Signal in Noise

Scott Zrebiec, Ph.D.
LexisNexis
Insurance Data Services, Modeling & Analytics
April, 2014

Outline

Smoothing methods allow the creation of extremely predictive data out of signal that would otherwise be hidden in the noise.

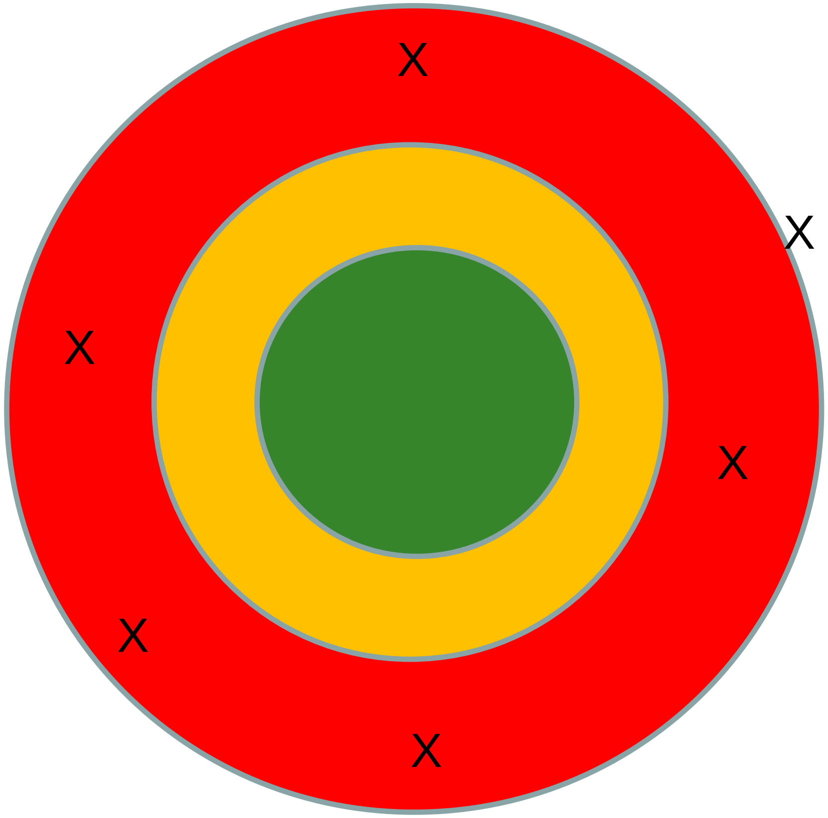
1. Hierarchical Credibility
2. Mathematical Approaches
3. Spatial smoothing approaches

Methods:

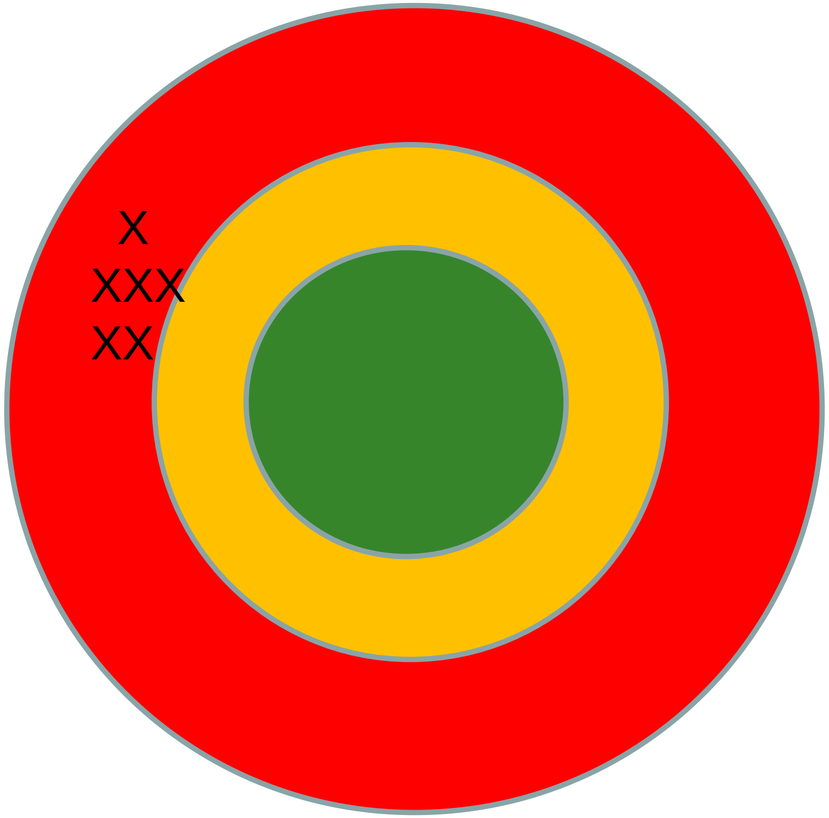
Noisy & Accurate → Accurate & Precise

Accuracy vs Precision

Perfect Accuracy



Biased but Precise



Goal: Accurate and Precise!

