Application of Statistical Inferencing Techniques to Building Inventory Compilation

Pooya Sarabandi[1]

Anne S. Kiremidjian[1], Ronald T. Eguchi[2]

[1]*Advanced Global Risk Assessment Laboratory (AGRAL)*

Stanford University, USA

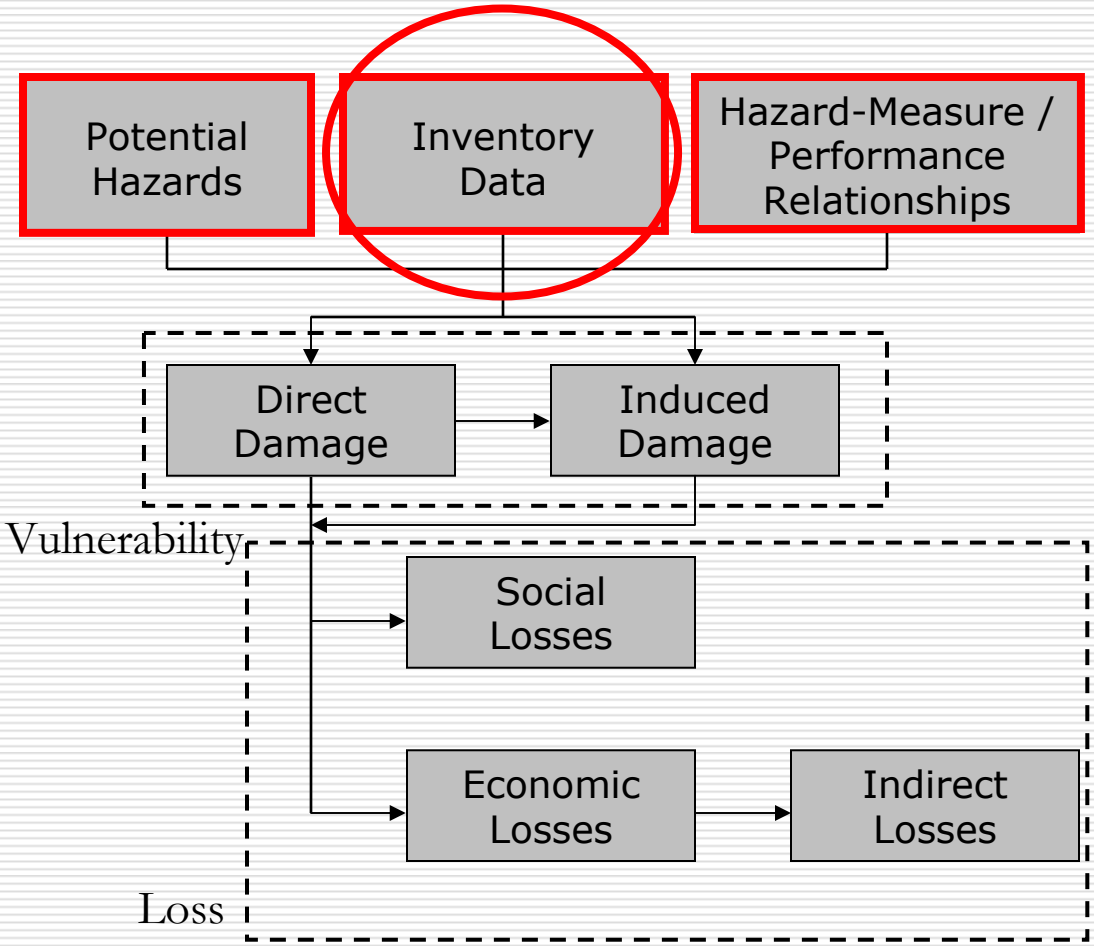
[2]*ImageCat Inc., Long Beach, CA, USA*

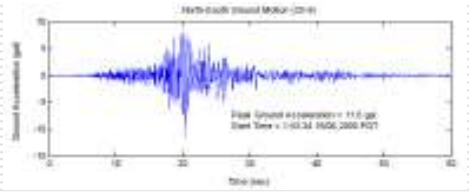
Outline

- Overview of Loss Estimation Methodologies
- Problem Statement and Objectives
- Statistical Structural Attribute Modeling
  - Classification Tree Models
  - Multinomial Logistic Models
  - Datasets Used to Train the Models
  - Examples

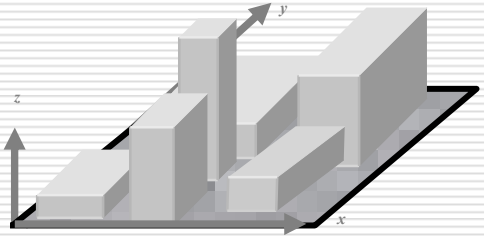
Overview of Loss Estimation Methodologies

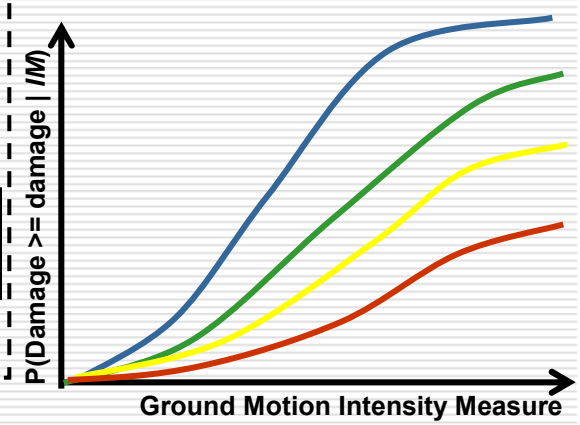
- **Overview**
- Problem Statement
- Statistical Structural Attribute Modeling





