Data Sources

Rob Curry
Roosevelt Mosley

DATA SOURCES

Rob Curry, MAAA, FCAS

Roosevelt Mosley, MAAA, FCAS

Learning Objectives

At the end of this presentation, you will be able to understand:

- Data internal to insurance companies that is used in predictive modeling
- Insurance-related data external to insurance companies used in predictive modeling
- Non-insurance based data being used for insurance purposes
- What is involved in appending the data
- How the data is validated for accuracy
- How data is used in production

Agenda

- Data internal to insurance companies
- Insurance-related data external to insurance companies
- Non-insurance-based data being used for insurance purposes
- Appending the data
- Validating for accuracy
- How data is used in production

Internal Data – Personal Auto

- Age
- Gender
- Location
- Vehicle use
- Vehicle type
- Miles driven
- Prior claims (could be external source too)
- Other coverages/insurance purchased

Internal Data - Homeowners

- Construction
- Fire protection
- Pool, trampoline, etc (could be external)
- Home business
- Smoker
- Distance to water
- Additional building features (could be external)
  - # of rooms,
  - # of bathrooms
  - Age of home
  - Age of roof
  - Age of utilities

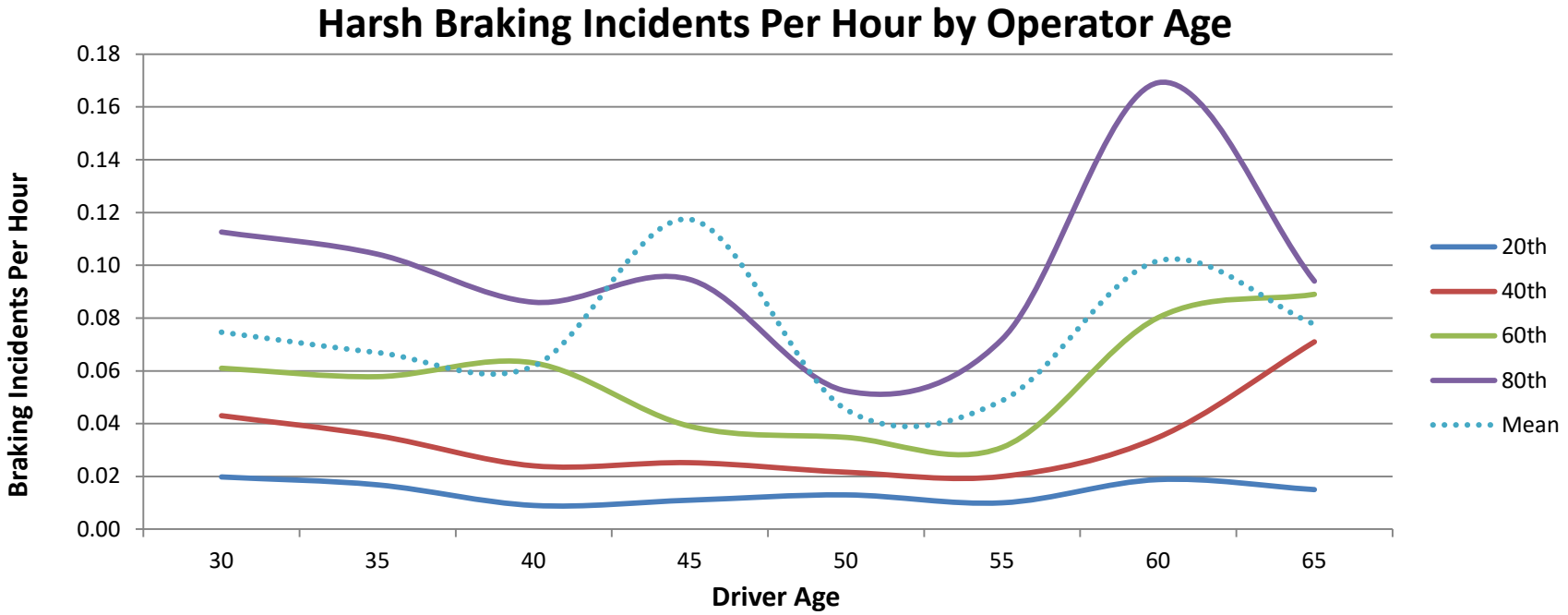
External Data – Insurance Related

- Loss history (internal too)
  - For other coverages
- MVR
- Building characteristics (some internal)
- Prior coverage
- Replacement cost estimates
- Vehicle symbols

