates
    - Severe: Fresno, Sacramento
    - Moderate: San Jose
- Rural Areas
  - Arithmetic Model Overestimates
    - Severe: Extreme Northern California away from the coast
    - Moderate: Extreme Northern California on the coast
  - Proximity Complement Performance
    - Excellent. Appears unbiased.
    - However, particularly in extreme Northern California, precision could be improved by extending the radius. The information density here is low and the LCG appears to be relatively flat.
    - There may be less of a need for a wider radius in rural Central and Southern California

Local Mixed Model Component Performance

- Property Damage Liability Frequency:
  - The LCG is usually not steep, so the problems that occurred in LA with respect to the proximity complement are not repeated.
  - Arithmetic Model
    - Over-predicts again for inland and coastal extreme Northern California. But the problem is much less pronounced.
    - Modest over-prediction for San Jose.
- Bodily Injury Liability Severity:
  - LCG is not steep. No major proximity complement problems.
  - Arithmetic Model
    - Central Orange County: Modestly under-predicted.
    - Oakland/Berkeley: Moderate over-prediction.
    - Part of Marin County: Underestimated
    - Santa Rosa: Underestimated
    - Sacramento: Modest underestimate
    - Part of Desert Area: Underestimated
    - Santa Barbara: Underestimated

Local Mixed Model Component Performance

- Property Damage Liability Severity:
  - LCG usually not steep. No major proximity complement problems.
  - Arithmetic Model
    - Southwest Orange County: Extreme under-prediction
    - Sacramento: Significant under-prediction
    - Part of San Diego County: Small overestimate.
    - Oakland/Berkeley: Modest underestimate
    - Extreme Northern California Inland: Modest overestimate.

Mixed Model Component Conclusions

- Regression Model Conclusions
  - Ideally, simple binary variables will not need to be introduced, and other continuous causal variables could be introduced that would reflect these differences.
  - Failing that, should try to define boundaries of geographical binary variables that correspond with court jurisdiction groupings
  - Bodily Injury Liability Frequency:
    - Appears significant local improvement in fit could be achieved by adding binary geographical rating variables for the following areas
      - Inland Extreme Northern California
      - Fresno
      - Sacramento
      - San Jose
      - Oakland/Berkeley

Mixed Model Component Conclusions

- Property Damage Liability Frequency
  - Improvement in fit could be achieved by adding binary geographical rating variables in the following areas:
    - Inland Extreme Northern California
    - San Jose
- Bodily Injury Liability Severity
  - Improvement in fit could be achieved by adding binary geographical rating variables in the following areas:
    - Central Orange County
    - Oakland/Berkeley
    - Part of Marin County, Santa Rosa, and Santa Barbara (these areas are similar in nature)
    - Sacramento
    - Part of Desert Area

Mixed Model Component Conclusions

- Property Damage Liability Severity
  - Improvement in fit could be achieved by adding binary geographical rating variables in the following areas:
    - Southwest Orange County
    - Sacramento
    - Oakland/Berkeley
    - Extreme Northern California Inland
- Proximity Complement Conclusions
  - A dynamically determined radius would dramatically improve performance.
    - Information Sparseness = increase radius
    - Information Density = decrease radius
    - Steep LCG = decrease radius
    - Flat LCG = increase radius
  - In Appendix C of the paper, we compare our proximity complement performance with Hunstad for BI frequency, using mean absolute deviation for each CAARP territory.

Agenda

- Introduction
- Risk Classification Challenges to Territory Analysis
- Homogeneity vs. Credibility
- Mixed Model Approach
- Mixed Model Component 2: Arithmetic Model
- Mixed Model Component 3: Proximity Complement
- Mixed Model Results
- **Constrained Cluster Analysis**
- Final Results
- Future Research
- Conclusions

Our Objectives in Creating Groupings

- California requires that for each coverage, zip codes be grouped into frequency and severity bands. Up until very recently, a maximum of 10 bands have been allowed per coverage. In our analysis we group frequency and severity into 10 bands for bodily injury liability and property damage liability.
- The use of professional judgment in creating territorial groupings is a frequent source of criticism: Barber (1929) [1], Casey et al. (1976) [26], Phase I (1978) [19], Shayer (1978) [34].
- Our goal is to objectively group zip codes into bands that accurately reflect their expected relative frequency and severity rates.
- We wish to be able to impose various social and regulatory acceptability constraints on the grouping process
- One of the reasons for grouping in the first place, a complement of credibility, is less of a concern for us because we have already incorporated complimentary information from the arithmetic model and from the surrounding zip codes

Desired Features of Groupings

- It makes sense to specify our decision variables as binary, arrayed in a matrix of 10 columns and 1,502 rows, with the columns corresponding to frequency or severity bands, and each row corresponding to a zip code.
- Only one column in each row can take on a value of “1”, meaning that the zip code belongs to that band. The remaining columns of the row must have “0” values.
- These might be setup as follows:

