The Hanover Insurance Group®

Cutting Edge Tools for Pricing and Underwriting Seminar

Integrating External Data into the Decision Making / Predictive Modeling Process

Casualty Actuarial Society
Ron Zaleski, Jr., FCAS, MAAA
Fall 2011

CAS Anti-Trust Notice

- The Casualty Actuarial Society is committed to adhering strictly to the letter and spirit of the antitrust laws. Seminars conducted under the auspices of the CAS are designed solely to provide a forum for the expression of various points of view on topics described in the programs or agendas for such meetings.

- Under no circumstances shall CAS seminars be used as a means for competing companies or firms to reach any understanding – expressed or implied – that restricts competition or in any way impairs the ability of members to exercise independent business judgment regarding matters affecting competition.

- It is the responsibility of all seminar participants to be aware of antitrust regulations, to prevent any written or verbal discussions that appear to violate these laws, and to adhere in every respect to the CAS antitrust compliance policy.

The Hanover: About Us

- Property and Casualty Insurance Company
- Founded over 150 years ago
- Among the largest property and casualty companies with revenues of $2.8+ billion
- Best of both national and regional companies
- The Boston Globe named us the #1 publicly traded financial services business in Massachusetts
- Both The Boston Globe and Business Insurance named us to their list of 2010 Best Places to Work

AGENDA

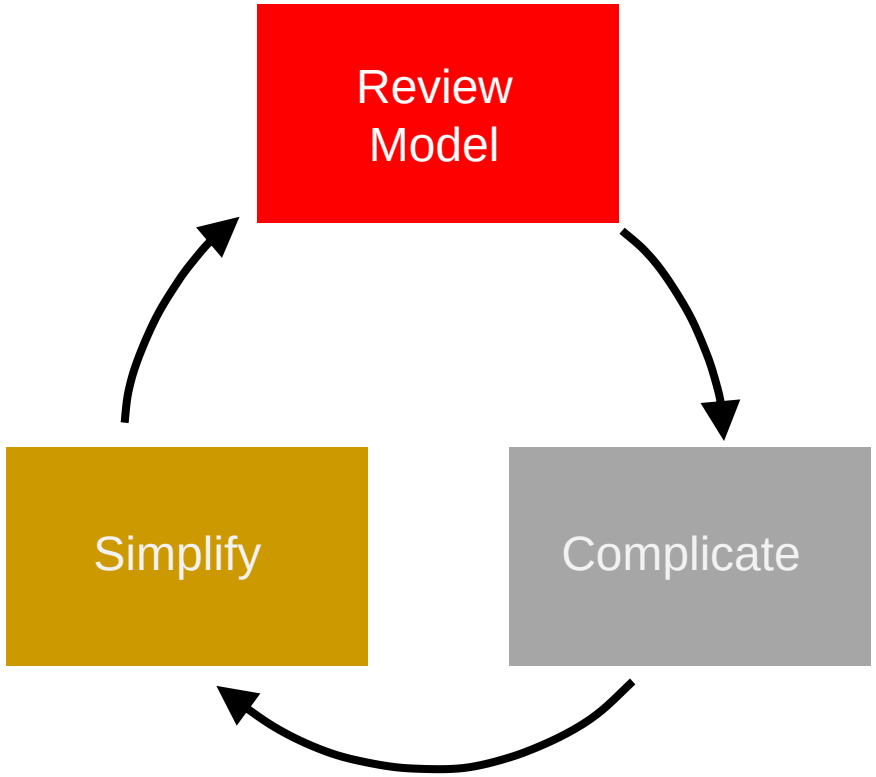
- **Background**

- **Case Study: Territory Definitions & Factors**
  - Selecting & Handling External Data
  - Incorporating Competitor Data
  - Supplementing with Industry Data

- **Summary**

Reminder

- Modeling is an iterative process
- How does the analyst decide which factors are most valuable?
  - Parameters/standard errors
  - Consistency of patterns over time or random data sets
  - Type III statistical tests (e.g., chi-square tests)
  - Judgment (e.g., do the trends make sense)
- Focus of the section is on gathering data NOT analysis



This presentation will focus on ways to select external data for modeling and evaluation of a territory project

Case Study:
PL Auto Territory Development

- Select analytical basis and approach
  - Geographic Unit: i.e. Census Tract
  - Target Variable: i.e. Loss Ratio ex. Territory Factors
  - Modeling Approach: i.e. GLM w. Spatial Smoothing
- Develop internal data
  - Experience data (exposures, premiums, losses)
  - Existing rating plan variables and derivations
- Identify and incorporate any external data, if needed
  - Measures that describe geographic unit to be used in the model
  - Supporting data to guide modeling effort and inform final decision process, especially where internal data is thin

