Modeling and Analytics in a UBI Setting

CAS Ratemaking and Product Management (RPM) Seminar

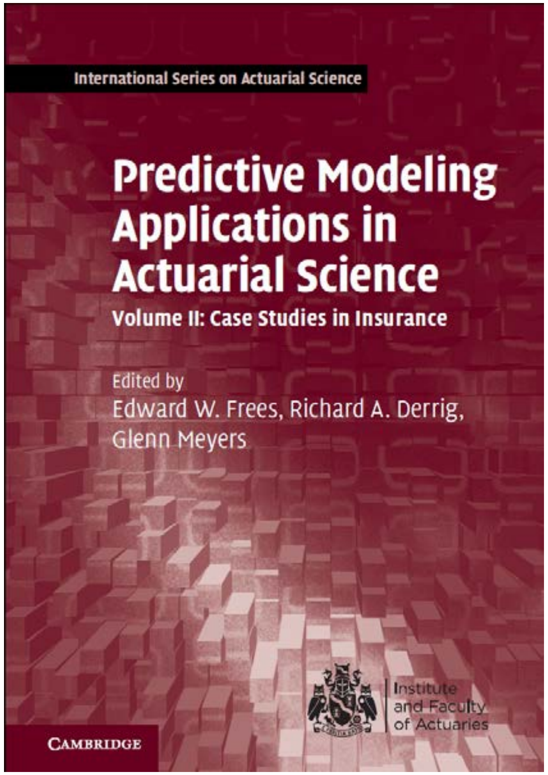
March 28, 2017 ● San Diego, CA

Verisk Analytics

SERVE | ADD VALUE | INNOVATE

State Farm®

Now available



International Series on Actuarial Science

Predictive Modeling Applications in Actuarial Science

Volume II: Case Studies in Insurance

Edited by
Edward W. Frees, Richard A. Derrig,
Glenn Meyers

Institute and Faculty of Actuaries

CAMBRIDGE

Chapter 11: Predictive Modeling for Usage-Based Insurance (Makov, Weiss)

Is UBI data ‘big’?

Several observations per second

Continuously refreshing dataset

Large number of collected metrics

Relationships to other dynamic databases

Potential data sources

Little big data

Sample telematics data point:

Data Element	Data Value
Vehicle Identification Number	1234567890ABCDEFG
Date	11/01/2015
Time (UTC)	02:06:34
Global Positioning System Latitude	41.773312800000000000
Global Positioning System Longitude	-87.715253500000020000
Avg. Speed (MPH, since prior obs.)	21.30
Accelerometer Axis X Readings (g-force)	{ 0.00, 0.11, 0.21, 0.11, 0.05 }
Accelerometer Axis Y Readings (g-force)	{ -0.21, -0.23, -0.26, -0.14, -0.01 }
Accelerometer Axis Z Readings (g-force)	{ -1.00, -1.00, -0.99, -0.98, -0.99 }
Odometer (miles)	18,246

Example is purely illustrative and **not** intended to suggest, for instance, that 5 Hz is optimal accelerometer sample rate.