Desired Features of Groupings

- $x_{ij} \in [0,1] \in \mathbb{N}$ (2.3)
- $\sum_i x_{ij} = 1$ (2.4)
- Where i ranges from 1 to 1,502, and j ranges from 1 to 10. Desirable L_2 or L_1 objective functions might be:

$$\min \sum_i \sum_j \left[\left(R_i - \frac{\sum_b x_{bj} R_b E_b}{\sum_c E_c x_{cj}}\right) x_{ij} E_i\right]^2 \qquad (2.5)$$

$$\min \sum_i \sum_j \left[abs\left(R_i - \frac{\sum_b x_{bj} R_b E_b}{\sum_c E_c x_{cj}}\right) x_{ij} E_i\right] \qquad (2.6)$$

- R_i is the computed mixed model relativity. E_i is the number of exposures in zip code i.

Desired Features of Groupings

- We also want to impose constraints
  - No band can consist of a land area of less than 20 square miles
  - We may wish to impose a minimum exposure or claim count for each band for credibility purposes
  - We may wish to impose factor weight constraints
- The 20 square mile constraint could be setup as follows, with Li representing the land area for zip code i.

$$\sum_i L_i x_{ij} \geq 20 \qquad (2.7)$$

Cluster Analysis Review

- Cluster Analysis would appear to be natural choice
- Cluster Analysis literature is vast, diverse and somewhat unorganized. It developed somewhat independently under the auspices of different academic disciplines
- The two standard texts are Kaufmann and Rouseauww (KR in sequel) (1990) [46] and Everitt, Landau and Leese (2001) [43]. Han, Kamber and Tung (HKT in sequel) (2001) [45] also provide a remarkably brief introduction.
- Use of Cluster Analysis for our purpose was mentioned once before in the actuarial literature in Phase I (1978) [19]. However, the authors ended up manually grouping zip codes into bands.

Cluster Analysis Review

- One major divide in Cluster Analysis techniques is the distinction between Hierarchical and Partitioning (KR) / Optimization (Everitt et al.)
- KR claim that Partitioning / Optimization methods will tend to arrive at the best groupings for a fixed number of clusters.
- Since we are interested in a fixed number of clusters – 10 bands, this would incline us to look into Partitioning / Optimization clustering.
- KR also emphasize robust methods. L1 norms are considered more robust. So this would incline us to prefer (2.6) to (2.5).

Cluster Analysis Review

- Recall that we also wish to impose constraints.
- Imposition of constraints is a very new topic in cluster analysis. It is not even mentioned in KR. Everitt et al. discuss it but focus on proximity/contiguity constraints and certain constraints related to hierarchy.
- HKT have a broader discussion of pioneering work being done. In particular they refer to Tung et al (2001) [52]

Cluster Analysis Review

- Tung et al divide constraints into six types: Existential, Universal, Existential-Like, Parameter, Summation, and Averaging.
- We are interested in summation constraints. Averaging constraints are very similar to summation constraints.
- Summation constraint involves the sum of some quantity tied to the units being grouped. In our case land area would be an example. Each zip code has a land area, and we constrain land area for each band to exceed 20 square miles.
- Factor weight constraints or minimum claim or exposure counts (for credibility purposes) are similar.
- Unfortunately, Tung et al do not provide a method of solution, and furthermore discuss the difficult nature of the problem.

Cluster Analysis Review

- Berkhin (2006) [38]
  - Provides very recent survey of recent advances in cluster analysis, including constrained cluster analysis.
  - Unfortunately, references HKT and Tung et al, which we have already covered.
  - Since HKT and Tung et al. both discuss how difficult *summation* constraints will be to solve, this leaves us in a bit of a pinch with respect to the cluster analysis literature.
- Teboulle et al. (2006) [51]
  - Indicates that most partitioning/optimization problems in cluster analysis involve non-convex objective functions. Draws relationship between *k-means* cluster analysis and nonlinear programming gradient-type method.

Cluster Analysis Review

- Cluster Analysis literature provides no answers for summation constraints at this time.
- Given that a relationship between partitioning / optimization cluster analysis and nonlinear programming has been made, it would seem we should look to nonlinear programming to see if it offers a solution.
- A review of our objective function and initial constraints reveals that it can be considered a nonlinear programming problem from operations research. See Hillier and Lieberman (1995) [60].