FOCUS

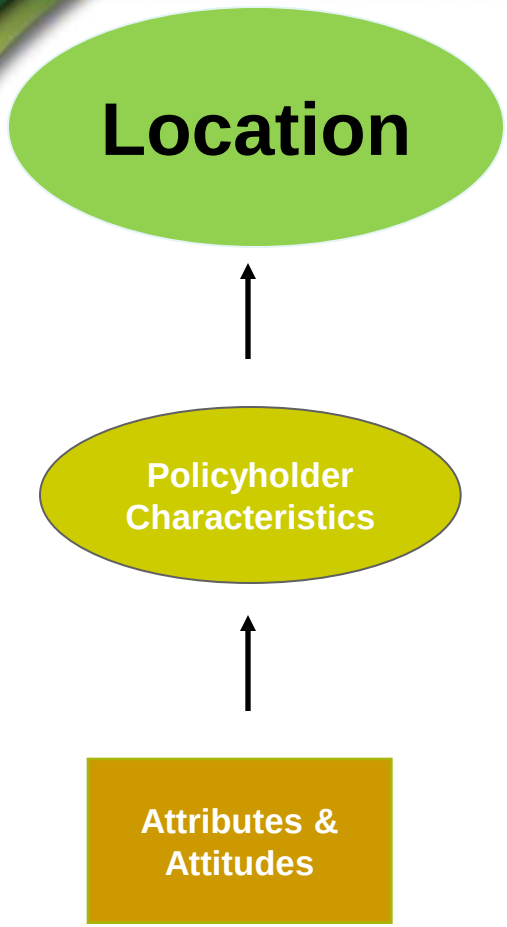
Questions Addressed

Location Proxy Data	What types of data can we use to represent geographic units in a model framework?
Credibility	How can we utilize external information to provide ballast when our internal data is thin or non-existent?
Competitor Analytics	How can we indentify the appropriate competitor data to use in the decision-making process?

AGENDA

- **Background**
- **Case Study: Territory Definitions & Factors**
  - Selecting & Handling External Data
  - Incorporating Competitor Data
  - Supplementing with Industry Data
- **Summary**

External Data: Location



Goal: Append external data that represents similarity between geographic units beyond proximity

Location-Proxy Variables

- U.S. Census Data (Demographics)
- Traffic Statistics (NHTSA)
- Other data providers, such as EASI

Competitor Information

- Rate Filing Research
- Competitor Rate Engines(InsurQuote / Quadrant)

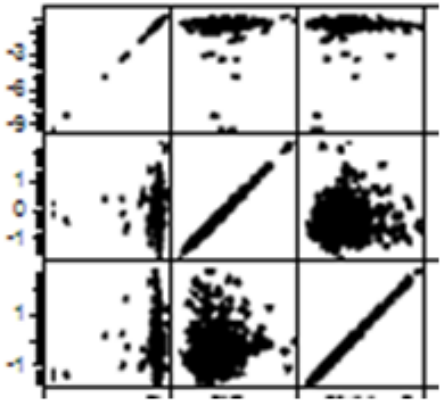
Industry Data

- ISO Data Cubes
- IIHS/HLDI Data

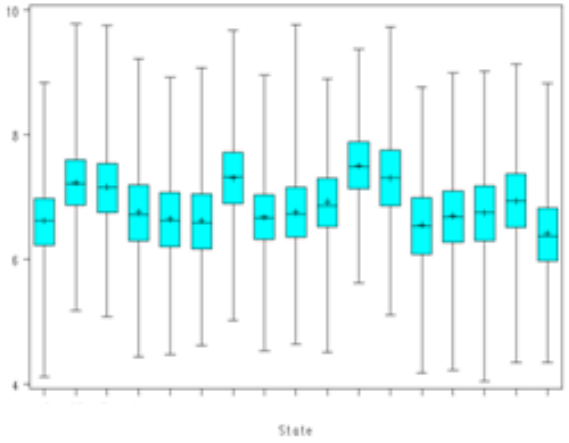
External Data:
Variable Inspection

- After appending external data, spend time with exploratory analyses to understand relationships between variables
  - Correlation Tests, such as Cramer's V
  - X-by-X plots (Unsupervised), such as Scatter Plots, Box-Whisker and 2-Way Plots to detect patterns