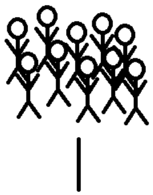
Hierarchical Credibility Theory

-Practical way to improve data
-Works with any hierarchy
-Great performance

1. Credibility

Simplistic view of Credibility:

- Employs some independence assumption
- Uses a simple hierarchy:



Large “Credible” sample

Similar “non-Credible” sample

The strength of credibility is in its practicality: reducing variance of estimates.

1. Miscellaneous Rant

Theorem: ("Central Lie of Mathematics").

If $\{X_j\}$ is a sequence of i.i.d. random variables,

and if $E[X_j]$ = mu $< \infty$,

and if $0 < Var[X_j] = \sigma$^2 $< \infty$,

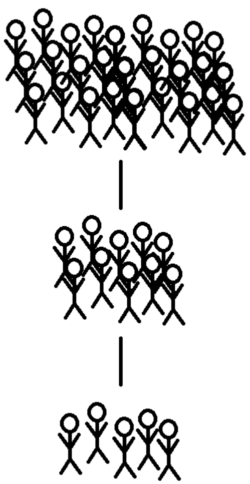
Then $\lim_{n\to\infty} \sqrt{n} \sum_{j=1}^{n} \frac{(X_j - \mu)}{n} \rightarrow N(0, \sigma)$

Observation: Independence is not in general true.

1. Hierachical (Credibility) Smoothing

Question why stop at 2 levels:

Multi level
Hierarchy



Census:

State
|
County
|
Tract
|
Blockgroup

Postal:

State
|
County
|
County & Zip3
|
County & Zip 5

Smoothed data is precision and accurate

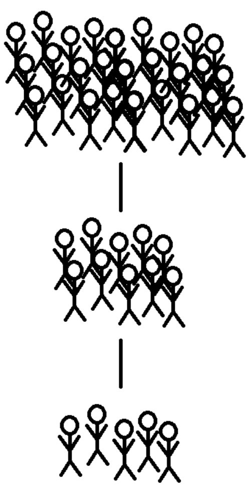
Alternate structures: adjacency, similarity, clustering

1. General, n-level hierarchy

Theorem: (Bühlman, Gisler)

Hypotheses (short version): i.i.d. at second highest level, conditionally independent given same leaf.

Multi level Hierarchy



Cond. Variance

τ_3^2
|
τ_2^2
|
τ_1^2

Mean

μ
|
μ_h
|
μ_i

Credibility Estimate:
(see text for details)

n/a
|

$$\mu_h'' = \alpha_h^{(2)} B_h^{(2)} + (1 - \alpha_h^{(2)}) * \mu,$$

$$B_h^{(2)} = \sum \frac{\alpha_i}{w_h} * \mu_i^{(1)}$$

$$\alpha_i = \frac{n}{n + \frac{\tau_1^2}{\tau_2^2}}$$

|

$$\mu_i'' = \alpha_i^{(2)} \mu_i + (1 - \alpha_i^{(2)}) * \mu_h''$$

1. A Noisy Accurate Data Element

Consider by peril, regional loss statistics

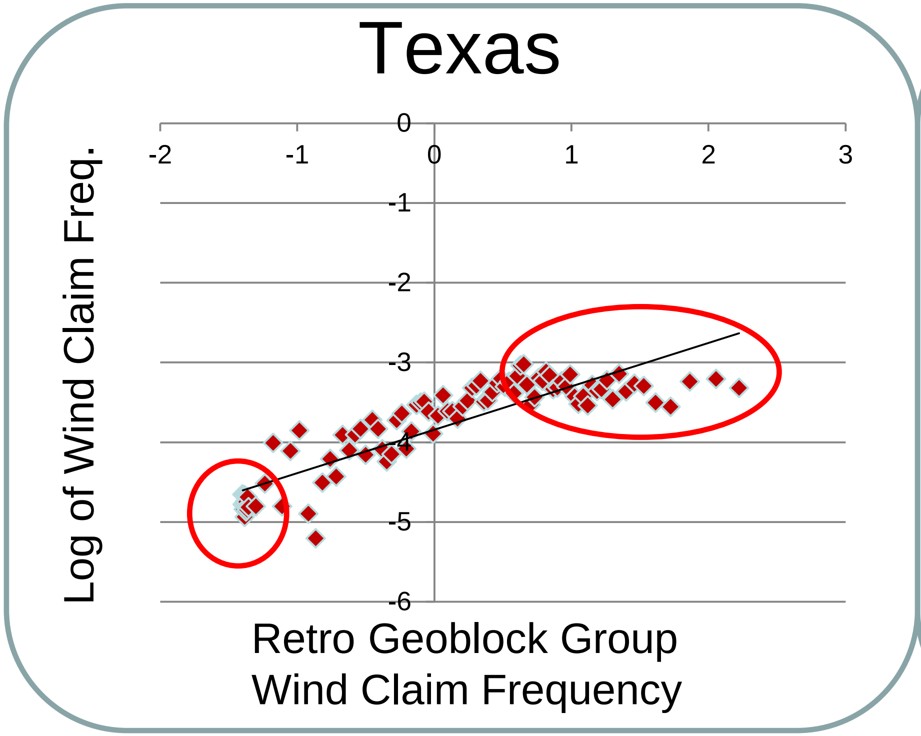
- $Frequency = \frac{Claim\ Counts}{Earned\ Exposure}$
- $Severity = \frac{Loss\ Amounts}{Claim\ Counts}$
- $Loss\ Cost = \frac{Loss\ Paid}{Claim\ Counts}$