Technology considerations

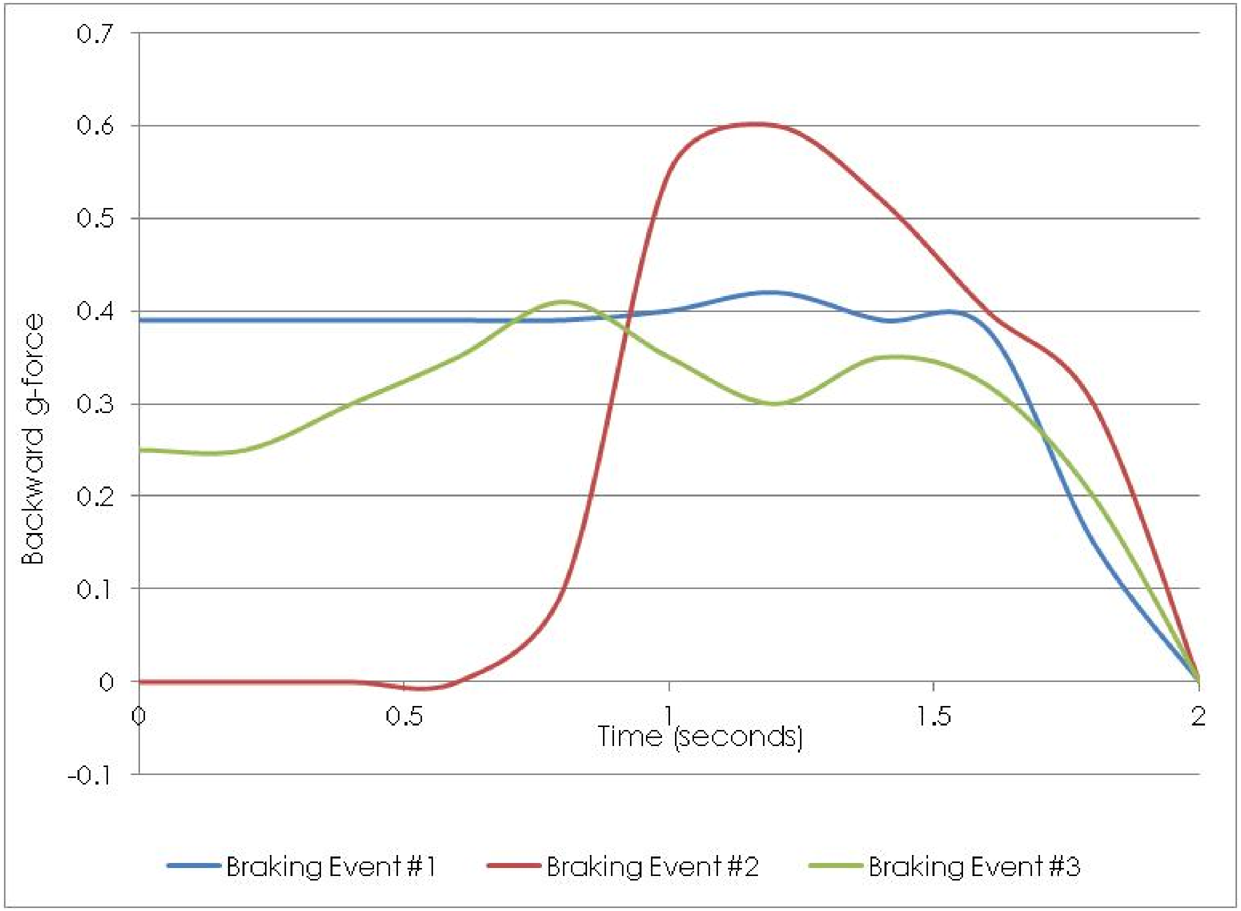
- Orientation – how do we determine what’s forward, backward, up and down?

- Calibration – how do we avoid hills, etc. registering as backward g-force?

- Device movements – how do we stop jostling the hardware from registering events?

- Event identification – how do we ensure different technologies record events in the same manner?

Give me a brake



If we define 'harsh braking' as 0.4 (backward) g-force exceedance, then these three braking events would be treated similarly

Modeling challenges

- Complexity – how do we transform 0s and 1s into something more predictive for insurance?
- Depth –100,000s of rows per risk … how do we compress with minimal loss of predictive power?
- Dimensionality – what do we do when the number of columns is in tens of thousands?
- Overlap – some classes are riskier … how do we avoid double-counting known effects?

Examples of context

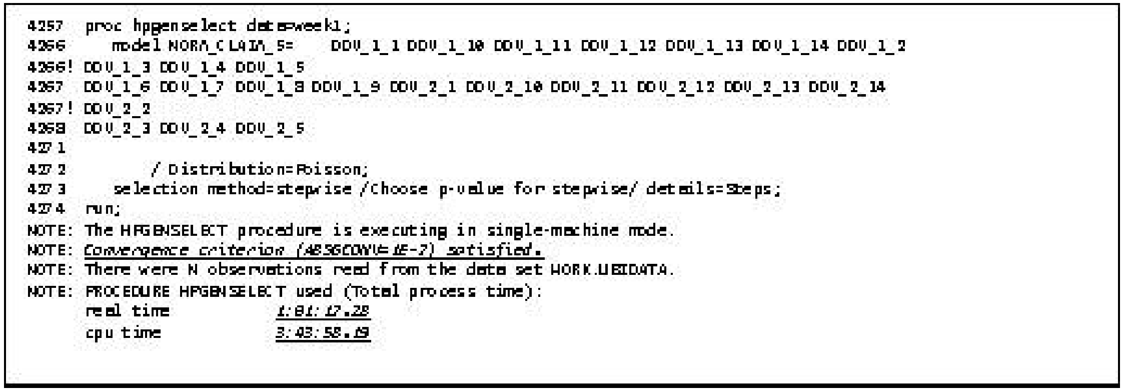
- Time of day (telematics data feed)
- Speed vehicle traveling (telematics data feed)
- Visibility and traction (weather database)
- Number of lanes (road atlas database)
- Speed vehicle *supposed* to be traveling (traffic database)