Nonlinear Programming Approach to Constrained Clustering

- Non-convex objective function
- Binary decision variables
- Linear / binary type constraints
- R, which we have been using up to this time does not have pre-programmed packages for handling this type of problem.
- The problem is too large to be handled by the standard spreadsheet solver.
- Fortunately, Frontline Systems, Inc., distributes an advanced solver that can plug right into the spreadsheet

Nonlinear Programming Approach to Constrained Clustering

- Constrained non-convex pure integer programming problem
- As originally configured, our problem is too large to be solved in a reasonable amount of time
- The size of the problem can be significantly reduced, and its structure made more clear with a few steps
  - Sort the zip codes, from smallest mixed model indication to largest
  - Remember that a zip code can only be assigned to one band
  - Quickly becomes apparent that many of the decision variables are irrelevant. For example, the rightmost rows are clearly irrelevant for zip codes with low mixed model indications – an optimal solution will never assign those zip codes to one of the high bands. And, the leftmost rows are clearly irrelevant for zip codes with very high mixed model indications – an optimal solution will never assign those zip codes to one of the lowest bands. So a considerable amount of pruning can be done which reduces the size of the problem.

Frontline KNITRO^{tm}

- KNITRO^{tm} appeared to be the best solver engine to use for our problem.
- All integer programming type problems employ branch & bound.
- KNITRO^{tm} uses one of three methods each time it conducts a minimization step
  - Interior Point Algorithms (Barrier Methods): Byrd, Gilbert and Nocedal (2000) [56], Byrd, Nocedal and Waltz (2003) [58]
    - Conjugate Gradient
      - Has a step which improves feasibility
      - Has a tangential step which improves optimality. Uses projected conjugate gradient iteration.
    - Direct
      - Primal-dual KKT system solution via direct linear algebra
  - Active Set (Sequential Linear Quadratic Programming): Byrd, Gould, Nocedal and Waltz (2004) [57]
    - First stage identifies constraints that are “active” for the first solution of the problem, which is a linear approximation within a trust region.
    - Second stage is quadratic approximation using projected conjugate gradient, subject only to constraints identified as “active” in first stage.

Problem Setup

- Starting with BI frequency, we began by dividing up the matrix of decision variables into roughly equal length sections in terms of number of zip codes (rows)
- Then we pre-assigned the decision variables “0” or “1” values in discrete columns.
- The first zip codes, numbered i=1 to 148, were assigned to frequency band “1”, which means that the first of the ten columns (j=1) were assigned the value “1” while the remaining columns (j=2 to 10) were assigned “0” values. For i=1 to 296, the column j=2 was assigned values of “1” while the columns corresponding to j=1 and j=3 to 10 were assigned values of “0”. And so forth.
- As we discussed earlier, the problem as specified is far too large to be solved with a practicable amount of time or computer resources.

Reducing the Size of the Decision Variable Matrix

- As we discussed earlier, we trimmed the width of the decision variable matrix by pruning off decision variables which clearly would not be assigned a “1” value in any optimal solution. As an example, the cell at (1,10) would have been among the first removed, since certainly the zip code with the lowest mixed model indication was not going to be assigned to the highest frequency band.
- Even after pruning back the size of our problem considerably, it was still too large.
- Through successive experimentation we found that the problem had to be restricted both in terms of width around the “trial solution” and in terms of the number of zip codes considered at one time.

Reducing the Complexity of the Problem

- We also found that it was advantageous to use our knowledge of the structure of the problem and what an optimal solution will look like
- We know that since we sorted the zip codes by mixed model indication, from smallest to largest, an optimal solution will tend to have "1" values which march forward in discrete columns.
- By incorporating this structure into a system of constraints, we can save computational time, preventing the computer from evaluating a lot of solutions which clearly will not be optimal.

Reducing the Complexity of the Problem

- We prevent consideration of band assignments that move “backwards” through the following system of constraints:

$$0 \le \sum_{j=1}^{10} j\left[x_{(i+1),j} - x_{i,j}\right] \le 1 \; for \; i \; from \; 1 \; to \; 1{,}501 \qquad (3.5)$$

- This corresponds to the entire range of decision variables. When we reduce the size of the problem as we discussed in slide 55, we can reduce this constraint to the same dimensions

Final Model Formulation

- Our final method of solution is a sequential one, which breaks the problem down into manageable pieces.
- We present the initial model formulation for BI frequency below, and then discuss the sequential solution procedure.
- We began by only considering decision variables in the following limited range:

$$x_{ij}\ for\ i \leq 148, j \leq 2\ and\ for\ 149 \leq i \leq 296\ j \leq 3, and\ 297 \leq i \leq 444, 2 \leq j \leq 4$$

Final Model Formulation

- We elected to use the L1 objective function (2.6), which converted to the range specified above is:

$$\min\left[\begin{array}{l} \sum_{i=1}^{148}\sum_{j=1}^{2}\left[abs\left(R_i-\frac{\sum_b x_{bj}R_bE_b}{\sum_c E_cx_{cj}}\right)x_{ij}E_i\right] \\ +\sum_{i=149}^{296}\sum_{j=1}^{3}\left[abs\left(R_i-\frac{\sum_b x_{bj}R_bE_b}{\sum_c E_cx_{cj}}\right)x_{ij}E_i\right] \\ +\sum_{i=297}^{444}\sum_{j=2}^{4}\left[abs\left(R_i-\frac{\sum_b x_{bj}R_bE_b}{\sum_c E_cx_{cj}}\right)x_{ij}E_i\right] \end{array}\right] \qquad (3.7)$$

Final Model Formulation

- In our initial attempts we decided to ignore the minimum land area constraint (2.7). Should a solution ever violate or threaten the constraint we would backtrack and add the constraint.