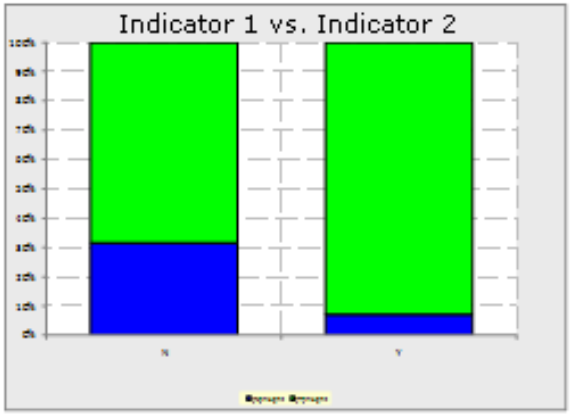
X-X Scatterplot



Box-Whisker



Two-Way Plot



Software Snapshots: JMP, SAS, and EMBLEM

External Data: Dealing with Correlation

- Principal Components Analysis
  - Unsupervised learning technique that seeks to explain the variance in the X's
  - Reduces a large number of continuous variables into a manageable smaller set that are a linearly independent, linear combination of the underlying larger set set
- Partial Least Squares
  - Similar to PCA, except the technique is supervised learning, seeking to maximize the covariance between the X's and the dependent Y
  - The advantage is that the PLS variables are extracted in order of importance based on relationship to the target (not each other)
  - The disadvantage is that it is supervised and therefore the outcome depends on the target variable.

AGENDA

- **Background**
- **Case Study: Territory Definitions & Factors**
  - Selecting & Handling External Data
  - Incorporating Competitor Data
  - Supplementing with Industry Data
- **Summary**

Competitor Territory Models

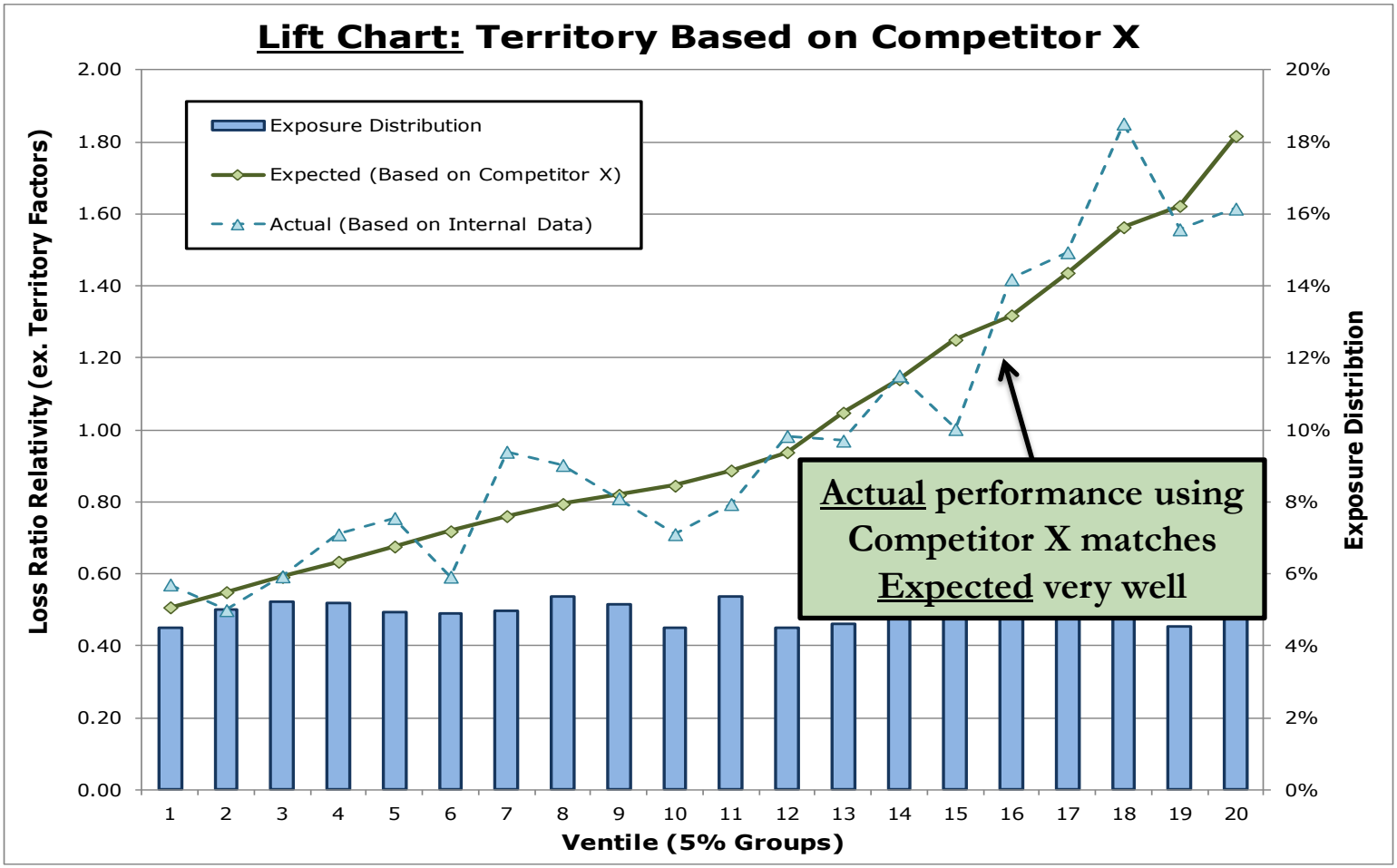
- At other companies, actuaries are selecting territory definitions and factors, too…
- They're performing similar analyses on the same metrics…
- They're working on another sample of the population…
- So let's view these territories as competing models to ours!