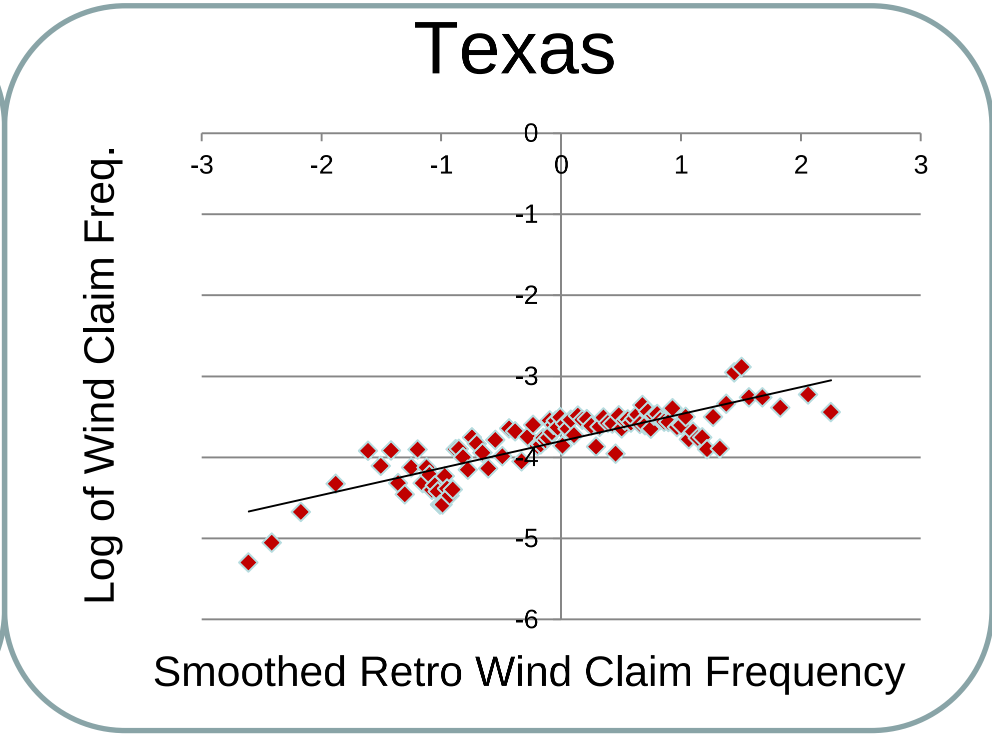
Statistics are easy to compute, and accurate.

At the finer levels they are too noisy to be useful.

1. Credibility Smoothing Results

Weighted estimates are stable and accurate





Precision gained by weighting with similar data.

Mathematical Smoothing Techniques

-Identify similarity
-Smooth IDW Average
-Creates new data

2. Metrics Identify Where to Weight

Metrics quantify similarity/distance between objects.

Lots of types of metrics:

- "Euclidean" Distance
- Distance between houses using characteristics
- Distance between areas using statistics

2. How to Creating Metrics

Creation of a metric/component metric

- Transform to segment
  - e.g. Year built is great at segmenting post 1960
  - $Distance\ YB\ between\ 2\ prop. = |\ \Delta\ Rescaled\ Year\ Built\ |$
- Rescale/ data to be comparable

Combine component metrics using L^p metrics

- $H.Distance = \sqrt{\sum c_j * Distance\ for\ Characteristic\ j^2}$

Optimize c_j and transformation based on needs.

2. IDW averages

IDW averaging smooths data by putting the most weight on the most similar data

- $IDW\ Avg\ of\ X\ for\ Obs\ j = \frac{\sum w_i * X_i}{\sum w_i}$
- $w_{i,j} = \frac{1}{Distance\ from\ obs.j\ to\ obs\ i}$

Uses: Weather data, Property Characteristics, high dimensional metric space.