Non-Insurance Data

- In some cases a more direct measure of risk than internal or external insurance data
  - replacement or incremental over existing variable
- Credit history
- Weather
- Census
- Business locations
- Crime
- Telematics data
- Traffic density
- Education
- Smart home data
- Social media
- Vehicle history
- Property permit data

Telematic Info

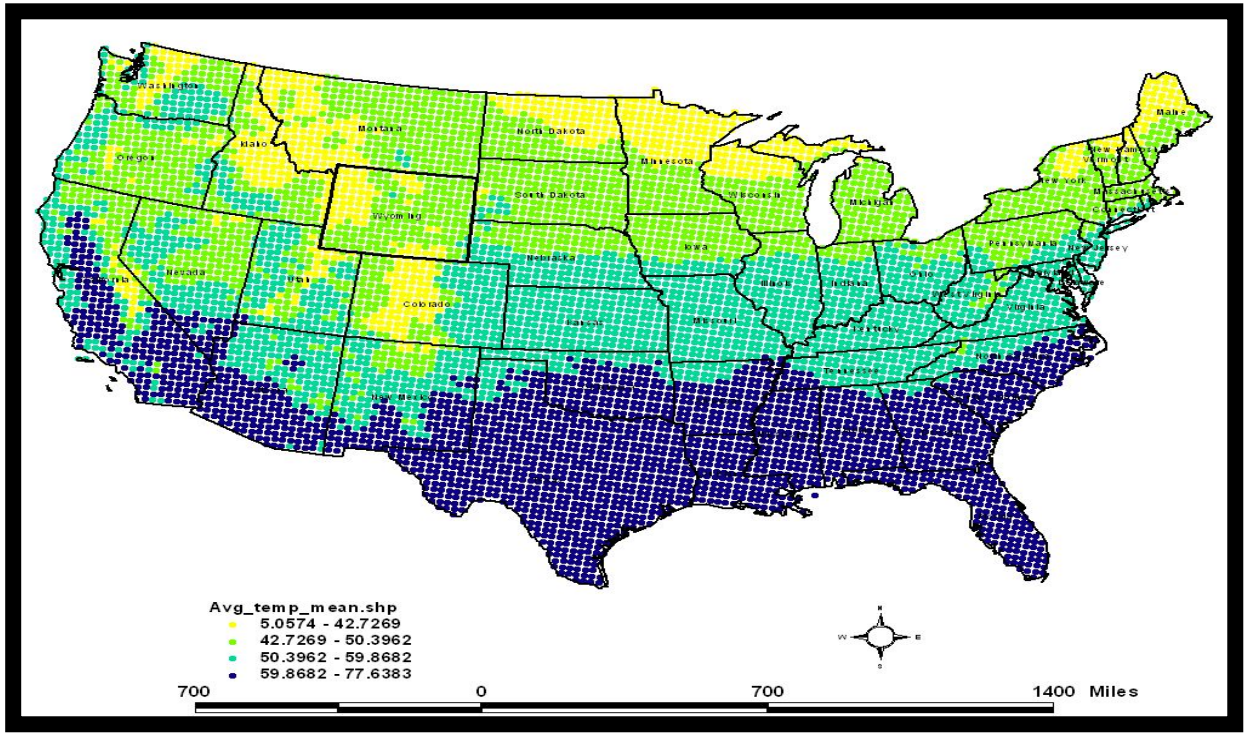


External Data Weather

(Indicators, daily, consecutive days, number of days)

- **Temperature**
  - Below freezing / high temperatures
  - Variations / average / min / max / deviation
- **Precipitation, wind and snow**
  - With / without
  - Average / min / max / deviation
- **Interactions**
  - Weight of snow (snow + temp)
  - Ice (rain + temp)
  - Fire (no rain, high temp + high wind)
  - Blizzards (snow + wind)