. Ground Shaking (*Sa*,*PGA*, *MMI*, …)
. Ground Failure (*PGD*, …)





- **Overview**
- Problem Statement
- Statistical Structural Attribute Modeling

Building Inventory Data Structure

- Address or Lon/Lat
- Height or Number of Stories
- Footprint Area
- Structural Type
- Occupancy Type
- Age or Year of Construction
- Roof Type and Rooftop Material
- Cladding or Façade Material
- Shape Irregularity
- Height Irregularity
- Mass Distribution
- Foundation Type
- Content Value
- Dominant orientation of city blocks

Building Inventory Attributes

- From Imagery (Remote Sensing)
  - *Location (Longitude and Latitude)*
  - *Height (Number of Stories)*
  - *Footprint Area*
  - *Perimeter and Total Cladding Area*
  - *Roof and Façade Material*
  - *Height and Plan (Shape) Irregularity*
- From Statistical Inferencing
  - *Structural Type* ✓
  - *Occupancy Type* ✓
- Other Sources
  - *Age or Year of Construction*
  - *Mass Distribution*
  - *Content Value*
  - *Foundation Type*

Problem Statement and Objectives

- Overview
- **Problem Statement**
- Statistical Structural Attribute Modeling

Develop a Statistical Framework to Infer Engineering Attributes of Buildings, Critical in Catastrophe Modeling (i.e. Occupancy Type and Structural Type) from Spatial, Geometric and Spectral Characteristics of the Buildings.

- Overview
- Problem Statement
- **Statistical Structural Attribute Modeling**

Statistical Structural Attribute Modeling

Known Attributes from 3-D models (Known)

- *Location (Longitude and Latitude)*
- *Height (Number of Stories)* ← Quantitative
- *Footprint Area* ← Quantitative
- *Perimeter and Total Cladding Area* ← Quantitative
- *Roof Type* ← Categorical
- *Height and Plan (Shape) Irregularity* ← Categorical

From Statistical Models (to be Predicted)

- *Structural Type* ← Categorical
- *Occupancy Type* ← Categorical

- Overview
- Problem Statement
- **Statistical Structural Attribute Modeling**

Statistical Structural Attribute Modeling

- Statistical Models Should Incorporate Categorical Data
- Response Variables Are Not Normally Distributed
  - In Case of Two Categories: Response Variables Have A Binomial Distribution

$$P(y_i = 1 \mid \mathbf{x_i}) = \pi$$
$$P(y_i = 0 \mid \mathbf{x_i}) = 1 - \pi$$

  - In Case of Multiple Categories: Response Variables Have A Multinomial Distribution

$$P(Y_{i1} = y_{i1}, \ldots, Y_{ik} = y_{ik} \mid \mathbf{x_i}) = \frac{n_i}{y_{i1}! \ldots y_{ik}!} \pi_{i1}^{y_{i1}} \ldots \pi_{ik}^{y_{ik}}$$

where $n_i = \sum_K Y_{ik}$

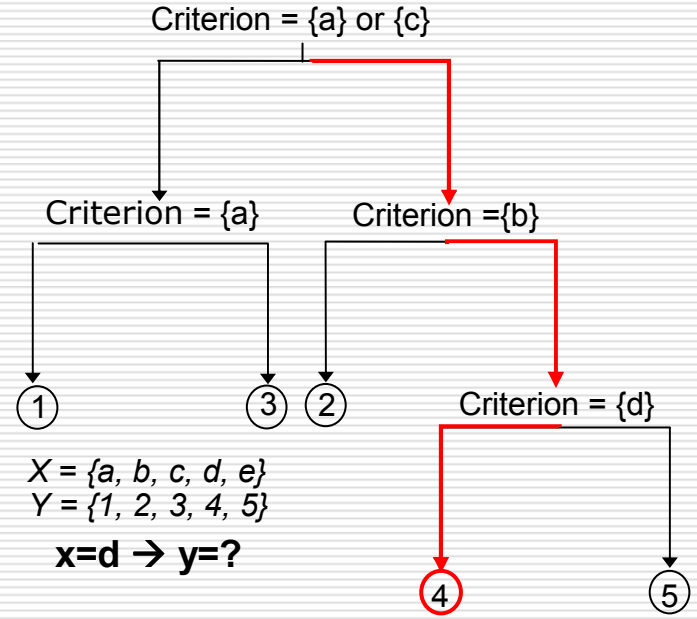
- Implemented Models:
  - Classification Tree Models
  - Multinomial Logistic Models

Statistical Structural Attribute Modeling

- Overview
- Problem Statement
- **Statistical Structural Attribute Modeling**
  - **Classification Tree Models**
  - Multinomial Logistic Models
  - Datasets
  - Examples

Classification Tree Models

- Attractive Knowledge Discovery Tools for Initial Data Exploration
- Simple
- Identify Most Important Variables in the Data
- Optimal at Each Node (the overall tree might not be optimal)
- Deviance Measure in Regression Trees: "Sum of the Squares of Residuals" of the Response Variable
- Splitting Rule: Values of Independent Variables which Give the Maximum Reduction to the Deviance of the Response Variable
- Deviance Measure in Classification Trees: *Information Index* As the Deviance (Impurity) Measure



$$D = \sum_i (y_i - \bar{y})^2$$

$$D = \sum_i (y_i - \bar{y})^2$$

$$\sum_{k \neq k'} p_{mk} p_{mk'} = \sum_{k=1}^{K} p_{mk}(1 - p_{mk}) = 1 - \sum_{k=1}^{K} p_{mk}^2$$

$$= \sum_{i \in L} (y_{Li} - \bar{y}_L)^2 + \sum_{i \in R} (y_{Ri} - \bar{y}_R)^2 - \left[N_L (\bar{y}_L - \bar{y})^2 + N_R (\bar{y}_R - \bar{y})^2 \right]$$