2. Example-Identifying Comps

Goal: provide a default value for missing data

Adaptive Distance: Measures similarity of two properties using:

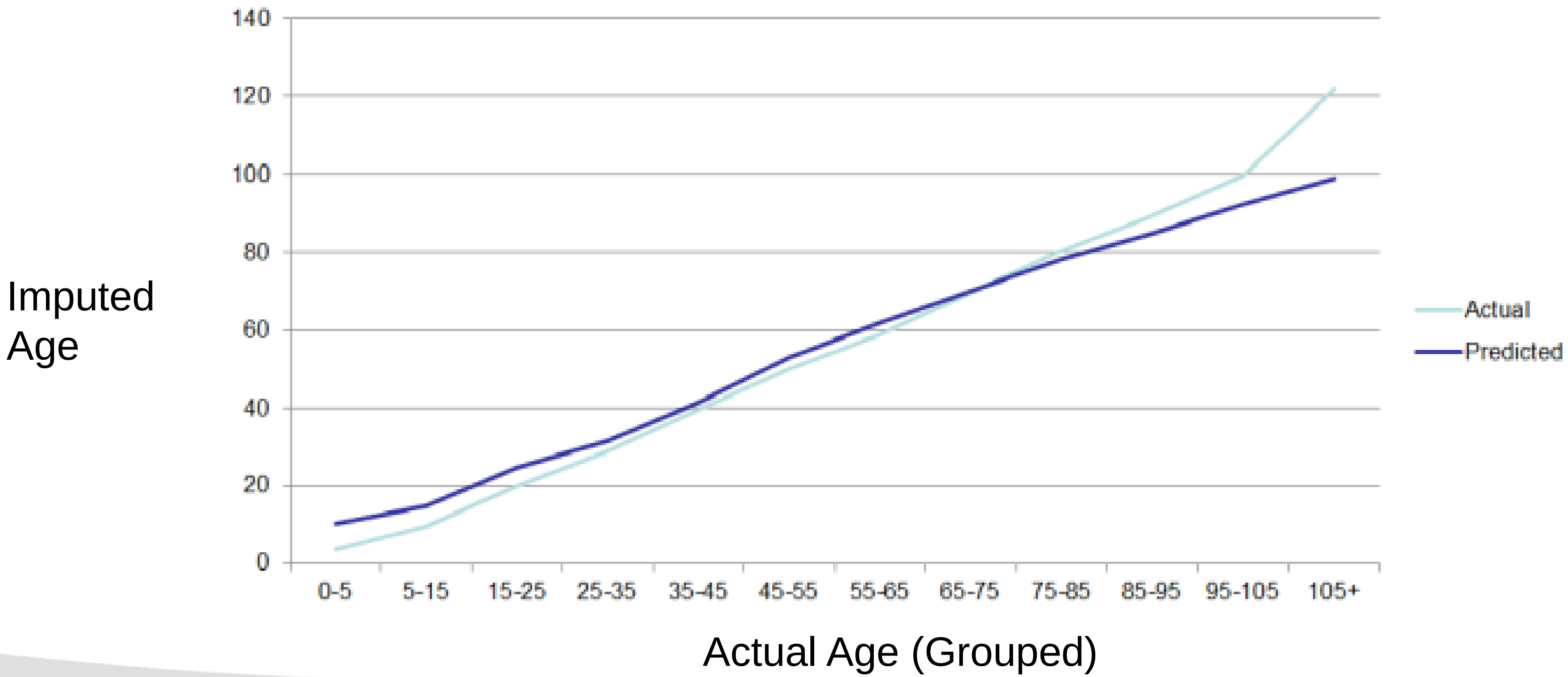
- "Distance" between two properties based on 10 characteristics
- Uses the data that is present

Property	Base	Best Match	Next Best Match	Worst Match
Value	65,900	65,800	NA	350,000
Baths	1	1	1	3
Area	NA	NA	1124	NA
Story	1	1	1	2
Garage	Carport	Carport	Carport	Attached
A.D.	**0.0**	**0.6**	**0.6**	**16**