How to create ~10,000 variables in seconds

- As a first step, determine true vs. false
  - Exceedance of thresholds (…0.39, 0.4, 0.41…)
  - Presence of various condition sets e.g.
    - Morning rush hour?
    - Four lane road?
    - Visibility < 1 mile?
    - Traveling between 46-50 MPH?
- Sum exceedance counts and exposure over every possible condition set
- Aggregate exposure/counts to vehicle level
- Determine incidence rates of exceedance (per exposure unit) for each condition set

Variable selection approach

- Group variables thematically
- Software (HPGENSELECT, ‘step’)
- Staged stepwise → 535 candidates
- Final stepwise → 57 variables



```
4257  proc hpgenselect data=week1;
4266     model NORA_CLAIM_5=    DDV_1_1 DDV_1_10 DDV_1_11 DDV_1_12 DDV_1_13 DDV_1_14 DDV_1_2
4266! DDV_1_3 DDV_1_4 DDV_1_5
4267  DDV_1_6 DDV_1_7 DDV_1_8 DDV_1_9 DDV_2_1 DDV_2_10 DDV_2_11 DDV_2_12 DDV_2_13 DDV_2_14
4267! DDV_2_2
4268  DDV_2_3 DDV_2_4 DDV_2_5
4271
4272          / Distribution=Poisson;
4273     selection method=stepwise /Choose p-value for stepwise/ details=Steps;
4274  run;
NOTE: The HPGENSELECT procedure is executing in single-machine mode.
NOTE: Convergence criterion (ABSGCONV=1E-7) satisfied.
NOTE: There were N observations read from the data set WORK.UEIDATA.
NOTE: PROCEDURE HPGENSELECT used (Total process time):
      real time           1:01:17.28
      cpu time            3:43:58.19
```

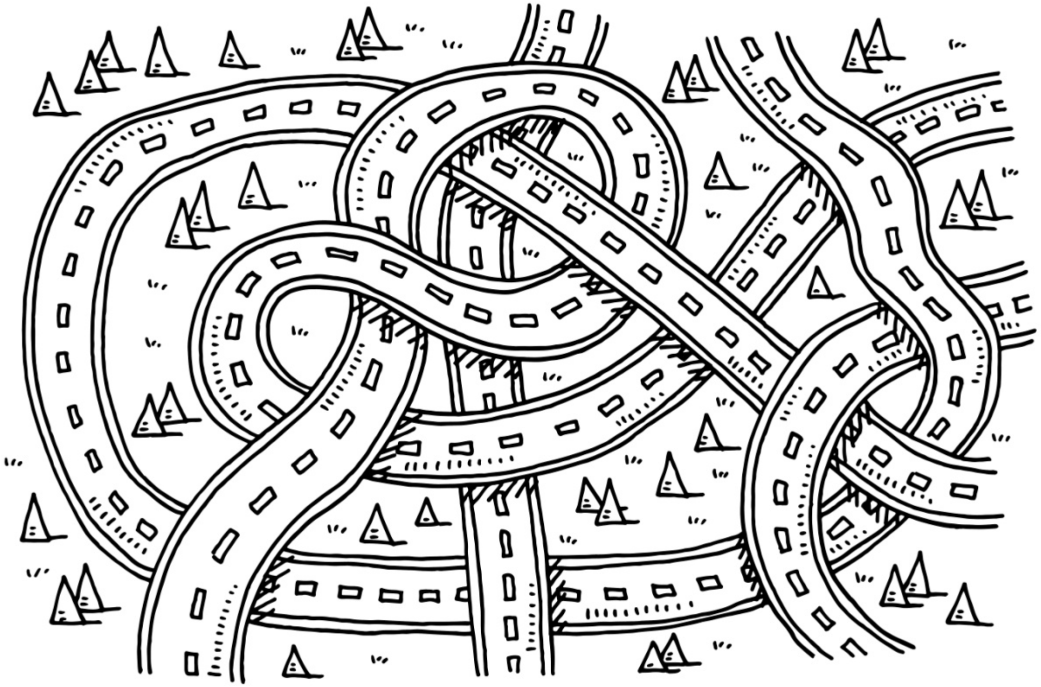
Poisson model

$$\text{E (claims)} = \exp \{ \theta + \sum_{j=1}^{J} \alpha\{j\}DDV1\{j\} + \sum_{k=1}^{K} \beta\{k\}DDV2\{k\} + \sum_{m=1}^{M} \vartheta\{m\}DDV3\{m\} + \sum_{n=1}^{N} \varphi\{n\}DDV4\{n\} \}$$

where …
θ: intercept term
DDVs1-4{j-n}:
 1-4 : thematic groupings i.e. braking, cornering, etc.
 j-n: individual DDVs i.e. incidence rates for condition sets
J,K,L,M: number of significant variables in family
DDV1-4{i}: normalized incidence rate of i^{th} variable in DDV family
ɑ(i), β(i), etc.: coefficient applicable to DDV1-4{i}