← *new within group dev.*

← *between group dev.*

$$\min_{j,s} \left[\min_j \left(\sum_{i|(x_i)_j \in s} (y_{Li} - \bar{y}_{Lj})^2 \right) + \min_j \left(\sum_{i|(x_i)_j \in s} (y_{Rj} - \bar{y}_{Rj})^2 \right) \right]$$

$$\sum_{i|(x_i)_j \in s} p_{mk} = \frac{n_{mk}}{N_m}$$

Statistical Structural Attribute Modeling

- Overview
- Problem Statement
- **Statistical Structural Attribute Modeling**
  - Classification Tree Models
  - **Multinomial Logistic Models**
  - Datasets
  - Examples

Multinomial Logistic Model

- Generalization of Binomial Models
- Variance of Response Variable Is Dependent on Its Mean Value
- Underlying Probability Distribution is Multinomial Such that: $P(y_i = k \mid \mathbf{x_i}) = \pi_k(\mathbf{x_i}) = \pi_{ik}$
- Log-Odds Ratio Can Be Modeled As A Linear Function of Input Variables

$$\ln\left(\frac{\pi_{ik}}{\pi_{iK}}\right) = \sum_j \beta_{jk} x_{ij} = \alpha_k + \beta_{1k} x_{i1} + \ldots + \beta_{pk} x_{ip}$$

- Probability Model:

$$\pi_{ik} = \frac{\exp\left(\sum_j \beta_{jk} x_{ij}\right)}{1 + \sum_{k=1}^{K-1} \exp\left(\sum_j \beta_{jk} x_{ij}\right)} \; ; \; i = 1:N \, , \, j = i:p$$

Statistical Structural Attribute Modeling

- Overview
- Problem Statement
- **Statistical Structural Attribute Modeling**
  - Classification Tree Models
  - **Multinomial Logistic Models**
  - Datasets
  - Examples

Parameter Estimation:

Maximization of the Log-Likelihood of the model over all of the independent observations

$$\ell_i = \ln[f(y_i \mid \theta_i, \phi)] = \frac{y_i\theta_i - a(\theta_i)}{b(\phi)} + h(y_i, \phi)$$

$$a(\theta_i) = -n\ln(1-\pi)\ ,\ b(\phi) = 1\ ,\ h(y_i, \phi) = \ln\binom{n}{y_i}$$

$$L(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{N} \ell_i = \sum_{i=1}^{N} \frac{y_i\theta_i - a(\theta_i)}{b(\phi)} + \sum_{i=1}^{N} h(y_i, \phi)$$

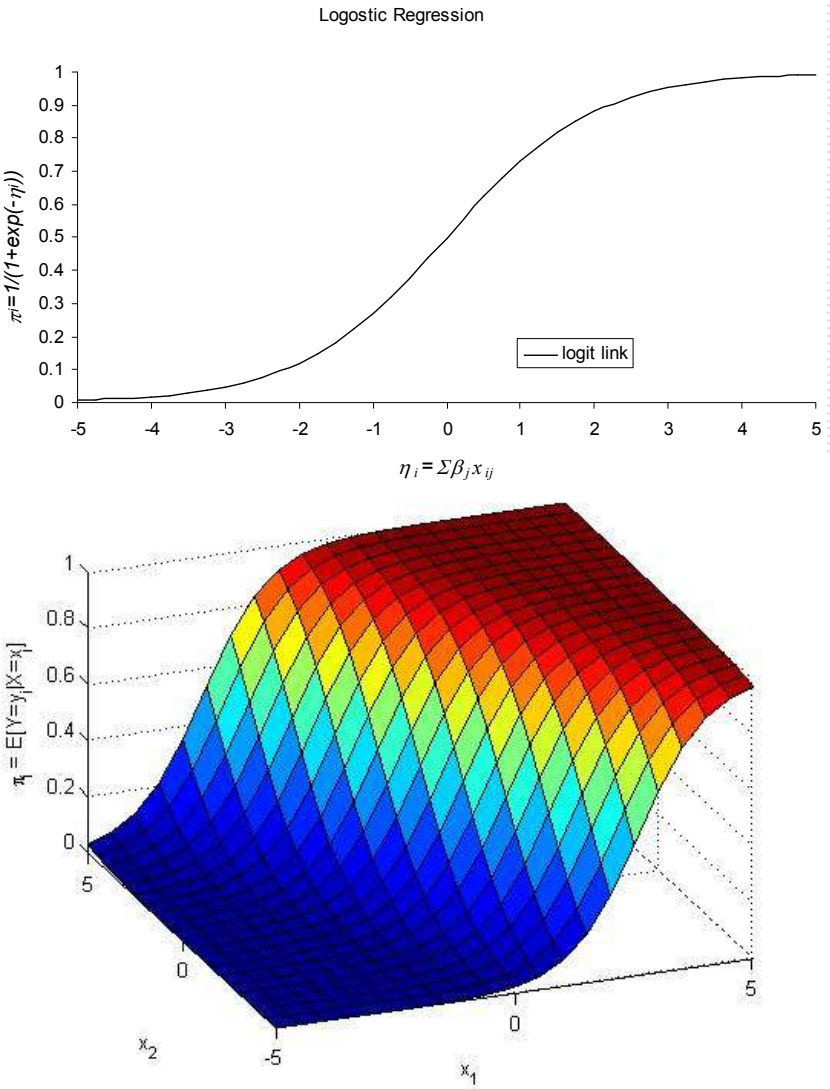
$$\max(L(\hat{\boldsymbol{\beta}})) \rightarrow E[\hat{\boldsymbol{\beta}}] = \beta$$