2. IDW Averaging Results

Imputation: Accurate Default Values

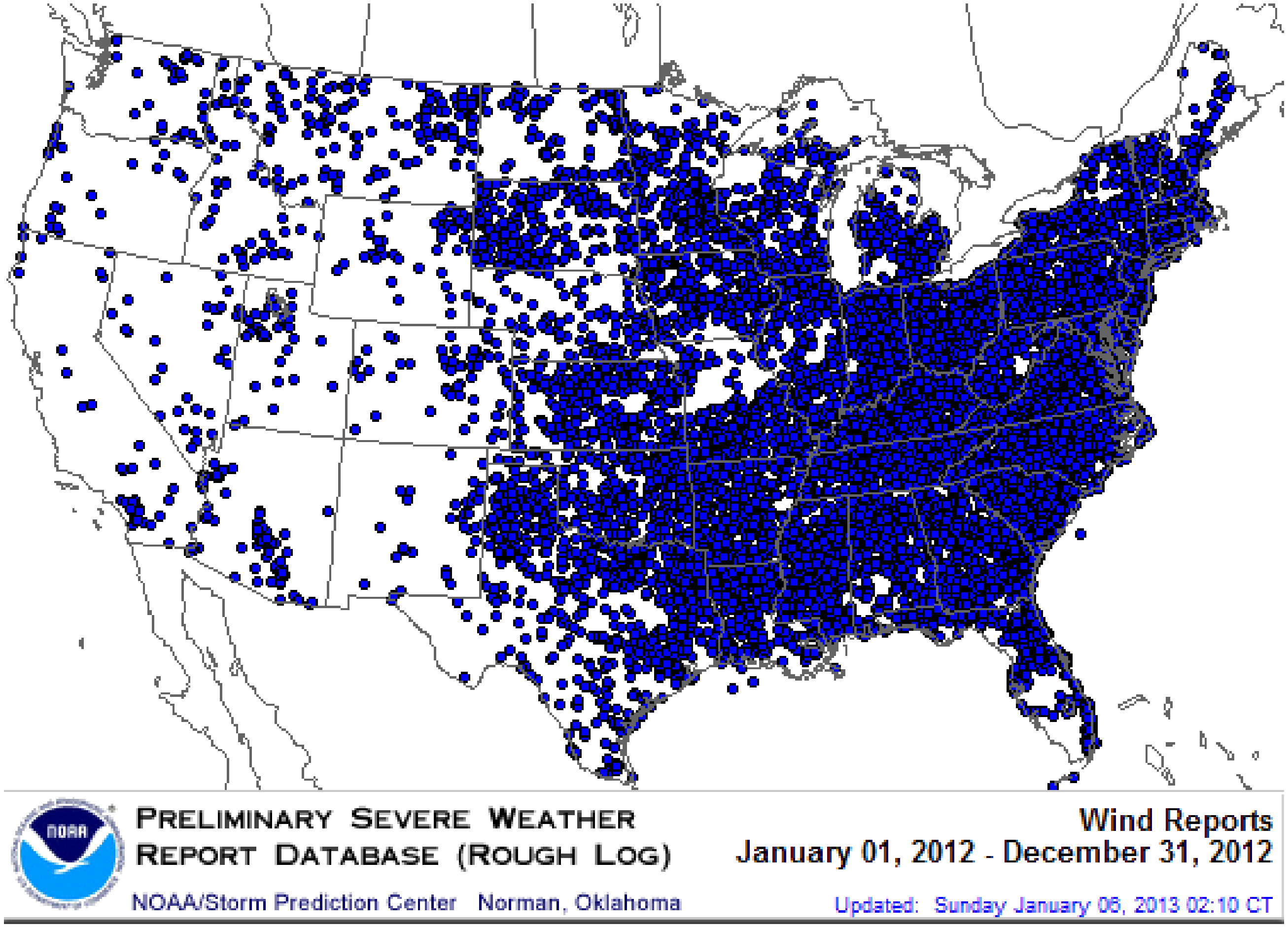
- Results are accurate and precise
- Outliers are slightly biased towards the mean



Spatial Smoothing Approaches

-Point → Region → Observations
-Kernel and Kriging Methods
-Results

3. Point data



Source: NOAA Storm Prediction Center; http://www.spc.noaa.gov/climo/online/monthly/2012_annual_summary.html#