This section will cover the following:

- How can we identify the best competitor model for comparison?
- How can we use the competitors' territories directly in our analysis to strengthen predictions?

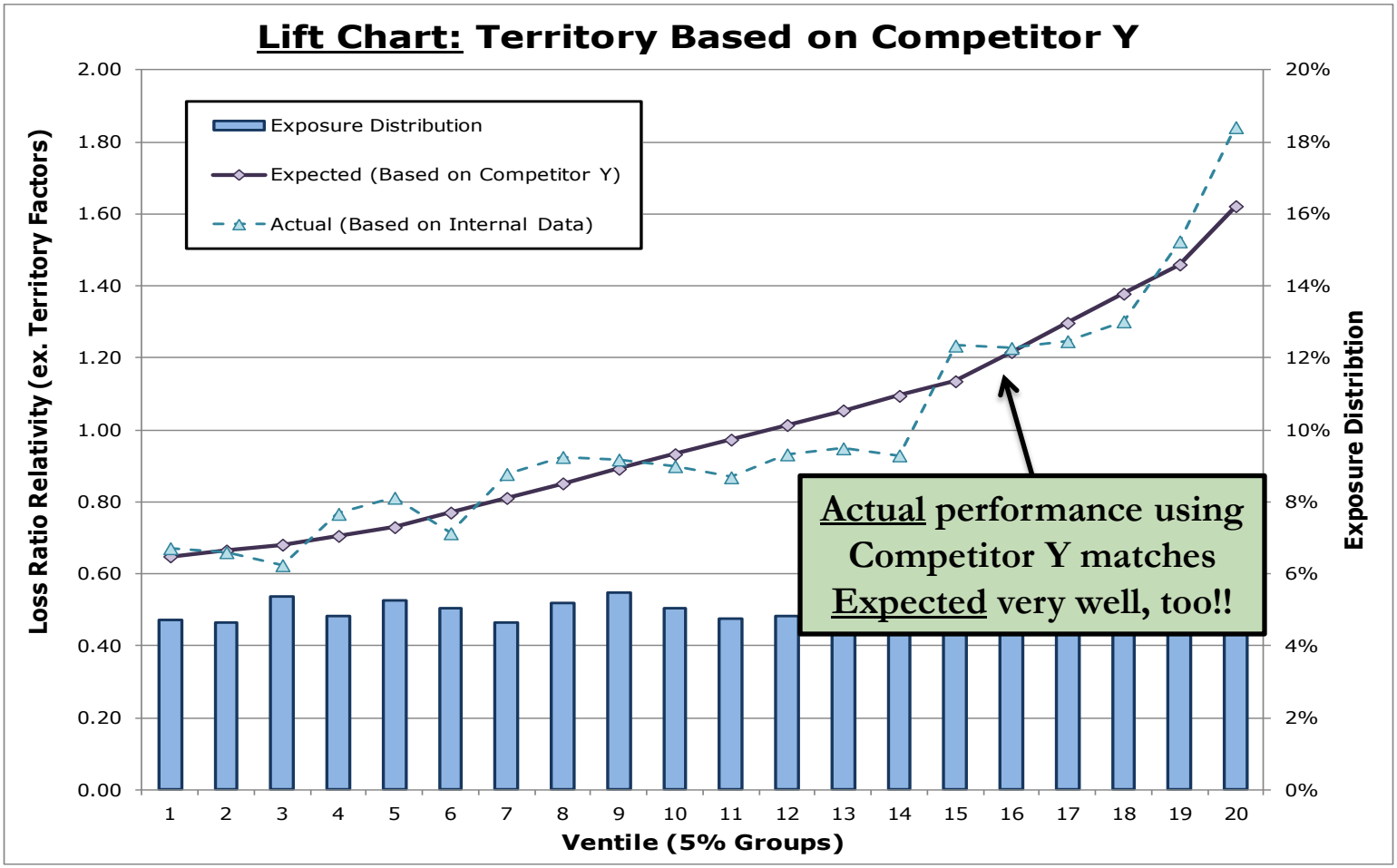
Competitor Evaluation:
Lift Charts

Using a traditional model evaluation technique, such as a lift chart, you can judge the appropriateness of a competitor's territory on your own data.