Sequential Solution Procedure

- The sequential solution procedure essentially involves moving downward and to the right through our original range of decision variables.
  - Initial Solution Stage
  - Solution Check Stage
    - Turns out this was not necessary. Solutions are stable.
  - Sequential Advancement Stage
  - Reaching the final band
- A summary of our setup and solutions in sequence for BI Frequency follows.

	i range	FB1	FB2	FB3	FB4	FB5	FB6	FB7	FB8	FB9	FB10
Setup1	1 to 148	1	0								
	149 to 296	0	1	0							
	297 to 444		0	1	0						
Solution1	1 to 116	1									
	117 to 275		1								
	276 to 444			1							
Setup2	117 to 275		1	0							
	276 to 444		0	1	0						
	445 to 592			0	1	0					
Solution2	117 to 276		1								
	277 to 453			1							
	454 to 592				1						
Setup3	277 to 453			1	0						
	454 to 592			0	1	0					
	593 to 740				0	1	0				
Solution3	277 to 474			1							
	475 to 628				1						
	629 to 740					1					

Setup4	475 to 628	1	0				
	629 to 740	0	1	0			
	741 to 888		0	1	0		
Solution4	475 to 637	1					
	638 to 766		1				
	767 to 888			1			
Setup5	638 to 766		1	0			
	767 to 888		0	1	0		
	889 to 1036			0	1	0	
Solution5	638 to 794		1				
	795 to 927			1			
	928 to 1036				1		
Setup6	795 to 927			1	0		
	928 to 1036			0	1	0	
	1037 to 1184				0	1	0
Solution6	795 to 928			1			
	929 to 1067				1		
	1068 to 1184					1	

Setup7	929 to 1067	1	0			
	1068 to 1184	0	1		0	
	1185 to 1332		0		1	0
Solution7	929 to 1084	1				
	1085 to 1220		1			
	1221 to 1332				1	
Setup8	1085 to 1220		1		0	
	1221 to 1332		0		1	0
	1333 to 1485			0		1
Solution8	1085 to 1223		1			
	1224 to 1339				1	
	1340 to 1485					1

Elected KNITRO^{tm} Solver Parameters

- Solution Method
  - As we have indicated, there are three solution methods: The direct and conjugate gradient interior point methods, and the active set method. The software's default setting is to allow the software itself to elect the best method at each stage in the process. Alternatively, the user can specify which of the three methods is to be used.
  - We elected to keep the default setting. As we will discuss later, there were two instances where we had to modify our reliance on the default and make use of a particular solution method.
- Global Optimization of non-convex problems
  - Finding a global optimum is not usually guaranteed.
  - Sometimes it can be guaranteed in integer programming problems, but usually it would take to long to arrive at a guaranteed solution.
  - As a result, additional measures should be taken to make it likely that a good solution near the global optimum is arrived at:
    - Multi-Start Search
    - Topographic Search
    - We elected to use both of these features

Elected KNITRO^{tm} Solver Parameters

- Automatic Scaling
  - Poor scaling can reduce precision. Selecting automatic scaling can in some instances help. But an effort to properly scale the problem should be made.
    - We elected to use the automatic scaling feature.
- Derivatives
  - The interior point methods work best when they can use analytic second derivatives.
  - The software could not find solutions to the second derivatives; perhaps because of the absolute value in our objective function.
  - In this instance the software offers the option of using analytic first derivatives or finite differences
    - We elected the analytic first derivative option
- Sparse Optimization
  - Our problems are large. Using this option on sparse problems can save time. The software indicted our problem was sparse.
    - We elected to use sparse optimization option

Elected KNITRO^{tm} Solver Parameters

- Integer Tolerance
  - When solving integer programming problems, branch & bound can solve to a pre-determined level of tolerance from true integer values, when testing for optimality.
  - The default setting is 0.05, which we did not change.
  - If one were to select “0”, it is possible that the software could arrive at a guaranteed global optimal solution, although it might take quite a while.
- Remaining Parameters
  - The remaining parameters were of less importance.
  - We elected the default settings in all remaining parameters.