Overlap possibilities

- Driver classification
- Accidents and violations
- Credit history
- Territory
- Annual mileage

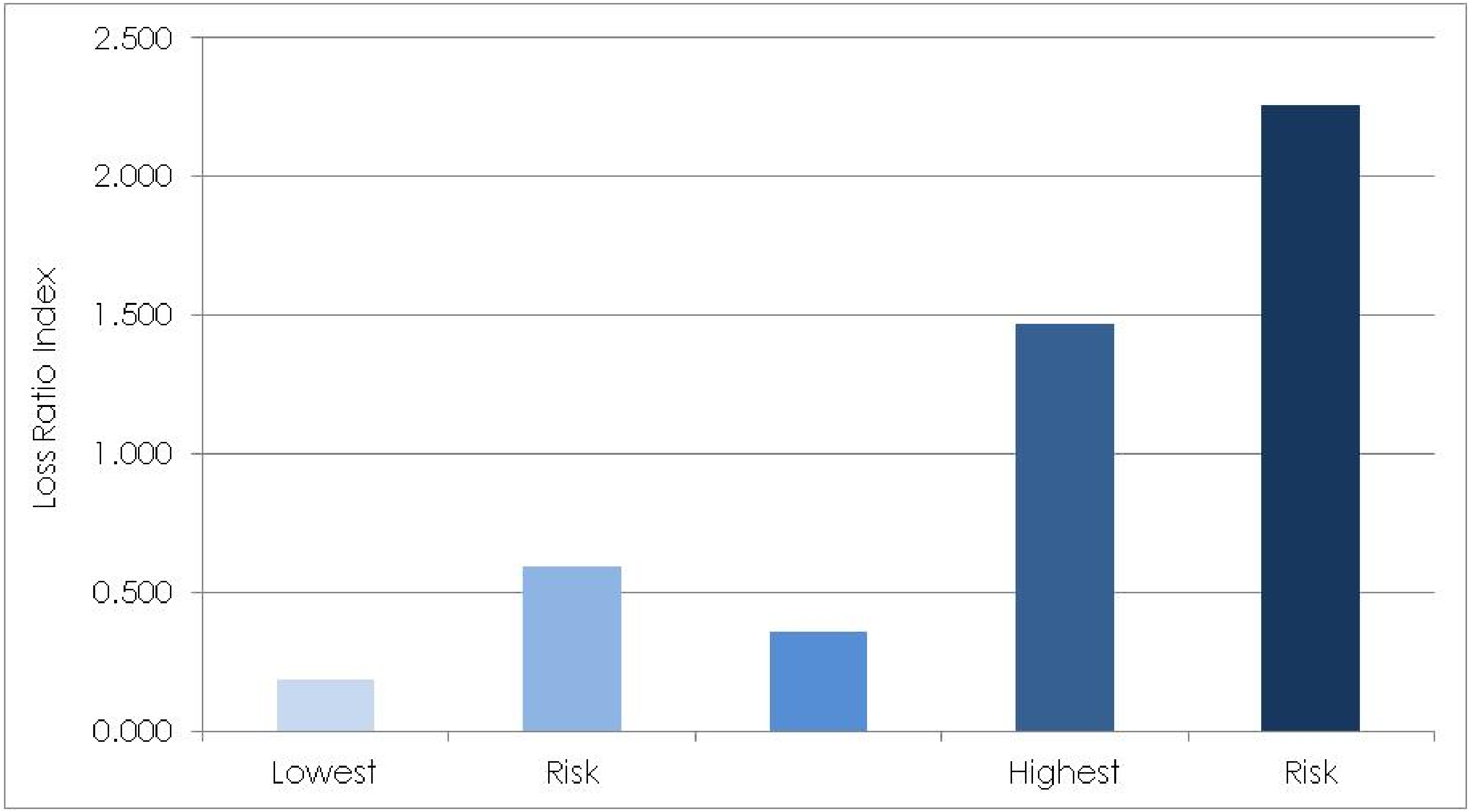


Approaches to overlap

Possibilities include:

- Assume independence of UBI v. trad'l
- Use existing variables as offsets/control
- Separate models by class
- Holistic model / machine learning
- UBI only approach