Prediction:

Calculating probability of falling within different classes of the response variable

$$\Pr(y \mid X = x) = \boldsymbol{\Pi} = \{\pi_1, \pi_2, ..., \pi_K\}$$

$$\Rightarrow y \in C \mid \max\{\boldsymbol{\Pi}\} = \pi_C$$

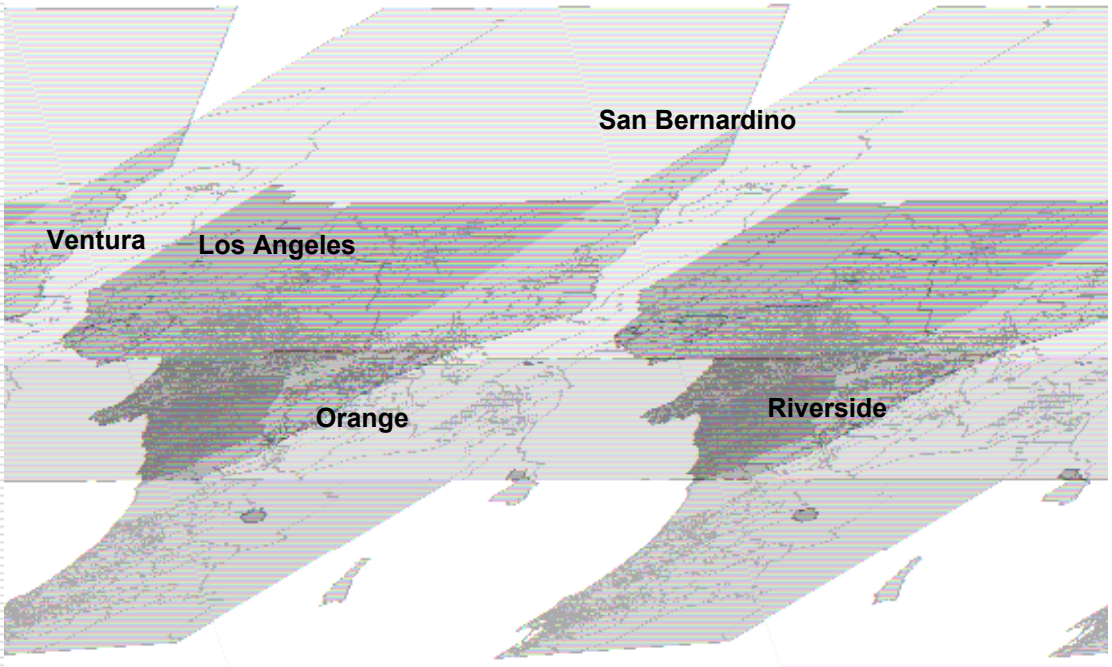


Statistical Structural Attribute Modeling

- Overview
- Problem Statement
- **Statistical Structural Attribute Modeling**
  - Classification Tree Models
  - Multinomial Logistic Models
  - **Datasets**
  - Examples

Dataset 1:

- Southern California
- Aggregate Tax Assessor Data
- Census Tract Level
- 1,570 Tracts
- 38,135 Buildings
- Attributes:
  - **Structural Type**
  - **Occupancy Type**
  - **Average Area**
  - **Height**
  - **Age**