Interior Point Methods with Branch & Bound

- There can be a problem with using interior point methods in combination with the branch & bound technique.
- The interior point methods can constrain the problem too tightly for the branch & bound to find a feasible solution.
- This is a danger when electing the default solution method in an integer programming problem, as we did, or when electing one of the specific interior point methods.
- We ran into this problem twice when conducting the cluster analysis for property damage liability severity

Interior Point Methods with Branch & Bound

- In our third problem setup for PD severity, repeated attempts resulted in failure to find feasible solution
  - In response we specifically elected the active set methodology.
    - Using this method, the algorithm ran much longer than we had ever encountered for our reduced-sized problems.
    - We could see that each iteration was bringing slight progress.
    - At this point, we elected to stop the process, leaving the interim solution in place.
    - Then we reran the problem, with the active set interim solution in place as an initial solution, and again elected the default setting which allows the software to pick which method to use at each step in the process.
    - The software then found a solution in a reasonable amount of time.

Interior Point Methods with Branch & Bound

- The problem repeated itself on the eighth and final setup for PD severity.
  - We repeated the same procedure we used before, except that we did not allow the active set method run so long before stopping and using the interim solution.

Agenda

- Introduction
- Risk Classification Challenges to Territory Analysis
- Homogeneity vs. Credibility
- Mixed Model Approach
- Mixed Model Component 2: Arithmetic Model
- Mixed Model Component 3: Proximity Complement
- Mixed Model Results
- Constrained Cluster Analysis
- **Final Results**
- Future Research
- Conclusions

Final Results

- Detailed information for BI frequency, PD frequency, BI severity and PD severity have all been placed on the CAS website
  - ○ Mixed model components
  - ○ Credibility assigned to each component
  - ○ Mixed model estimate
  - ○ A comparison of the new band assignment with the Frequency and Severity Band Manual Assignment

For BI frequency the Hunstad assignments modestly outperform mixed models with clustering. The mixed model outperforms the Hunstad result for bands 1 and 10, with results for the first band significantly better.

BI Frequency

	FB1	FB2	FB3	FB4	FB5	FB6	FB7	FB8	FB9	FB10
					Relatvities					
Mixed Model	0.5438	0.6180	0.6730	0.7253	0.7866	0.8602	0.9870	1.1386	1.3374	1.7544
Actual	0.4895	0.5775	0.6589	0.7232	0.7882	0.8619	0.9940	1.1488	1.3472	1.7708
Hunstad	0.5334	0.6715	0.7456	0.8037	0.8767	0.9795	1.0752	1.1856	1.3425	1.7393
					MAD					
New Cell	0.00105	0.00092	0.00047	0.00039	0.00037	0.00045	0.00058	0.00071	0.00109	0.00315
Hunstad Cell	0.00121	0.00041	0.00034	0.00029	0.00048	0.00035	0.00052	0.00052	0.00086	0.00319
New Total					0.00087					
Hunstad Total					0.00083					

For PD frequency, mixed models with clustering moderately outperformed the Hunstad result. Our approach again outperformed for bands 1 and 10.

PD Frequency

	FB1	FB2	FB3	FB4	FB5	FB6	FB7	FB8	FB9	FB10
					Relatvities					
Mixed Model	0.6548	0.7265	0.7853	0.8423	0.9171	0.9663	1.0127	1.0598	1.1247	1.3036
Actual	0.6132	0.7137	0.7827	0.8423	0.9173	0.9671	1.0140	1.0613	1.1271	1.3102
Hunstad	0.7301	0.8634	0.9297	0.9642	0.9965	1.0219	1.0492	1.0740	1.1117	1.2430
					MAD					
New Cell	0.00223	0.00094	0.00081	0.00074	0.00067	0.00049	0.00047	0.00044	0.00114	0.00299
Hunstad Cell	0.00261	0.00129	0.00048	0.00042	0.00030	0.00027	0.00027	0.00029	0.00060	0.00318
New Total					0.00082					
Hunstad Total					0.00097					

For BI severity, our approach significantly outperformed the Hunstad result, and again outperformed in bands 1 and 10.

BI Severity

	SB1	SB2	SB3	SB4	SB5	SB6	SB7	SB8	SB9	SB10
					Relatvities					
Mixed Model	0.8297	0.8777	0.9026	0.9267	0.9499	0.9805	1.0136	1.0422	1.0761	1.1268
Actual	0.8224	0.8728	0.8985	0.9253	0.9508	0.9833	1.0154	1.0427	1.0765	1.1293
Hunstad	0.8380	0.8902	0.9202	0.9525	0.9792	1.0049	1.0232	1.0445	1.0675	1.1156
					MAD					
New Cell	207.61	129.62	91.92	87.93	87.16	124.18	90.86	92.81	100.82	206.48
Hunstad Cell	229.64	100.22	113.12	158.01	210.82	171.97	139.16	144.30	145.46	243.90
New Total					117.85					
Hunstad Total					168.71					

For PD severity, our approach moderately outperformed the Hunstad result, and again outperformed for bands 1 and 10.