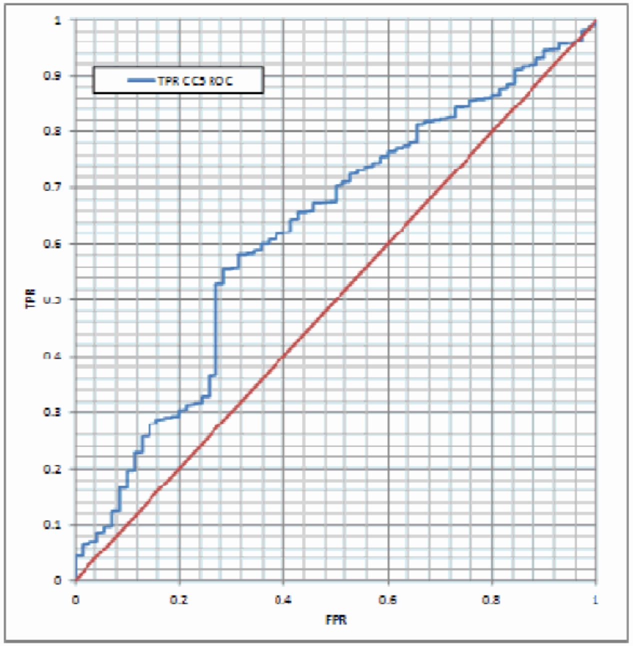
Loss ratio ‘lift chart’



Analysis performed using same vehicles used to train model, but separate period of 90 driving days to produce estimates. Chart suggests ‘All other things being equal,’ model identifies one in five that are >10x as risky.

ROC and 'area under curve'

- Thresholds for binary classification
- True vs. false positives and negatives
- FPR = FP ÷ (FP + TN)
- TPR = TP ÷ (TP + FN)
- AUC = 0.62



Graph and AUC relate to target variable of claims from holdout sample.

Alternative models

We tested whether the following alternative approaches to variable selection and modeling produced more parsimonious results

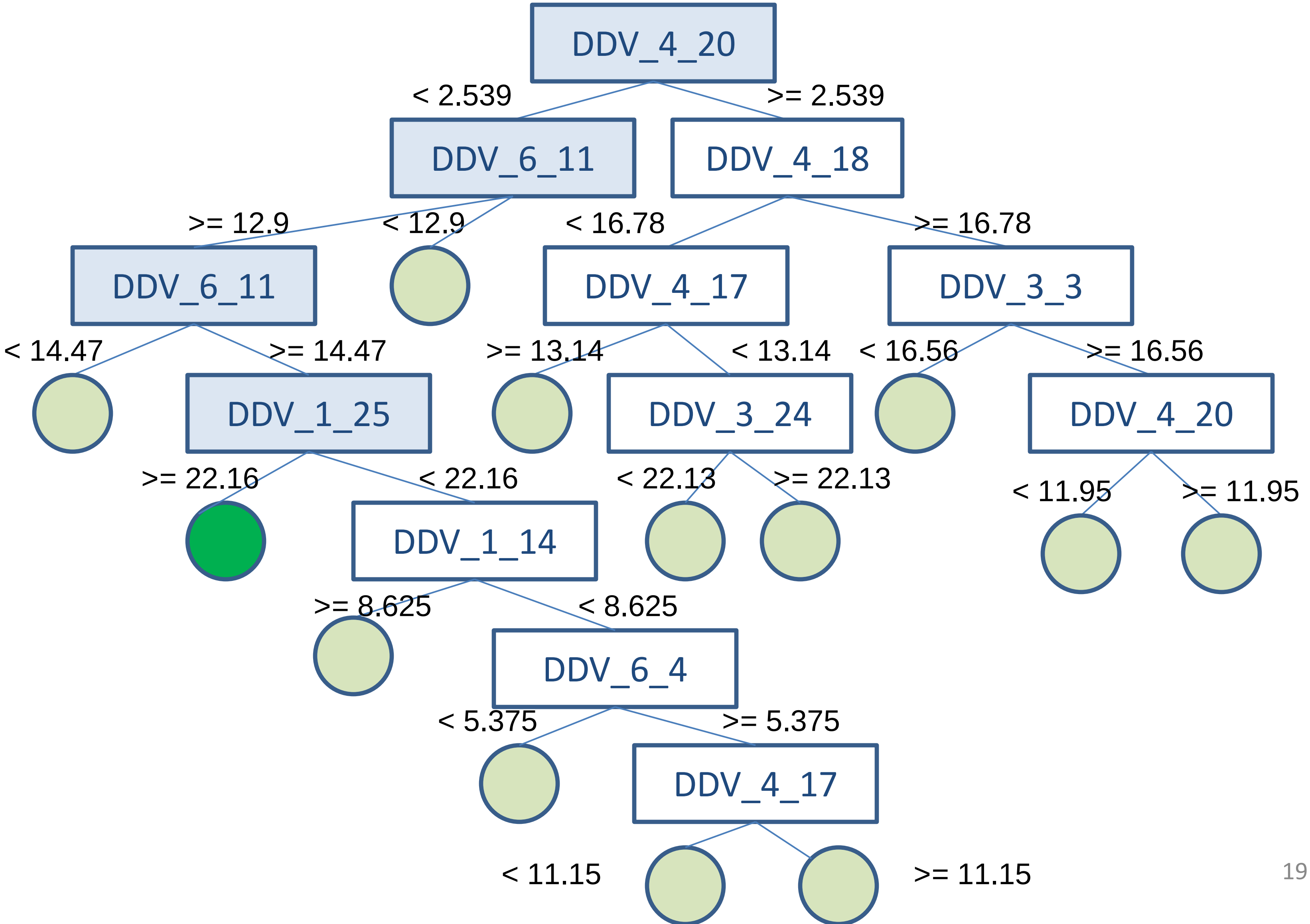
Variable Selection	**Model Form**
Stepwise*	Poisson Regression*
Tree	Tree
Tree	Poisson Regression
Tree then Stepwise	Poisson Regression
Stepwise, Tree, Stepwise	Tree