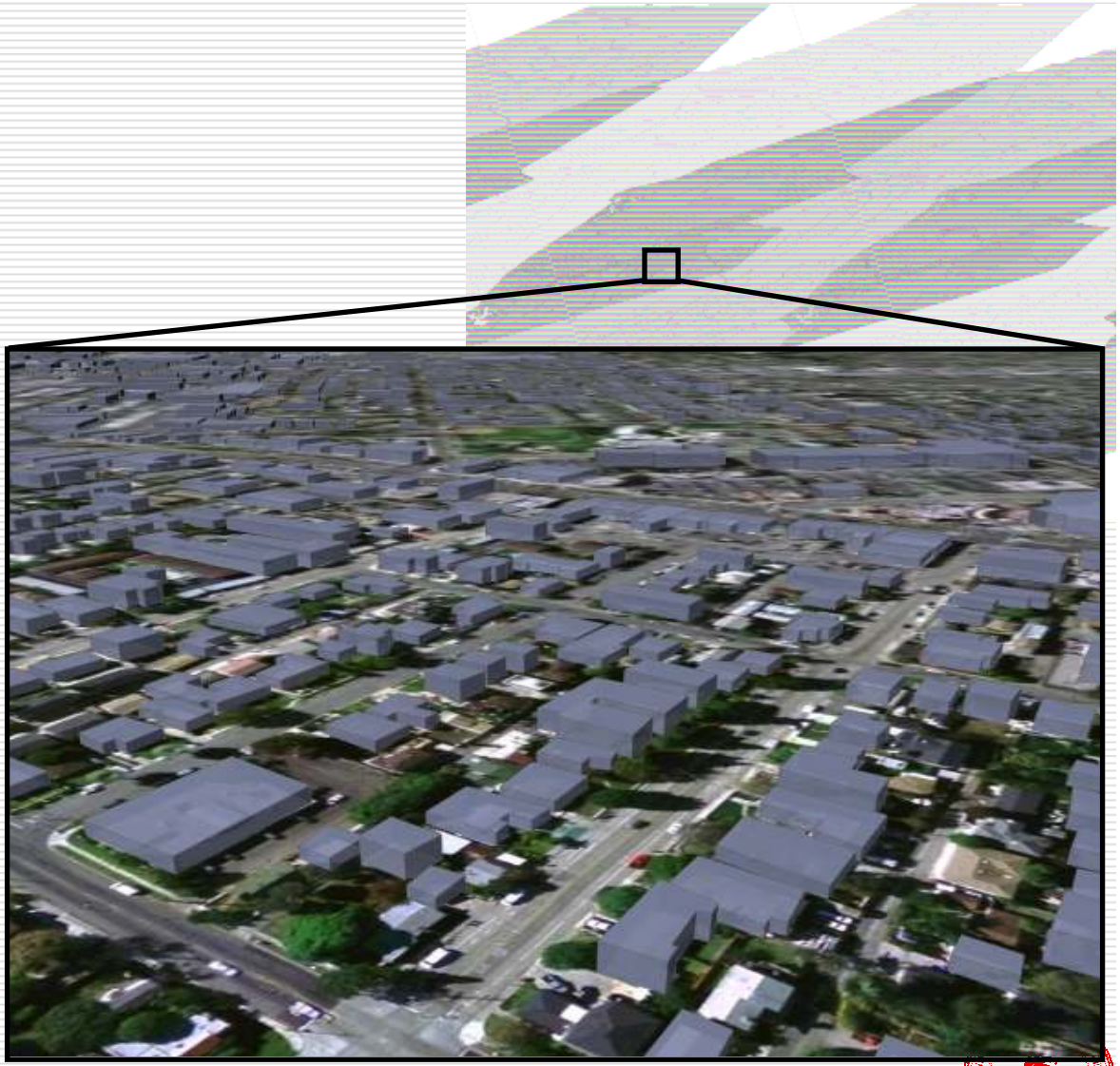


Statistical Structural Attribute Modeling

- Overview
- Problem Statement
- **Statistical Structural Attribute Modeling**
  - Classification Tree Models
  - Multinomial Logistic Models
  - **Datasets**
  - Examples

Dataset 2:

- Southern California
- Building Level
- 1,947 Building
- Attributes from 3-D building Models:
  - **Height**
  - **Area**
  - **Irregularity**
  - **Roof type**
- Attributes from Assessor Files:
  - **Structural Type**
  - **Occupancy Type**
  - **Age**



Statistical Structural Attribute Modeling

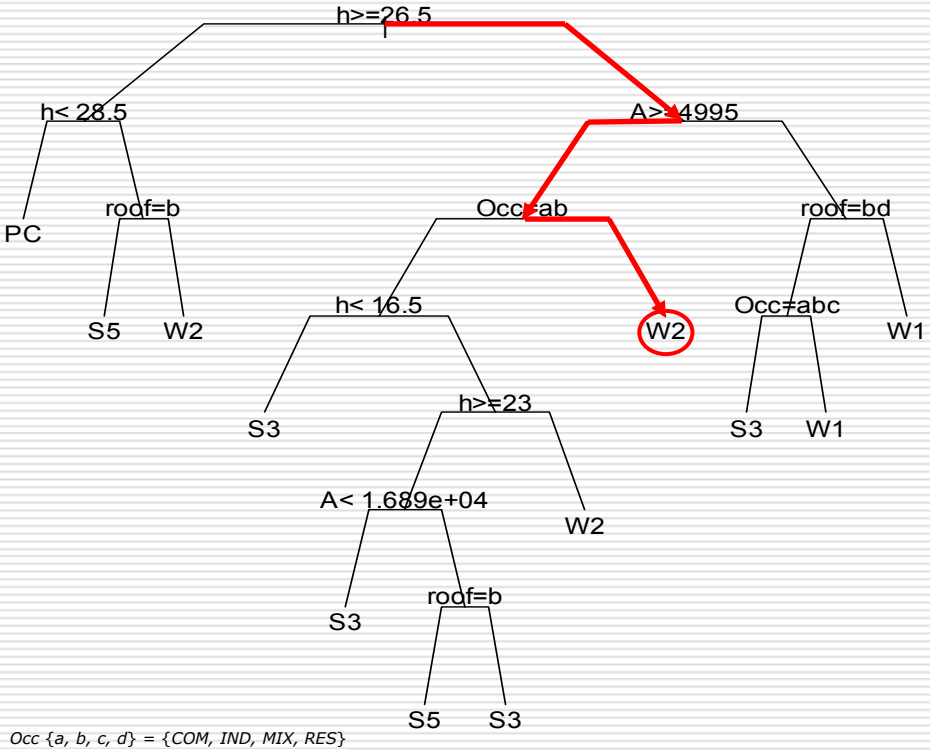
- Overview
- Problem Statement
- **Statistical Structural Attribute Modeling**
  - Classification Tree Models
  - Multinomial Logistic Models
  - Datasets
  - **Examples**