PD Severity

	SB1	SB2	SB3	SB4	SB5	SB6	SB7	SB8	SB9	SB10
					Relatvities					
Mixed Model	0.8387	0.8770	0.9078	0.9346	0.9615	0.9905	1.0181	1.0423	1.0803	1.1487
Actual	0.8355	0.8755	0.9076	0.9349	0.9625	0.9909	1.0181	1.0421	1.0807	1.1503
Hunstad	0.8505	0.8989	0.9406	0.9771	0.9983	1.0155	1.0283	1.0449	1.0700	1.1303
					MAD					
New Cell	28.94	11.79	11.40	12.11	13.73	12.68	8.33	9.46	19.58	35.53
Hunstad Cell	29.83	18.28	20.34	12.95	10.06	5.18	7.54	8.00	14.25	42.84
New Total					14.67					
Hunstad Total					17.01					

Agenda

- Introduction
- Risk Classification Challenges to Territory Analysis
- Homogeneity vs. Credibility
- Mixed Model Approach
- Mixed Model Component 2: Arithmetic Model
- Mixed Model Component 3: Proximity Complement
- Mixed Model Results
- Constrained Cluster Analysis
- Final Results
- **Future Research**
- Conclusions

Directions for Future Research

- Within the existing framework of Territory Analysis:
  - Refinement of the Arithmetic Model
  - Refinement of the Proximity Complement
  - Refinement of the Credibility Weighting Scheme
  - Refinement & Automation of Constrained Cluster Analysis
- Development of new Territory Analysis framework:
  - Introduction of New Geographical Rating Variables
  - Integrate Territory Analysis with parameterization of remaining Class Plan.
- Refinements to California Personal Automobile Ratemaking
  - Updated Frequency and Severity Bands Manual and Data
  - Constrained Cluster Analysis in lieu of Pumping and Tempering
  - Progressively supplant relative frequency and severity with new causal geographical variables to further achieve goals of Prop 103

Refinement of the Arithmetic Model

- Identify better and new causal geographical variable formulations.
- Introduction of a handful of binary geographical variables could substantially improve the result
- Spatially autoregressive model

Refinement of the Proximity Complement

- Methods of selecting elements of the complement:
  - Immediately contiguous complements (Tang)
  - Hierarchical cluster analysis with overlapping clusters
- Methods of weighting elements of complement:
  - For example, weight so that population or exposure weighted latitude and longitude (ideally at the census block level) for complement equals that of the atomic geographical unit (zip code) being complemented
  - Weight by distance from geographical unit being complemented (Hunstad suggestion)
  - Incorporate spline or graduation information into a proximity complement
- Use spatially autoregressive model (without all the independent variables) to generate values of proximity complement elements

Refinement of the Credibility Weighting Scheme

- Ideally the local geographical fit should influence the weight for both the arithmetic model and the proximity complement
- And the credibility weight for the zip code indication itself should be relative rather than the absolute 1,082 claim rule.
- More formal mixed model. Searle et al 1992, Rao 1997, etc.

Refinement of Constrained Cluster Analysis

- Alternative Method of Constrained Cluster Analysis should be investigated.
  - One of the other large-scale solver engines distributed by Frontline would appear to be especially applicable to our problem.
  - Large-Scale SQP^tm (Sequential Quadratic Programming) Solver
    - We did not have luck with this one in our initial experiments.
    - However, SQP^tm supports a special form of analysis that is particularly applicable to our problem.
      - Special Ordered Set (SOS) involve binary decision variables arrayed like ours and constrained via a system like our (2.4).
      - Introduced in Beale and Tomlin (1969) [54]
      - Would be particularly relevant if one were to increase the size of the problem, reducing or eliminating the sequential procedure of solving the problem in pieces which we have developed here.

Automation of Constrained Cluster Analysis

- Automate the Sequential Procedure
  - We used the plug-in to a spreadsheet, because this allows for a more interactive approach where one can experiment and learn with simpler setups.
  - After the procedure becomes a little more well established, it could probably be completely automated via one of the other implementations of Frontline.
  - When so automated, the Constrained Cluster Analysis would be incredibly efficient, dramatically improving the productivity of those involved in large-scale territorial revisions for many states

Introduction of New Geographical Rating Variables

- Traffic Density
  - Well Accepted.
  - Quantitative so can facilitate integration of Territory Analysis
  - Challenge is to find acceptable measure, which must incorporate spatial interaction.
    - In competitive markets, there will be the obvious incentives to come up with good measures.
    - In heavily regulated markets, regulators should come up with a measure or with the criterion for deriving an acceptable measure.
    - New information from mobile position-aware devices and remote sensing may soon allow for extremely accurate measurement.
    - Demand for Workers versus Supply of Commuters