External Data Weather



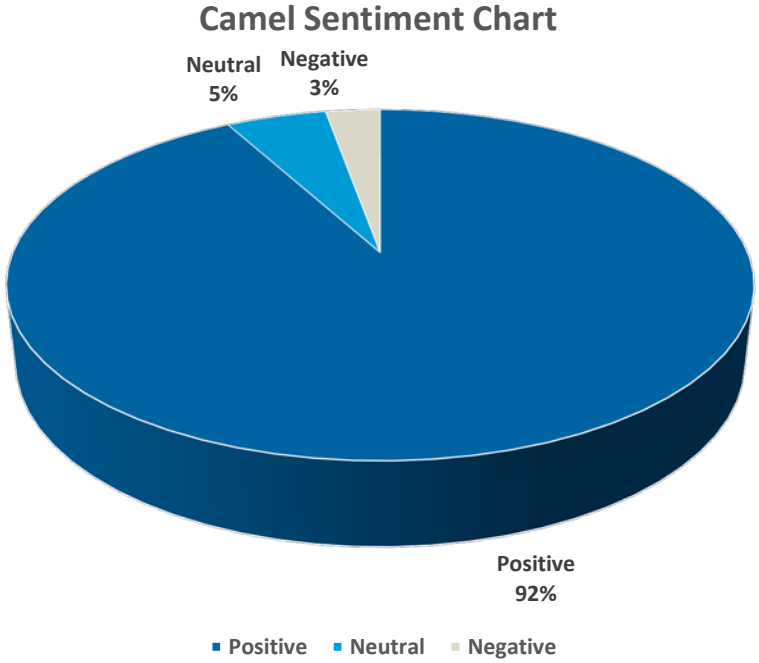
Mean of daily average temperature

Homeowners Telematics: Data Possibilities

 Usage	• Patterns of energy and water consumption • Water running when no occupants home • Which rooms are used, when, and for how long?
 Occupants	• Occupants: number, frequency of access • Number of smokers; frequency and time of day of smoke • Number of connected devices
 Contents	• Movement of contents in and out of house • Major appliance location • Sprinkler system detection
 Residence	• Roof age and condition; material; weight load • Wind speed and barometric pressure • Gas leak detection

Social Media Examples

Camel – Sentiment Chart



Census Data

- Levels
  - State
  - ZIP
  - Census

- Housing density
- Vehicle density
- Number of household members
- Gender
- Home occupancy
- Commuting
- Employment type
- Etc.

Data – General Issues

- There are some variables that insurers avoid like race and income
- Be aware of the potential for variable to be a proxy for or correlate with variables that you are trying to avoid
- Obtaining of raw data vs. scored data – reliance on other's work

Appending Data for Analysis

- What is the level of resolution for the data source?
  - Zip code
  - Census block, census block group, census track
  - Address
  - Name
- Have a plan for dealing with no-hits, because they will happen
  - An issue for analysis and production
  - Should no-hit be a valid value to model?