Sample Classification Tree from Dataset

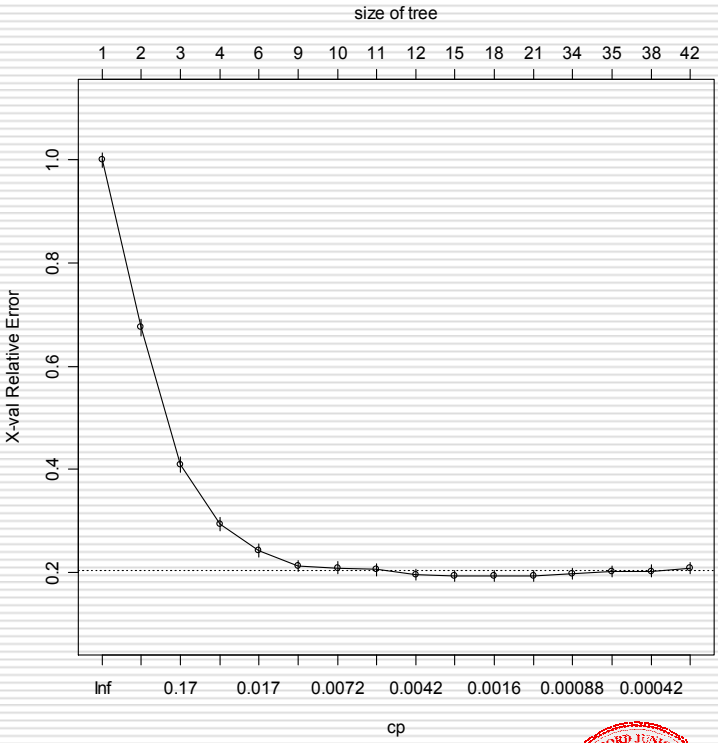
- Response Variable: Structural Type
- Independent Variables: Height, Area, Irregularity, Roof type, Occupancy Type

Example: Mid-rise, 5500 sqft, Residential → Structural Type?

Prediction Accuracy: 88.54%



Occ {a, b, c, d} = {COM, IND, MIX, RES}

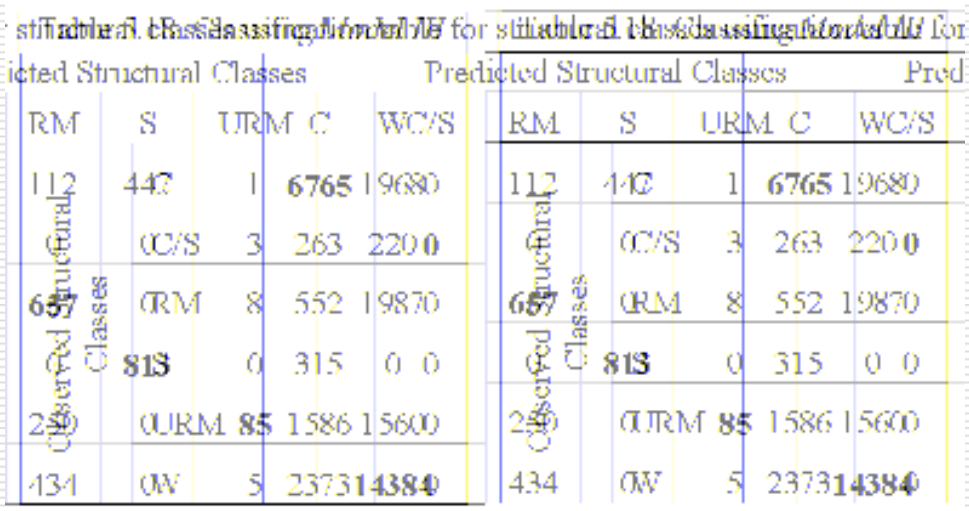


Statistical Structural Attribute Modeling

- Overview
- Problem Statement
- **Statistical Structural Attribute Modeling**
  - Classification Tree Models
  - Multinomial Logistic Models
  - Datasets
  - **Examples**

Prediction Accuracy Assessment

- Stratified 10-fold cross-validation scheme
- Prediction Error = Average and Standard Deviation Error Using Prediction Error of Each of the 10 Folds



Observed Structural Classes	RM	S	URM	C	WC/S
RM	112	442	1	6765	19680
S	0	C/S	3	263	2200
C	657	CRM	8	552	19870
WC/S	813	0	315	0	0
URM	250	URM	85	1586	15600
MW	434	MW	5	2373	14384

$$\varepsilon = 1 - \frac{sum[diag(T)]}{sum[T]}$$

Statistical Structural Attribute Modeling

- Overview
- Problem Statement
- **Statistical Structural Attribute Modeling**
  - Classification Tree Models
  - Multinomial Logistic Models
  - Datasets
  - **Examples**

Sample Multinomial Logistic Model from Dataset

- Response Variable: Structural Type
- Independent Variables: Height Category (L,M,H), Area

$$
\begin{cases}
\ln\left(\frac{\pi_{C/S}}{\pi_C}\right) = -2.216 - 0.324x_1 - 0.572x_2 - 0.000013x_3 \\
\ln\left(\frac{\pi_{RM}}{\pi_C}\right) = -0.574 + 0.337x_1 - 0.355x_2 - 0.000012x_3 \\
\ln\left(\frac{\pi_S}{\pi_C}\right) = -1.154 - 0.130x_1 - 1.502x_2 + 0.0000084x_3 \\
\ln\left(\frac{\pi_{URM}}{\pi_C}\right) = -0.401 + 0.382x_1 - 0.388x_2 - 0.000017x_3 \\
\ln\left(\frac{\pi_W}{\pi_C}\right) = -8.906 + 11408x_1 + 9.099x_2 - 0.000049x_3
\end{cases}
$$