* - previously outlined approach

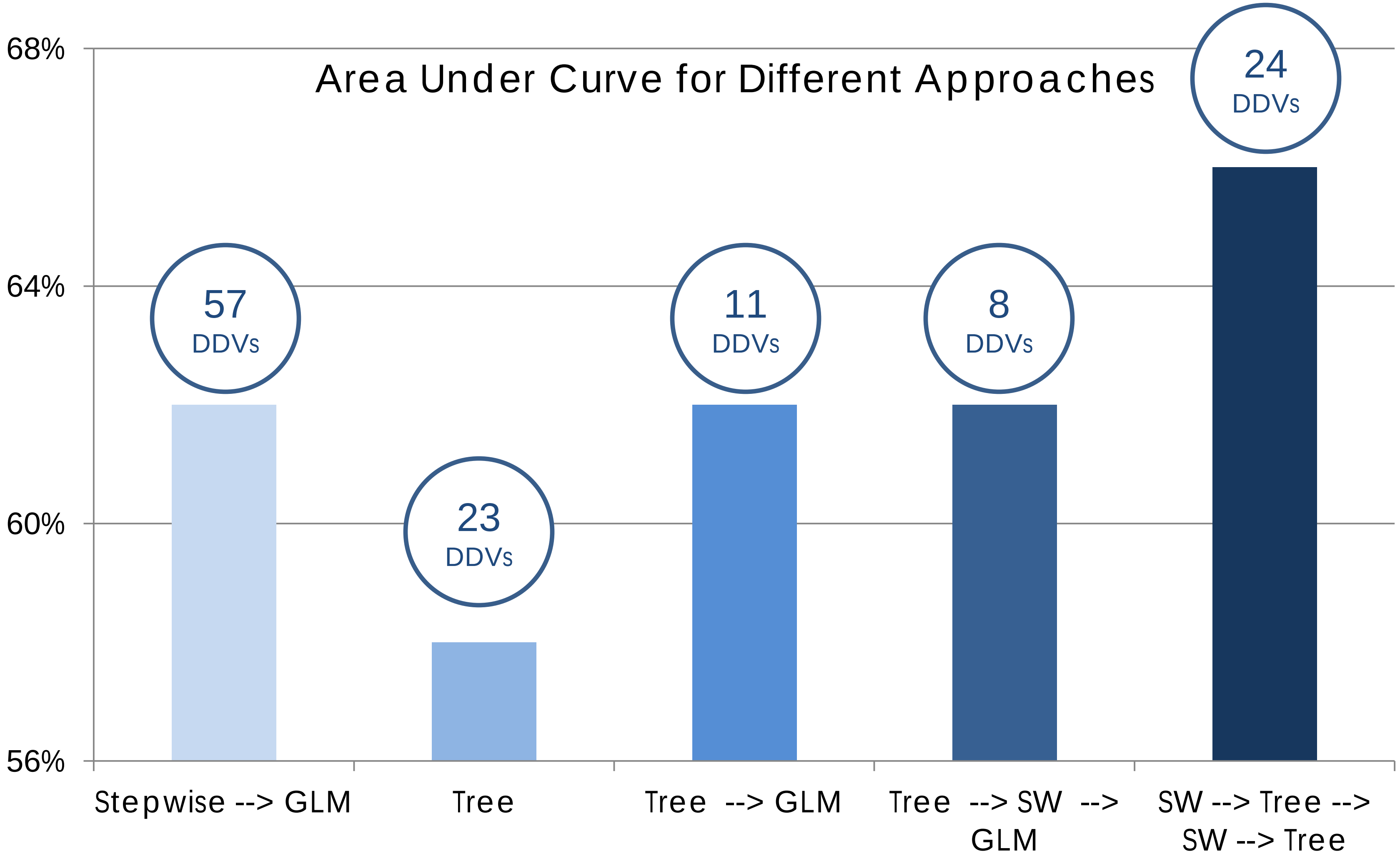
Data properties that may warrant using decision trees