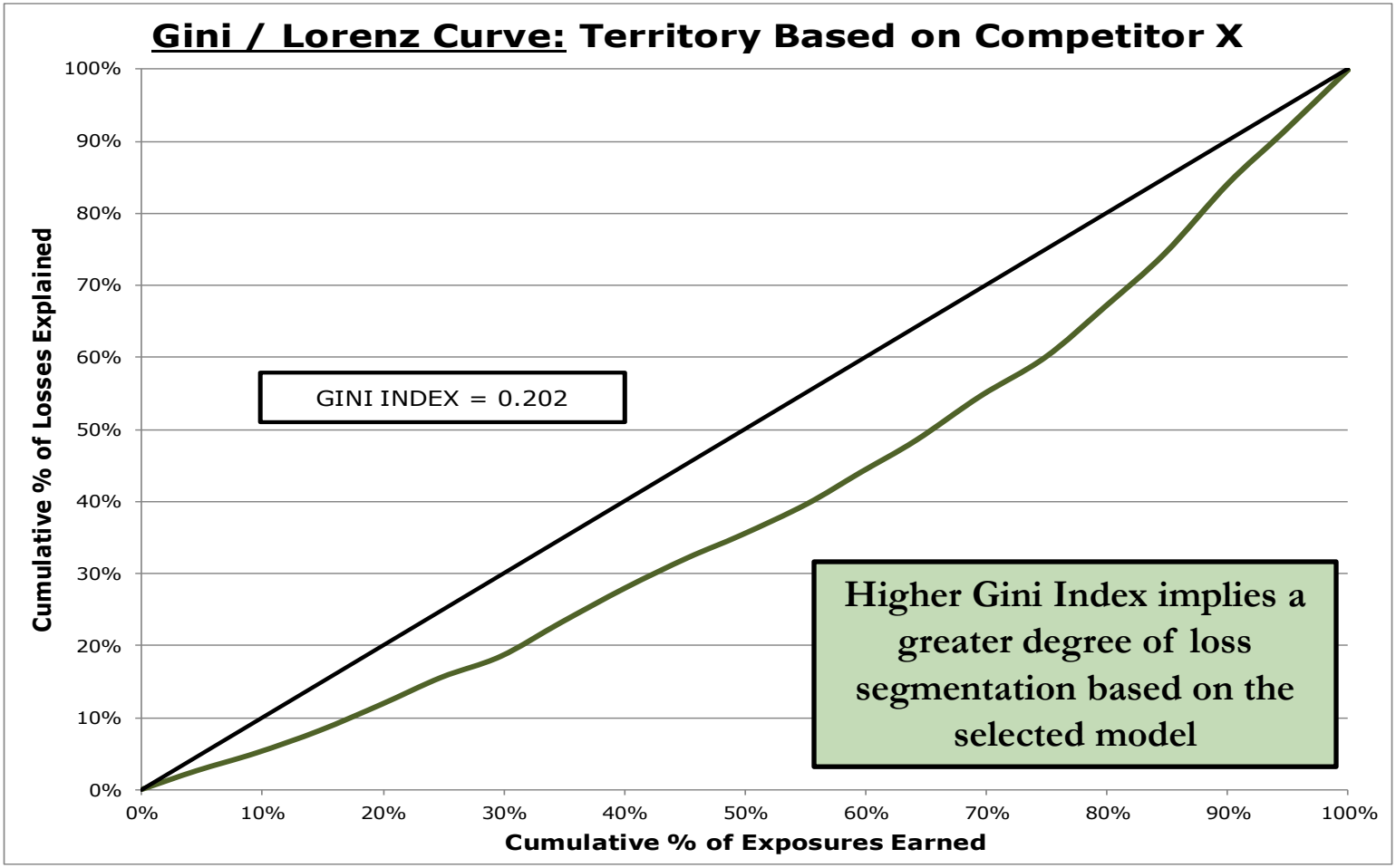
Competitor Evaluation:
Comparing Lift Charts

But what happens when a second competitor looks just as good?



Competitor Evaluation:
Lorenz/Gini Curve

An alternative view is to use a Lorenz curve and calculate a Gini Index to provide a quantitative measure to compare two models