Explanatory Variable	x_1	x_2
Height (High-rise)	0	0
Height (Medium-rise)	1	0
Height (Low-rise)	0	1

Statistical Structural Attribute Modeling

- Overview
- Problem Statement
- **Statistical Structural Attribute Modeling**
  - Classification Tree Models
  - Multinomial Logistic Models
  - Datasets
  - **Examples**

Example: Low-rise Structure with Average Area of 2176 sqft → Structural Type?

$$\begin{cases} \frac{\pi_{C/S}}{\pi_C} = e^{-2.817401} = 0.05976103 \\ \frac{\pi_{RM}}{\pi_C} = e^{-0.9549136} = 0.3848454 \\ \frac{\pi_{RM}}{\pi_C} = e^{-0.9549136} = 0.3848454 \\ \frac{\pi_{URM}}{\pi_C} = e^{-0.8268078} = 0.4374435 \\ \frac{\pi_W}{\pi_C} = e^{+0.8956142} = 2.4488395 \end{cases} \Rightarrow \begin{cases} \pi_C = \frac{1}{1+e^{-2.817401}+e^{-0.9549136}+e^{-2.63715737}+e^{-0.8268078}+e^{+0.8956142}} = 0.2271515 \\ \pi_{C/S} = \frac{e^{-2.817401}}{1+e^{-2.817401}+e^{-0.9549136}+e^{-2.63715737}+e^{-0.8268078}+e^{+0.8956142}} = 0.01357467 \\ \pi_{RM} = \frac{e^{-0.9549136}}{1+e^{-2.817401}+e^{-0.9549136}+e^{-2.63715737}+e^{-0.8268078}+e^{+0.8956142}} = 0.0874174 \\ \pi_S = \frac{e^{-2.63715737}}{1+e^{-2.817401}+e^{-0.9549136}+e^{-2.63715737}+e^{-0.8268078}+e^{+0.8956142}} = 0.01625607 \\ \pi_{URM} = \frac{e^{-0.8268078}}{1+e^{-2.817401}+e^{-0.9549136}+e^{-2.63715737}+e^{-0.8268078}+e^{+0.8956142}} = 0.09936458 \\ \pi_W = \frac{e^{+0.8956142}}{1+e^{-2.817401}+e^{-0.9549136}+e^{-2.63715737}+e^{-0.8268078}+e^{+0.8956142}} = 0.5562357 \end{cases}$$

$$\Rightarrow \Pi = \{0.2272 \ , \ 0.01357 \ , \ 0.0874 \ , \ 0.0163 \ , \ 0.0994 \ , \ 0.5562\}$$

$$\Pi = \{\pi_C, \pi_{C/S}, \pi_{RM}, \pi_S, \pi_{URM}, \pi_W\}$$

Example: High-rise, Commercial, with Average Area of 75'000sqft built in 1974 → Structural Type?

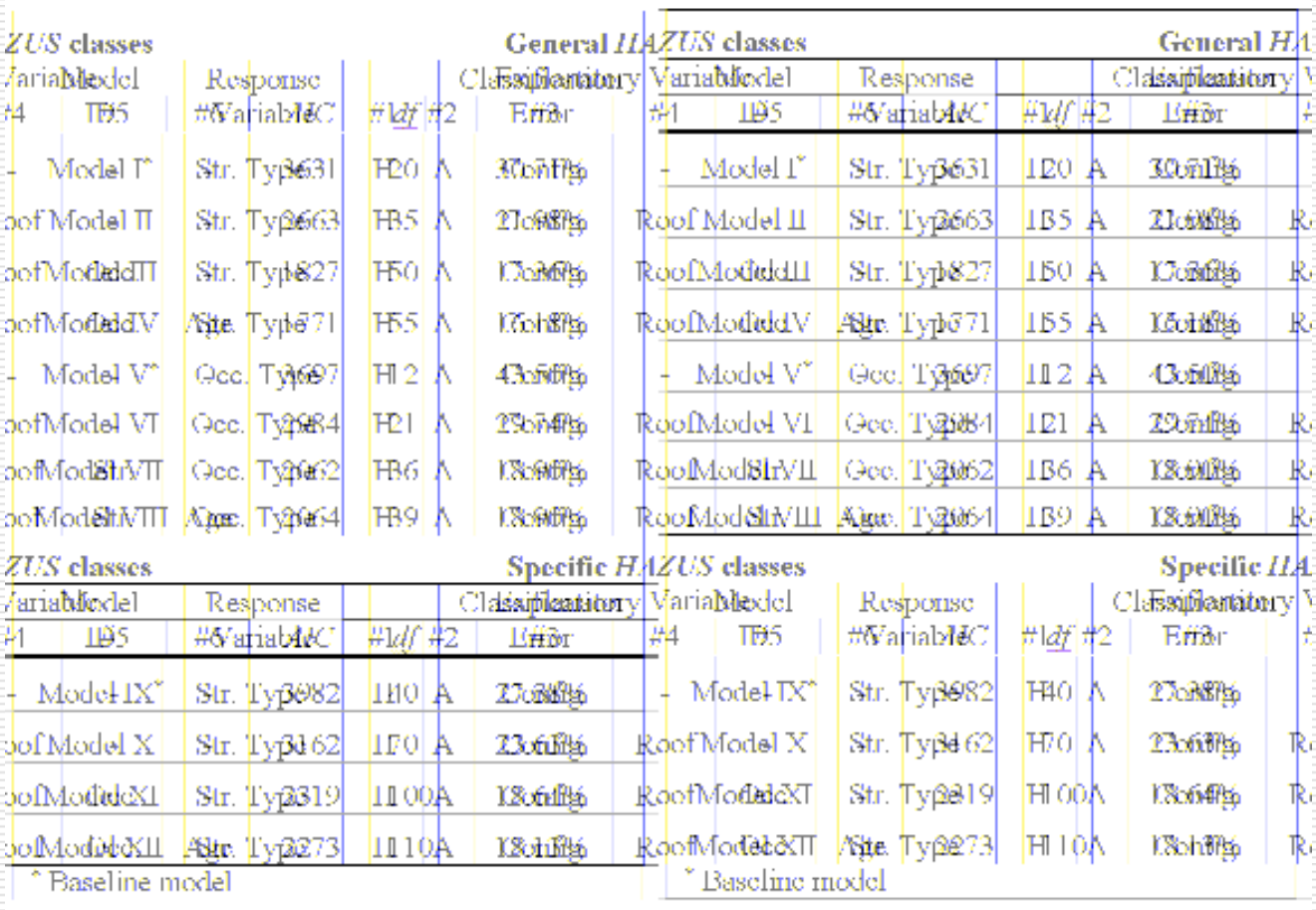
$$\Rightarrow \Pi = \{0.4837 \ , \ 1.1e-23 \ , \ 5.6e-20 \ , \ 0.5163 \ , \ 1.0e-21 \ , \ 1.2e-55\}$$