3. Kernel Smoothing

Point data is assigned to regions using Kernel smoothing

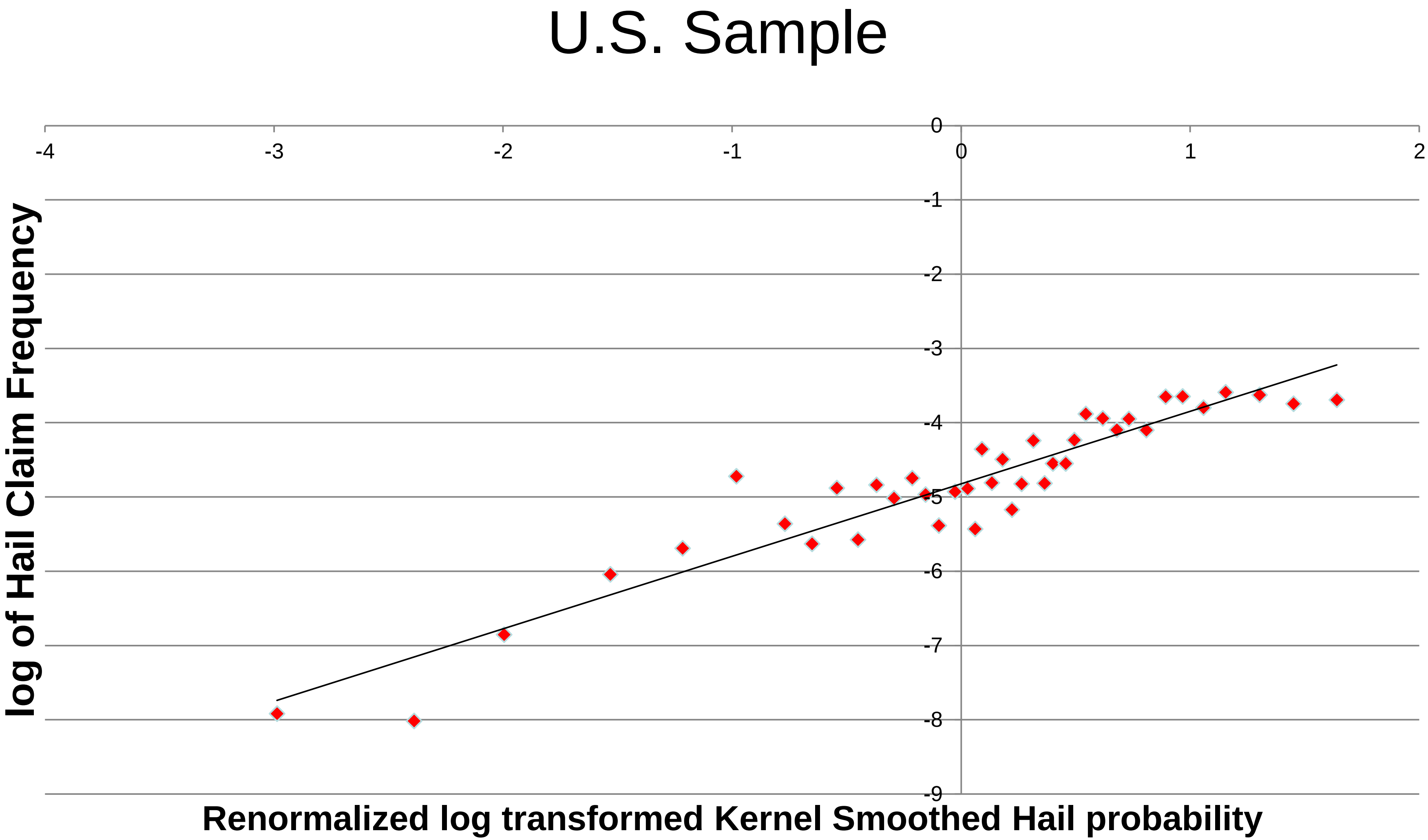
$$Hail\ Risk\ at\ x = \sum_{\{y\}} K_{\lambda}(x, y)$$

Where $f(x) = K_{\lambda}(x, y)$ is the pdf at x for a Random variable, e.g. Uniform, with μ=y and $\sigma = \lambda$.

Even simpler interpretation: Number of Storm events in X –miles in the past Y years