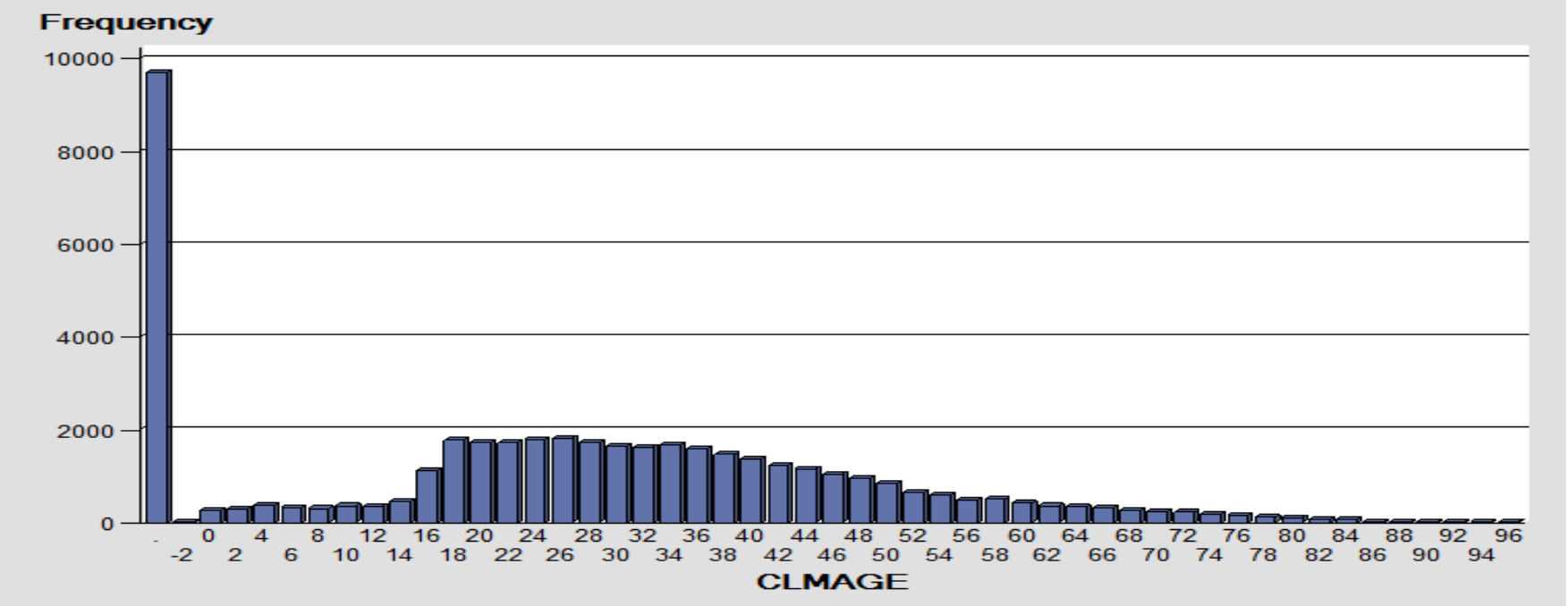
Data Source Validation

- ASOP No. 23 – Data Quality
- Edits on individual values
  - What are expected range of values
- Relationship edits
  - E.g., age of home = 5, age of utilities = 25
- Univariate review
  - Is the relationship intuitive?
  - Final result could be different when combined with other variables in a model
- Histogram of values
- Correlation matrix
  - Try to avoid using variables that are too correlated

Univariate example

Annual Mileage	Exposure	Claim Count	Loss Amount	Claim Severity	Pure Premium
Missing	40,239	1,779	8,649,340	4,861.91	214.95
Less Than 3,000	1,476	47	176,155	3,747.98	119.33
3,000 - 4,000	933	42	108,133	2,574.60	115.91
4,000 - 5,000	1,169	53	178,313	3,364.39	152.55
5,000 - 6,000	1,186	42	168,878	4,020.90	142.45
6,000 - 7,000	1,759	69	244,180	3,538.84	138.83
7,000 - 8,000	1,747	86	512,799	5,962.78	293.54
8,000 - 9,000	1,720	61	210,646	3,453.21	122.45
9,000 - 10,000	1,779	76	274,909	3,617.23	154.53
10,000 - 11,000	1,782	71	331,727	4,672.21	186.13
11,000 - 12,000	1,697	67	246,659	3,681.47	145.32
12,000 - 13,000	1,431	64	259,428	4,053.57	181.32
13,000 - 14,000	1,400	64	256,718	4,011.22	183.41
14,000 - 15,000	1,243	74	277,828	3,754.43	223.54
15,000 - 16,000	1,012	48	233,904	4,873.01	231.11
16,000 - 17,000	976	49	227,414	4,641.11	232.97
17,000 - 18,000	792	45	159,007	3,533.49	200.75
18,000 - 19,000	677	44	144,803	3,290.98	214.02
19,000 - 20,000	575	28	105,747	3,776.66	183.82
Over 20,000	3,065	174	788,571	4,532.02	257.24

Histogram Example



Correlation Matrix Example

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
V1										
V2	19%									
V3	16%	9%								
V4	20%	18%	17%							
V5	20%	16%	15%	29%						
V6	28%	19%	16%	73%	50%					
V7	17%	100%	70%	30%	14%	31%				
V8	16%	68%	48%	28%	13%	30%	44%			
V9	12%	32%	43%	10%	12%	19%	18%	20%		
V10	18%	38%	56%	23%	19%	25%	54%	46%	35%	

How Data is Used in Production

- Is data internally hosted?
- If need to get externally, what is response time?
- What is desired speed of underwriting decision?
  - Immediate in 4 steps on an app?
- How often is data updated?

Questions

?

For More Information

For more information, contact

Marc Rosenberg, senior casualty policy analyst,

At rosenberg@actuary.org or (202) 785-7865