Competitor Evaluation:
Ranking by Gini

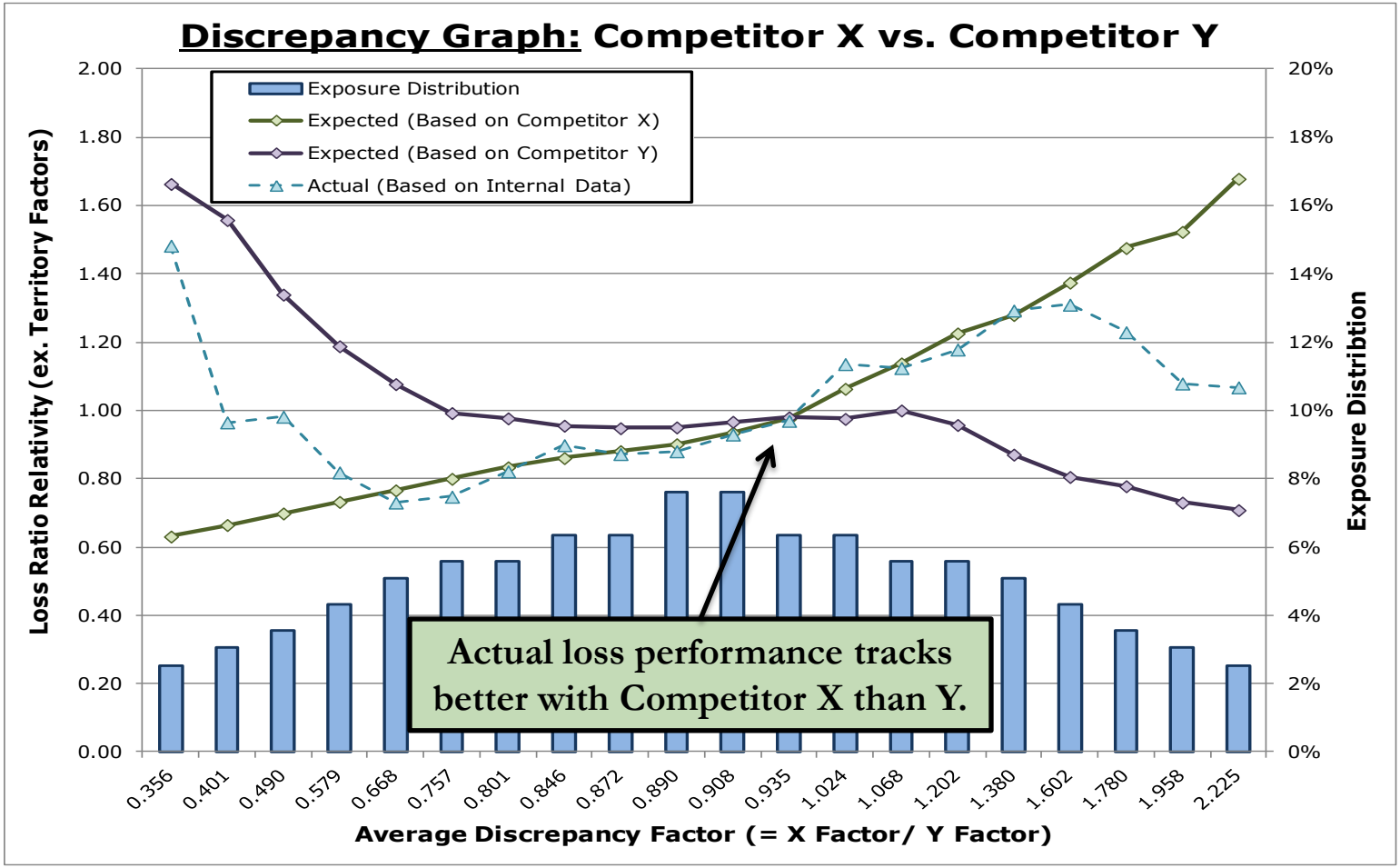
Ranking the performance of each of the Competitor Models by Gini Index will help guide your selection.

Competitor Name	Gini Index
Competitor X	0.202
Competitor Y	0.160
Competitor Z	0.084
Competitor W	0.080
Competitor U	0.064
Competitor V	0.056

Takeaway: Using quantitative measures, such as the Gini Index, makes determining the "best" model easier

Competitor Evaluation: The "Playoffs"

Another alternative visual comparison is the discrepancy or "X" graph that compares models against each other.



Competitor Territories:
Integration into Decision-Making

So Competitor X seems to perform best… now what?

Model Development

- Incorporate factors directly as variables in model
- Perform correlation analysis to identify other potential predictive variables

Benchmarks

- Consider Competitor X statistics, such as Gini, as minimum performance standards
- Compare models using Discrepancy "X" graphs

Credibility Complements

- Competitor X is determined to be the best competitor complement
- Integrate discrepancy in spatial/residual smoothing

AGENDA

- **Background**
- **Case Study: Territory Definitions & Factors**
  - Selecting & Handling External Data
  - Incorporating Competitor Data
  - Supplementing with Industry Data
- **Summary**