$$\Pi = \{\pi_C, \pi_{C/S}, \pi_{RM}, \pi_S, \pi_{URM}, \pi_W\}$$

Statistical Structural Attribute Modeling

- Overview
- Problem Statement
- **Statistical Structural Attribute Modeling**
  - Classification Tree Models
  - Multinomial Logistic Models
  - Datasets
  - **Examples**

Summary of Classification Accuracy Using Multinomial Logistic Regression and Dataset 2



ZUS classes						General HA
Variable	Model	Response			Classification	Explanatory
#4	#5	#6 Variable	#1 df	#2	Error	#3
-	Model I^	Str. Type	3631	H20	A	30.71%
Roof	Model II	Str. Type	2663	H35	A	21.98%
Roof	Model III	Str. Type	1827	H50	A	17.36%
Roof	Model IV	Str. Type	1771	H55	A	16.18%
-	Model V^	Occ. Type	3697	H12	A	43.50%
Roof	Model VI	Occ. Type	2984	H21	A	29.74%
Roof	Model VII	Occ. Type	2062	H36	A	18.90%
Roof	Model VIII	Occ. Type	2064	H39	A	18.90%

ZUS classes						General HA
Variable	Model	Response			Classification	Explanatory
#1	#5	#6 Variable	#1 df	#2	Error	#3
-	Model I^	Str. Type	3631	H20	A	30.71%
Roof	Model II	Str. Type	2663	H35	A	21.98%
Roof	Model III	Str. Type	1827	H50	A	17.36%
Roof	Model IV	Str. Type	1771	H55	A	16.18%
-	Model V^	Occ. Type	3697	H12	A	43.50%
Roof	Model VI	Occ. Type	2984	H21	A	29.74%
Roof	Model VII	Occ. Type	2062	H36	A	18.90%
Roof	Model VIII	Occ. Type	2064	H39	A	18.90%

ZUS classes						Specific HA
Variable	Model	Response			Classification	Explanatory
#1	#5	#6 Variable	#1 df	#2	Error	#3
-	Model IX^	Str. Type	3982	H40	A	27.38%
Roof	Model X	Str. Type	3162	H70	A	23.63%
Roof	Model XI	Str. Type	2319	H100	A	18.64%
Roof	Model XII	Str. Type	2273	H110	A	18.13%

^ Baseline model

ZUS classes						Specific HA
Variable	Model	Response			Classification	Explanatory
#4	#5	#6 Variable	#1 df	#2	Error	#3
-	Model IX^	Str. Type	3982	H40	A	27.38%
Roof	Model X	Str. Type	3162	H70	A	23.63%
Roof	Model XI	Str. Type	2319	H100	A	18.64%
Roof	Model XII	Str. Type	2273	H110	A	18.13%

^ Baseline model

- Overview
- Problem Statement
- Statistical Structural Attribute Modeling

Statistical Structural Attribute Modeling

Application in probabilistic seismic hazard analysis (*PSHA*)

$$\nu_L = P\left[L > l \mid t \in (0,1]\right]$$

$$= \sum_{\substack{\text{all the} \\ \text{events } i}} \nu_i \Big[\sum_{\substack{\text{all Occ.} \\ \text{classes}}} \sum_{\substack{\text{all Str.} \\ \text{classe}}} \int\int\int\int P\left[L > l \mid D, OCC\right] P\left[OCC = occ, STR = str \mid X_o\right]$$

$$\left. f_{D|IM,STR}(d|im,str).f_{IM|R,M}(im|r,m).f_{R|M}(r|m).f_M(m).dd.dim.dr.dm \right]$$

ν_L	=	Annual rate of exceeding a certain level of loss
L	=	Loss (or other risk metric)
t	=	Time (years)
ν_i	=	Mean annual rate of occurrence of earthquakes generated by source i
D	=	Damage state
OCC	=	Occupancy type
STR	=	Structural type
X_o	=	Spatial, geometric and spectral attributes from 3-D models and building inventories
IM	=	Ground motion intensity measure
R	=	Distance from source to site
M	=	Magnitude of event
$f_X(x)$	=	Probability density function of variable X

THANK YOU