Issues: observational bias, boundary effect, choice of λ

3. Kernel Smoothing Results



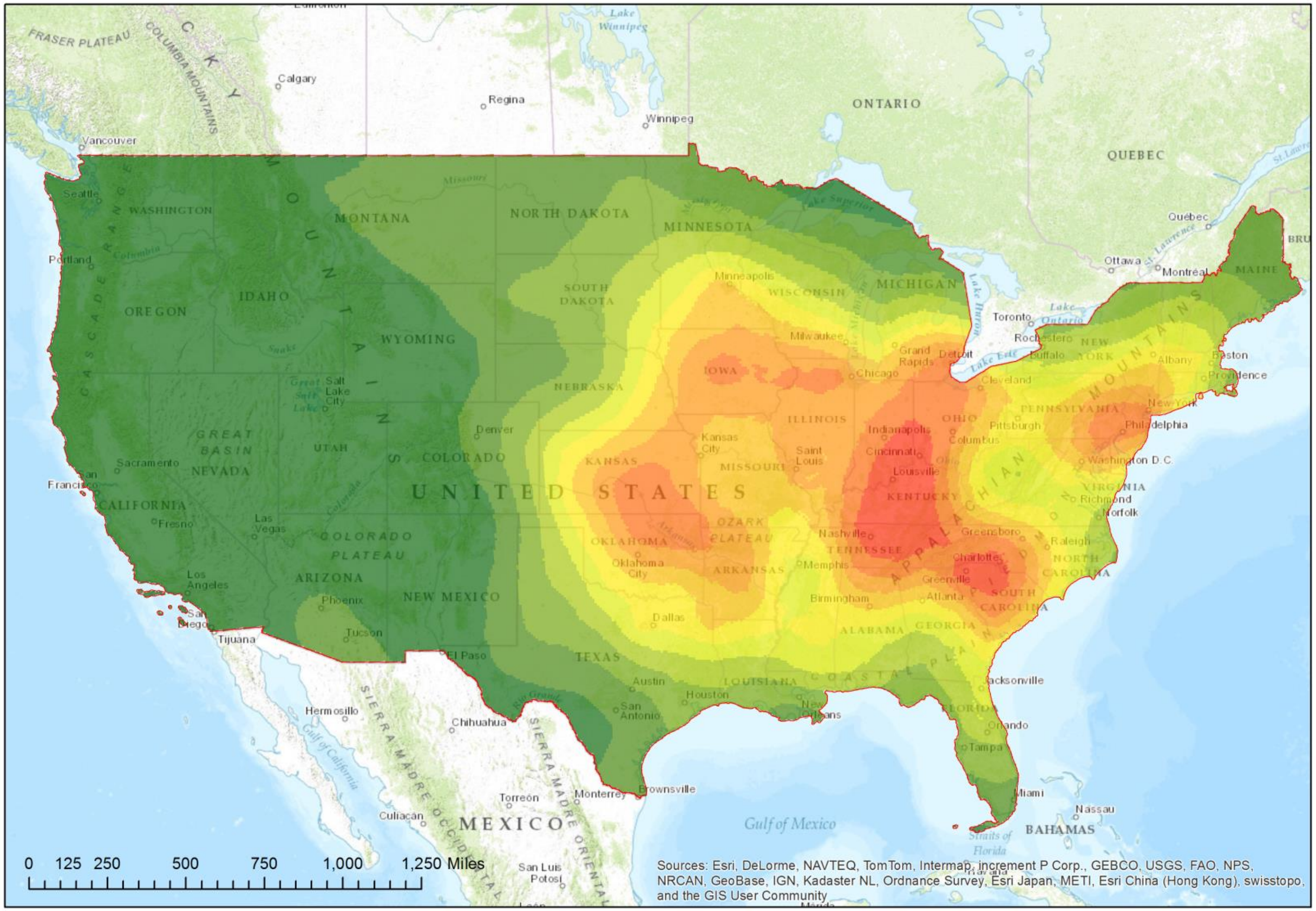
3. Kriging

Observation: Adjacent points have correlated geographic data.

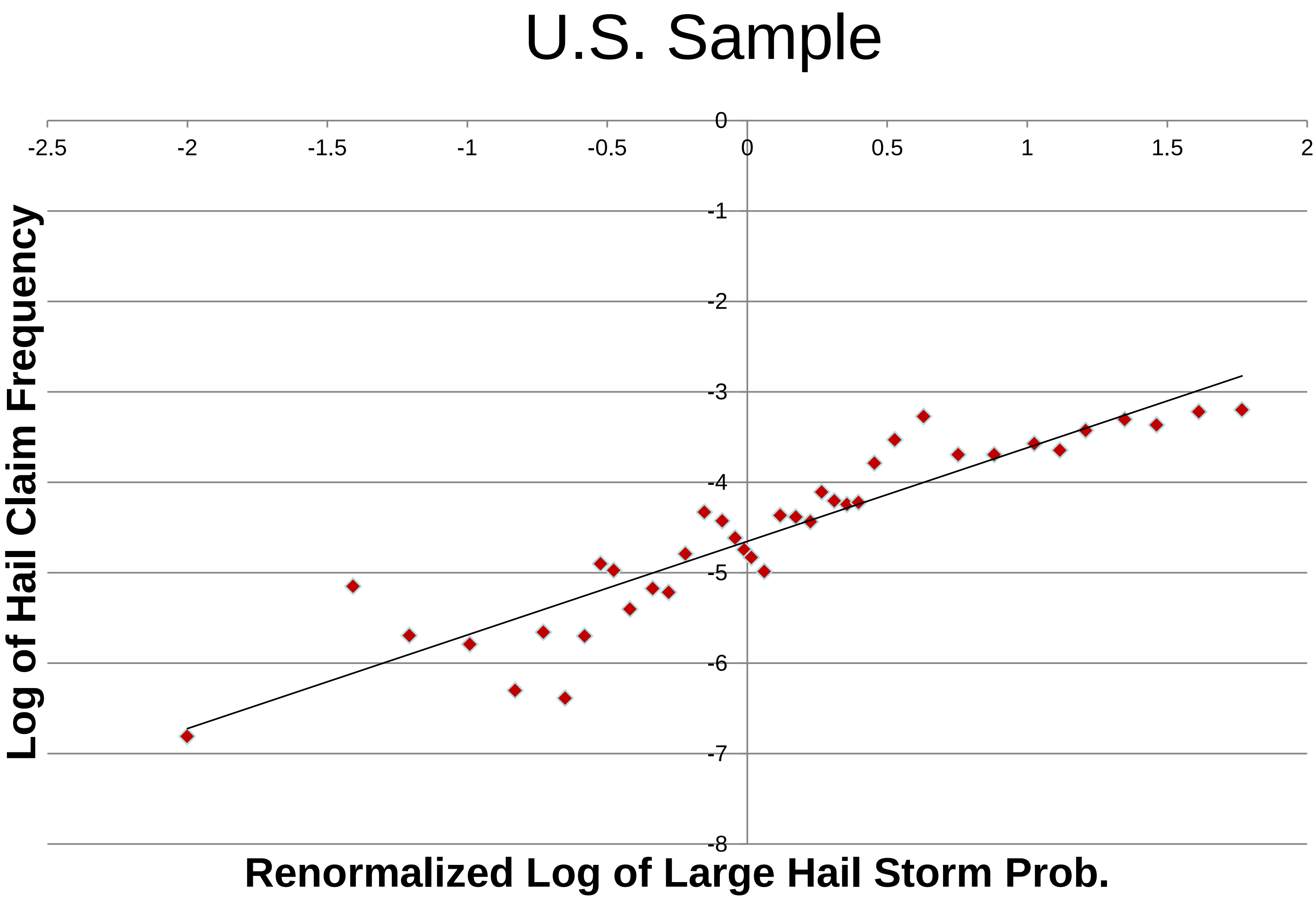
Kriging:

- Assumes a Gaussian field:
  - Each position associated with random variable
  - Spatial correlation
  - Either interpolation or statistical fit
- Smoothed average of nearby points.
- Produces "similar" results to kernel approaches

3. Map-Wind Storm Probability:



3. Kriging Results

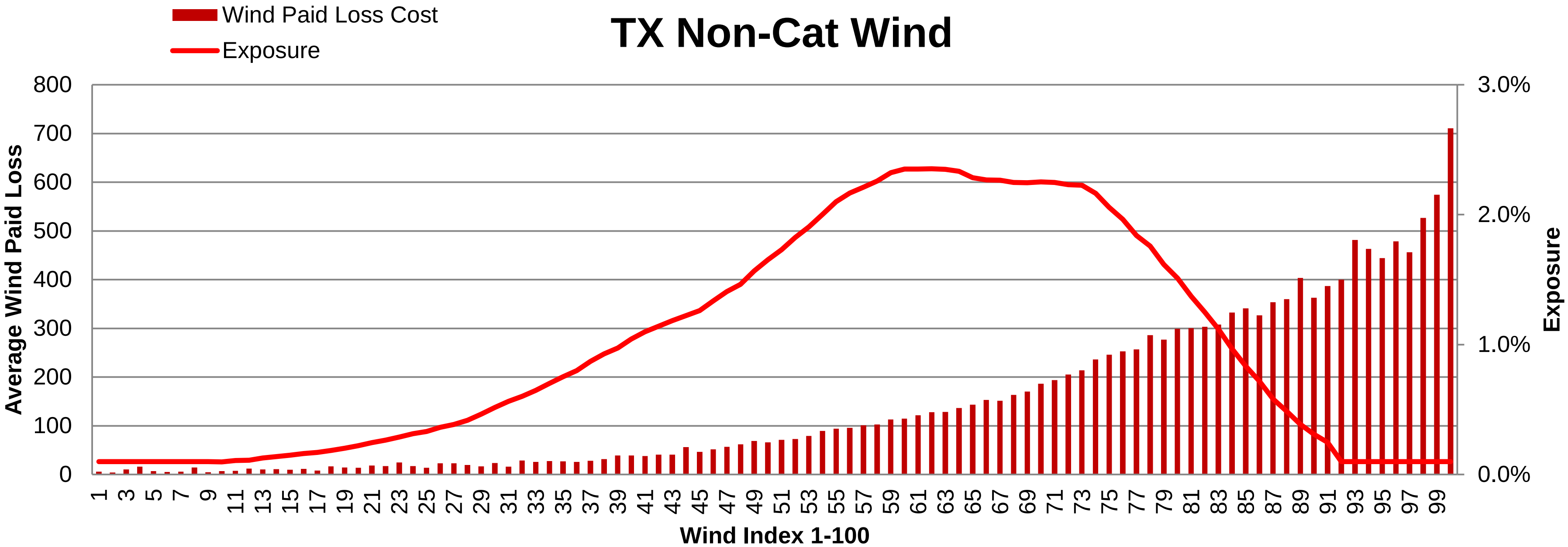


3. Good Data gives good models:

Houses

- in areas with many historic Wind & Hail Storms/Claim activity
- That have risky property characteristics

Tend to have high hail losses.



Conclusions

Smoothing methods create good data out of accurate garbage.

Consider smoothing methods whenever:

- Data is very predictive but very noisy
- Data is associated with a different class of objects
- Data is missing

Thank you

Scott Zrebiec, Ph.D.
Manager Statistical Modeling
LexisNexis Risk Solutions
scott.zrebiec@lexisnexis.com

References

- H. Bühlman, A. Gisler, *A course in Credibility theory and its applications*, Springer, 2008.
- T. Hastie, R. Tibshirani, J. Friedmann *Elements of Statistical Learning*, Springer, 2001.
- R. Bivand, E. Pebesma, V. Gomez-Rubio, *Applied Spatial Data Analysis with R*, Springer, 2008.