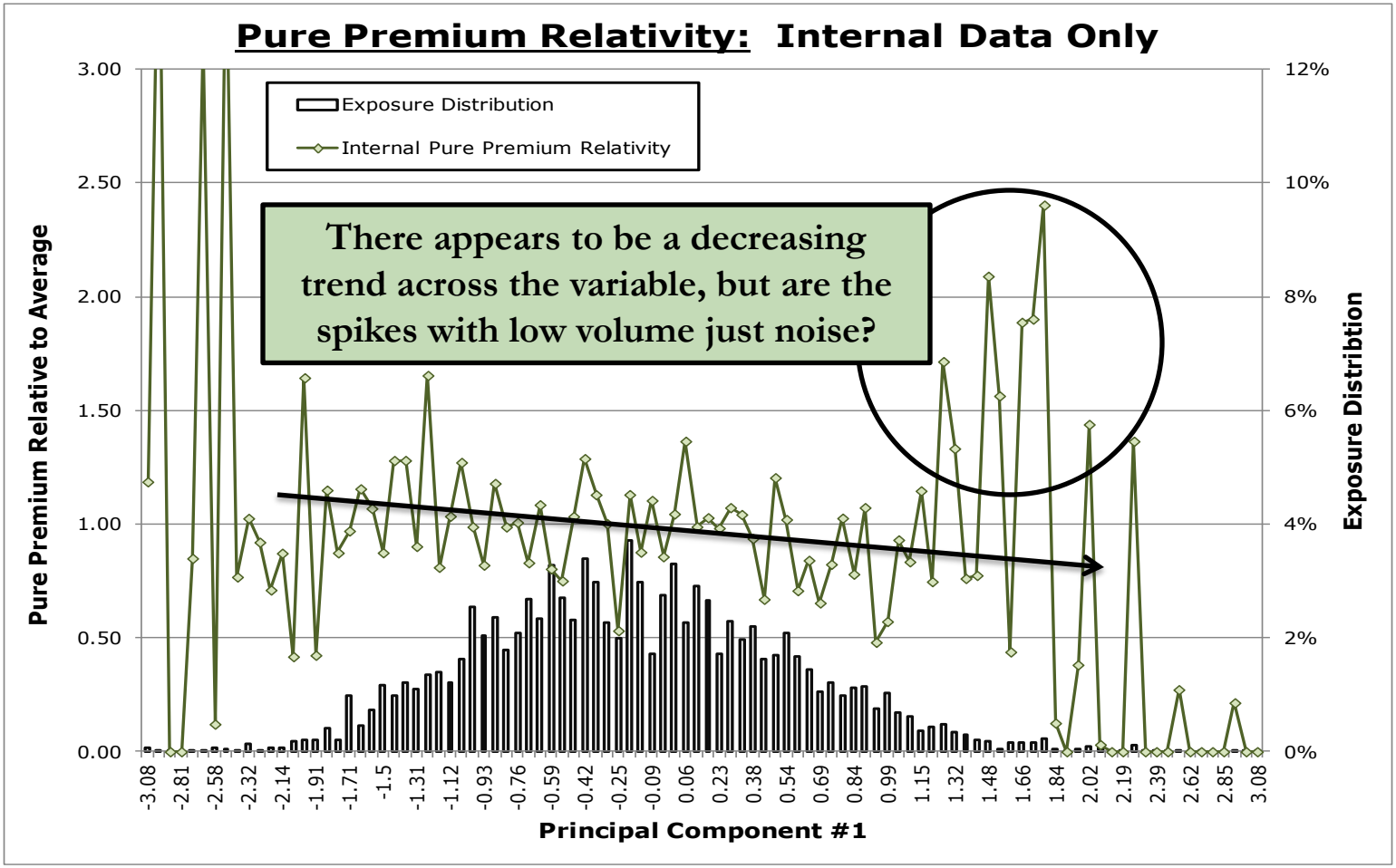
Industry Data

- In dealing with the common problem of low data volume, we constantly look for ways to supplement analysis to deal with the credibility in separating the signal form the noise

- A data source, such as HLDI, can provide both an early indication of which external variables are important as well as helping to detect the underlying signal in a noisy process