- Dataset described by fixed set of attributes
- Target function has discrete set of values
- ‘Disjunctive descriptions’ potentially required
- Noisy training data (sparse or variant)

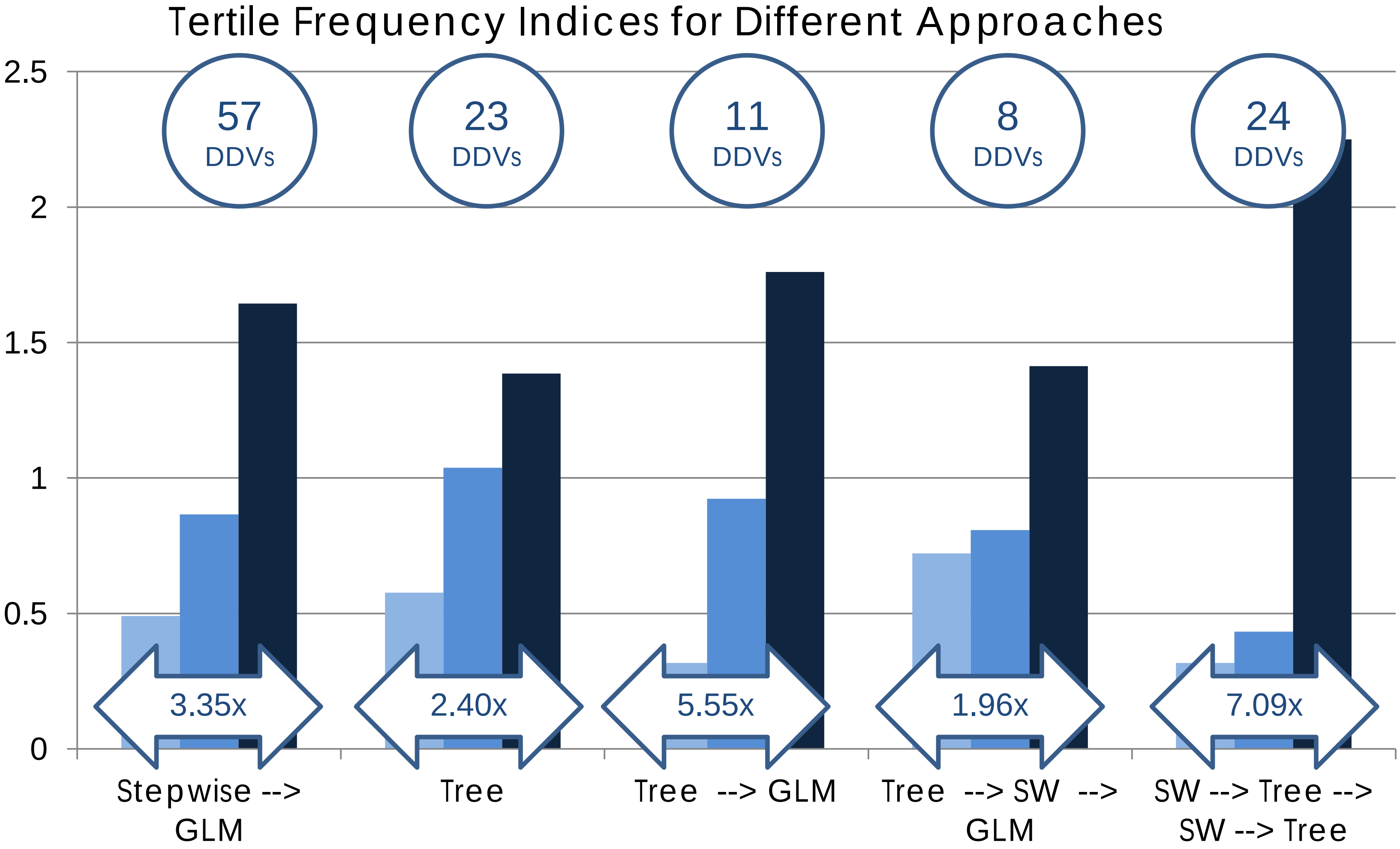
Classification trees for UBI



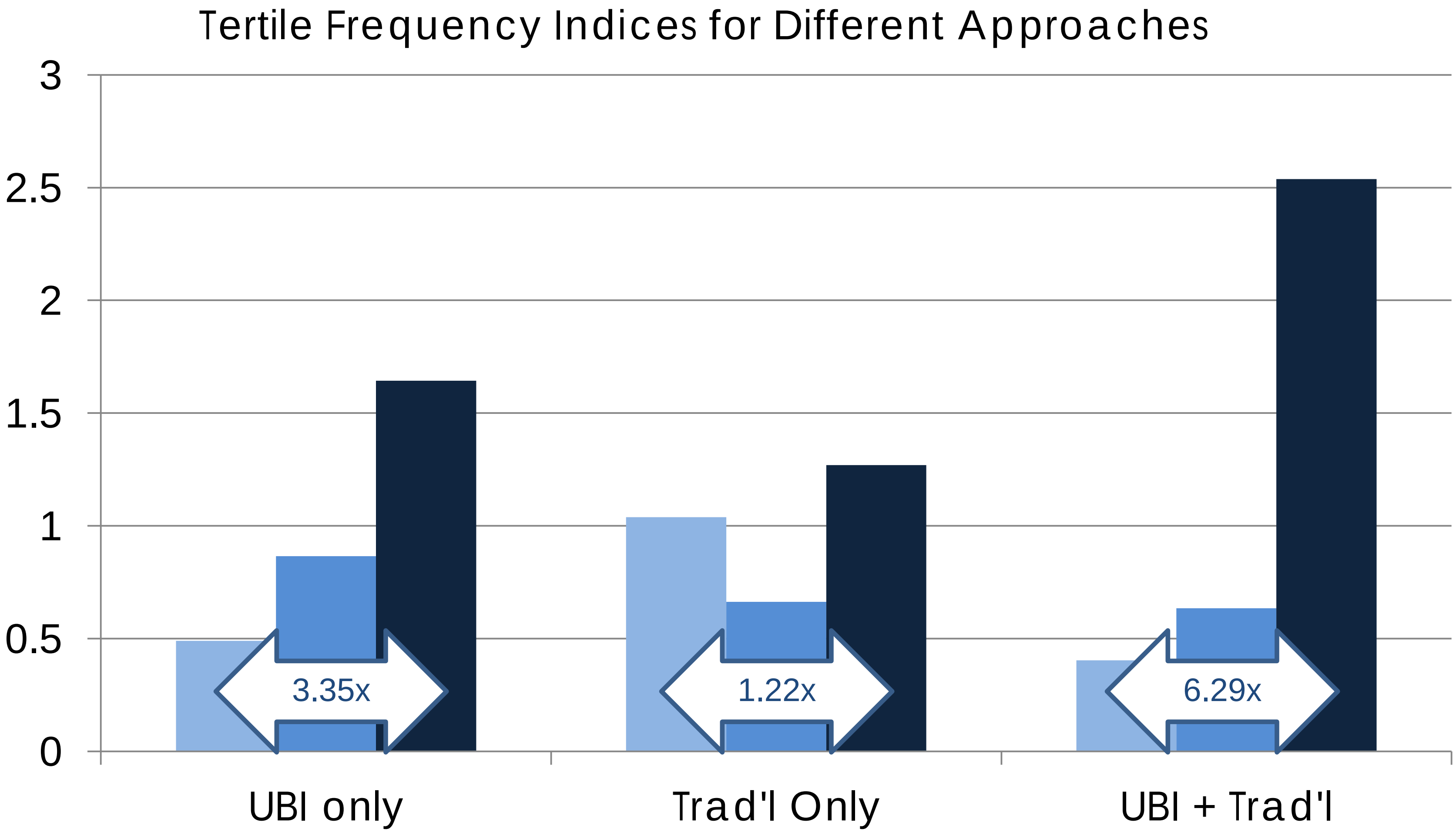
Revisiting variable selection



Combining trees, GLMs may yield strongest UBI results



Back with the old



Poor performance of traditional predictors may result in large part from relatively small data volumes.

Regulatory considerations

- Familiarity of approach
- Discounts vs. surcharges
- Confidentiality
- Observation period
- Support and policyholder challenges
- Privacy

Observation period

Considerations:

- Stability
- Predictive power
- Technology deployment
- Renewal management
- Behavioral modification

Areas for future exploration

- Pay per mile, trip, etc.
- Evolving data collection options
- Commercial lines / heavy trucks
- Changing ownership patterns
- Autonomous vehicles

Questions and remarks

Greg Hayward

greg.hayward.ajml@statefarm.com

Jim Weiss

jim.weiss@verisk.com