The Introduction of New Causal Geographical Rating Variables

- Traffic Enforcement
  - It is commonly accepted that increased enforcement reduces accidents
  - Phase II (1979) [20] constructed an enforcement ratio measure.
    - The measure is somewhat defective in that it measures the relationship between injury accidents and all driving incidents.
      - Since that time it has become increasingly recognized that the frequency of injury accidents is heavily influenced by claims environment.
  - Would make sense to re-investigate enforcement using property damage liability accidents in lieu of bodily injury liability accidents
  - Should a loss preventive effect be measured using an accurate enforcement measure, the rationale for introduction as rating variable would be extremely powerful.
    - economic incentives for actions that reduce the number of accidents
  - All of the data necessary to conduct such a study using property damage liability accidents is available in the appendices of the Phase II study.
  - Even more ideal would be release of more recent data set from the California DMV which was the original source for Phase II.

The Introduction of New Causal Geographical Rating Variables

- Legal Environment
  - Legal or claims environment might only be a good candidate for introduction in heavily regulated jurisdictions after several other causal geographical variables have successfully been introduced
  - Improved measures of lawyer density
  - Binary geographical rating variables that correspond to court jurisdictions
    - How would spatial interaction be reflected?
      - Mobility of vehicles
      - And choice of venue relatively flexible for auto liability.

The Introduction of New Causal Geographical Rating Variables

- Medical and Repair Cost Indices
  - Influence on loss costs is probably less than that of the other variables we have identified.
  - But probably an uncontroversial variable candidate
  - So if relationship between acceptable indices and severity can be demonstrated, acceptability likely.
- Integrate new causal geographical rating variables into GLM or other predictive model used to parameterize remaining classification plan

Refinements to California Personal Automobile Ratemaking

- A New Frequency and Severity Bands Manual For California
  - Updated with the release of more recent data from the same source, such as was used in Tang (2005)
  - The use of a mixed model technique, or Tang's new proximity complement might be in order
  - Or, the new data could be provided without a new manual.

Refinements to California Personal Automobile Ratemaking

- An Alternative to Pumping and Tempering in California
  - Pumping and Tempering
    - courts have criticized this procedure as arbitrary
  - Introduce factor weight as a constraint in the Cluster Analysis procedure
  - An investigational attempt to implement this form of constraint would be of interest

$$\frac{\sum_i \sum_j \left[abs\left(\frac{\sum_b x_{bj} R_b E_b}{\sum_c E_c x_{cj}} - 1 \right) x_{ij} E_i \right]}{\sum_d E_d} \leq M \qquad (3.9)$$

Refinements to California Personal Automobile Ratemaking

- New Causal Geographical Rating Variables in California
  - The introduction of causal geographical rating variables, combined with reductions in the scope of relative frequency and severity would improve accuracy and further achieve the objectives of Proposition 103
  - To see why, let's review Prop 103 and its origins.

Refinements to California Personal Automobile Ratemaking

- Intellectual Underpinnings of Proposition 103
  - Casey et al. (1976) [26]
  - Shayer (1978) [34]
  - Ferreira (1978a) [28]
  - Ferreira (1978b) [29]
  - Chang & Fairley (1978) [27]
  - Stone (1978) [35]
  - Phase I (1978) [19]
  - Phase II (1979) [20]

Refinements to California Personal Automobile Ratemaking

- Central argument against territorial rating by Prop 103s precursors
  - Not a causal variable
    - Introducing variables that the authors of precursor papers themselves recognized as causal is a means of eliminating this objection
    - Causality appeared determinative for Shayer for similarly situated variable.
  - Subjective / arbitrary procedures in grouping
    - Cluster Analysis is a means of eliminating this objection

Refinements to California Personal Automobile Ratemaking

- Procedure for introducing new causal geographical variables
  - The California Insurance Commissioner has the power to introduce new rating variables that have been demonstrated to have a "substantial relationship to the risk of loss."
  - Currently, two such geographical rating variables exist – relative claims frequency and relative claims severity
  - As causal geographical variables are introduced, the more "undesirable" geographical variation in frequency and severity, with no known cause, would be captured in the relative frequency and severity bands.
  - Sequential analysis of new variables would seem to be easy enough
    - Could occur after all other variables but before relative frequency and severity
  - Allowed scope of relative frequency and severity could be reduced as new causal geographical variables are introduced

Refinements to California Personal Automobile Ratemaking

- Traffic Density
  - Recognized as a causal variable for at least 90 years; even by critics of territory as a rating variable.
- Traffic Enforcement
  - CDI itself investigated this as causal variable in Phase II via enforcement ratio
  - Do another study of enforcement ratio.
    - Reconfigure, using PD liability accidents in lieu of injury accidents.
    - Data necessary for study is contained in appendices of Phase II.
    - Or new data of similar nature could be taken from DMV
  - Powerful loss prevention argument for variable if it can be shown to influence losses
  - Assign enforcement ratio for each zip code every year or so. Conduct sequential analysis against that enforcement ratio.
  - Enforcement ratio already reflects spatial interaction
- Medical and Repair Cost Indices
  - Arrive at acceptable granular indices and test relationship to severity
  - Causality would be clear. Uncontroversial candidate for introduction as variable.