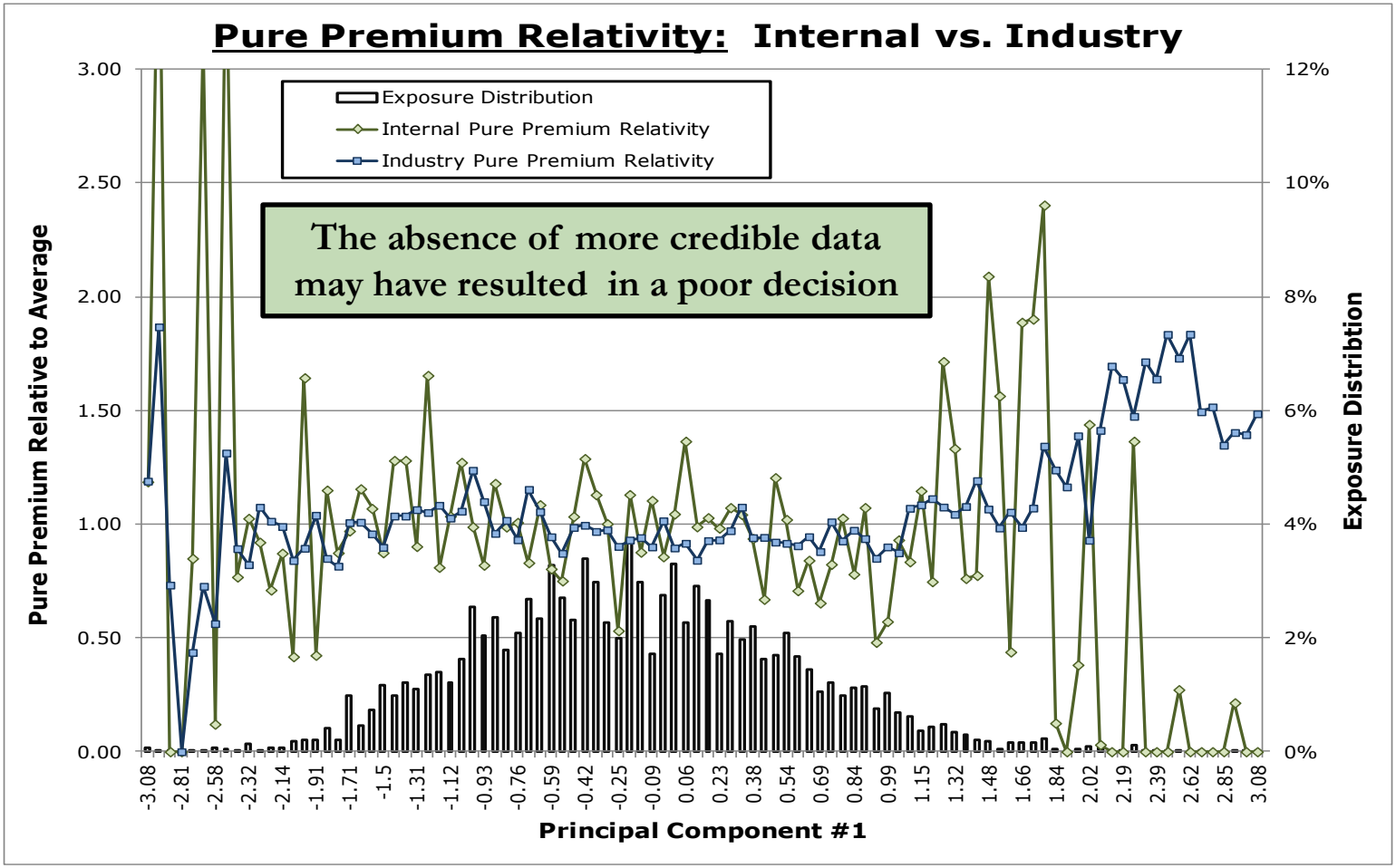
Identifying Signal:
Bring on the Noise…

Detection of signal is especially difficult where the data is thin. Consider the analysis of Principal Component below.



Identifying Signal:
The Advantage of Large Data…

More data helps to flatten "noisy" spikes and reveal the true signal, especially where internal data is thin. It is also a good sign when trends are consistent.



Identifying Signal: Suggested Approach

- Append external variables and transformations onto the industry data and build (partial) predictive models
  - Advantages:
    - More data (rows) = clearer signal; reduced noise pollution
    - Allows you to test sampling variance on a larger population
  - Disadvantages:
    - Less data (columns) = Difficult to reflect other class plan variables not available on the industry data to avoid OVB (Omitted Variable Bias)
    - Could be considering "cheating" from a sampling perspective since your data/signal may be included in the dataset

Final Result

- Industry Territory model that can be included in modeling dataset
- Modeling "Guide" -- Lists of variables with an importance measure for each coverage analyzed:

Parameter Estimates on Industry Data

Variable	Coverage A	Coverage B	Coverage C	Coverage D
Principle Component 1	0.140	0.410	0.230	
Principle Component 2	0.840	0.590	0.320	0.690
Principle Component 3	0.680	0.400		
Principle Component 4	**1.370**	0.070	**1.170**	**3.860**
Principle Component 5	**1.060**	0.160	0.740	
Principle Component 6	0.270		0.380	

Remember! The results on industry data could suffer from OVB, capturing signal that your underlying class plan would have already accounted for

AGENDA

- **Background**
- **Case Study: Territory Definitions & Factors**
  - Selecting & Handling External Data
  - Incorporating Competitor Data
  - Supplementing with Industry Data
- **Summary**

SUMMARY

- The foundation of every modeling project is developing the right data for the task
- External data is readily available and provides not only useful predictors, but also validation and benchmarks
- Creatively using competitor and industry data can strengthen the final decision by lending additional credibility to underlying data and providing another modeler's opinion.

Contact Details

Ronald Zaleski, Jr., FCAS, MAAA
The Hanover Insurance Group
AVP – Actuary
Personal Lines Research & Development
440 Lincoln Street
Worcester, MA 01653

(508) 855-8121
rzaleski@hanover.com

Useful Links

Census Data:	http://www.census.gov/
NHTSA:	http://www.nhtsa.gov/
EASI:	http://www.easidemographics.com/
IIHS-HLDI:	http://www.iihs.org/
ISO:	http://www.iso.com/