Agenda

- Introduction
- Risk Classification Challenges to Territory Analysis
- Homogeneity vs. Credibility
- Mixed Model Approach
- Mixed Model Component 2: Arithmetic Model
- Mixed Model Component 3: Proximity Complement
- Mixed Model Results
- Constrained Cluster Analysis
- Final Results
- Future Research
- **Conclusions**

Conclusions

- Our mixed model with clustering approach to Territory Analysis, which is entirely objective, generally outperformed the existing Proposition 103 California Frequency and Severity Band Manual in terms of mean absolute deviation. This is impressive because the implementation of the new concept was rudimentary.

Conclusions

- Significant further work can be done on improving each of the elements of the mixed model, which would substantially improve the accuracy of the result.

Conclusions

- And after the method is fine tuned and has matured, it would be a relatively easy matter to automate the sequential piecewise procedure employed in the Cluster Analysis. In that format, the approach could become extremely efficient, relative to the manual procedures currently involved when extensive territorial refinements are conducted.

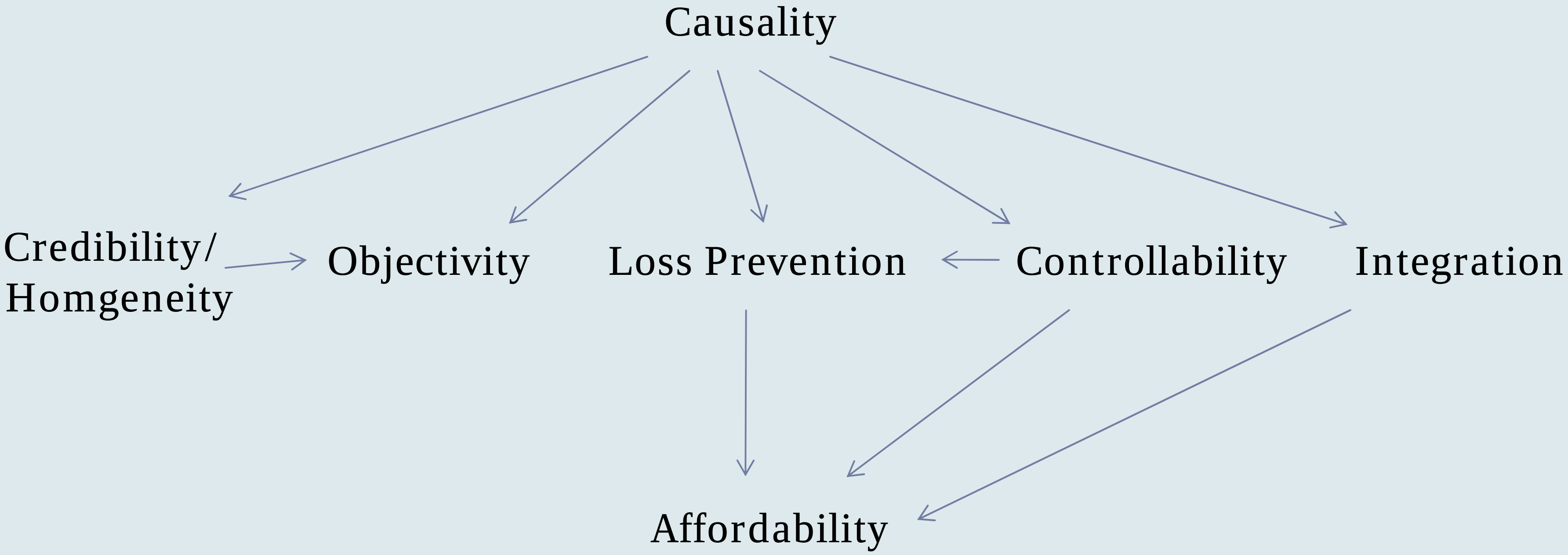
Conclusions

- The causal analysis of geographical variation in loss costs which could ensue from our approach could pave the way for the introduction of new causal geographical rating variables.

Conclusions

- In addition to eliminating criticisms regarding causality and potentially invigorating local loss prevention initiatives, this group of largely continuous variables could be integrated with the parameterization of the remaining classification plan via the extensive array of predictive modeling procedures that are being employed for that purpose.

Conclusions



Causality

Credibility/ Homgeneity

Objectivity

Loss Prevention

Controllability

Integration

Affordability

- Questions?
- Thank You