Application of Statistical Inferencing Techniques to Building Inventory Compilation

Pooya Sarabandi¹
Anne S. Kiremidjian¹, Ronald T. Eguchi²

¹Advanced Global Risk Assessment Laboratory (AGRAL)
Stanford University, USA

²ImageCat 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

- Potential Hazards
- Inventory Data
- Hazard-Measure / Performance Relationships

- Direct Damage
- Induced Damage

- Vulnerability
  - Social Losses
  - Economic Losses
  - Indirect Losses

- Loss

- Ground Shaking (Sa, PGA, MMI, ...)
- Ground Failure (PGD, ...)

5th International Workshop on Remote Sensing Applications to Natural Hazards
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

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.
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
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
    \[
    P(y_i = 1 \mid x_i) = \pi \\
    P(y_i = 0 \mid x_i) = 1 - \pi
    \]
  - In Case of Multiple Categories: Response Variables Have A Multinomial Distribution
    \[
    P(Y_{i1} = y_{i1}, \ldots, Y_{ik} = y_{ik} \mid x_i) = \frac{n_i}{y_{i1}! \cdots y_{ik}!} \pi_{i1}^{y_{i1}} \cdots \pi_{ik}^{y_{ik}}
    \]
    where \( n_i = \sum_k Y_{ik} \)
  - Implemented Models:
    - Classification Tree Models
    - Multinomial Logistic 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

\[
D = \sum (y_i - \bar{y})^2
\]

\[
D = \sum p_{mk} p_{mk'} = \sum L (y_{Li} - \bar{y}_L)^2 + \sum R (y_{Ri} - \bar{y}_R)^2 - \sum L (\bar{y}_L - \bar{y})^2 - \sum R (\bar{y}_R - \bar{y})^2
\]

\[
\min \left( \min \left( \sum_{i \in S} (y_{Li} - \bar{y}_L)^2 \right), \min \left( \sum_{j \in S} (y_{Ri} - \bar{y}_R)^2 \right) \right)
\]
Statistical Structural Attribute Modeling

**Multinomial Logistic Model**

- Generalization of Binomial Models
- Variance of Response Variable Is Dependent on Its Mean Value
- Underlying Probability Distribution is Multinomial Such that: \( P(y_i = k | x_i) = \pi_k(x_i) = \pi_{ik} \)
- 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:

\[
\pi_{ik} = \frac{\exp(\sum_j \beta_{jk}x_{ij})}{\sum_{k=1}^{K-1} \exp(\sum_j \beta_{jk}x_{ij})}; \ i = 1:N, \ j = i:p
\]
Statistical Structural Attribute Modeling

**Parameter Estimation:**

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

\[ \ell_i = \ln \left[ f(y_i | \theta_i, \phi) \right] = \frac{y_i \theta_i - a(\theta)}{b(\phi)} + h(y_i, \phi) \]

\[ a(\theta) = -n \ln(1 - \pi) \quad b(\phi) = 1 \quad h(y_i, \phi) = \ln \binom{n}{y_i} \]

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

Maximization leads to

\[ \max(L(\hat{\beta})) \Rightarrow E[\hat{\beta}] = \beta \]

**Prediction:**

Calculating probability of falling within different classes of the response variable

\[ \Pr(y | X = x) = \Pi = \{\pi_1, \pi_2, ..., \pi_K\} \]

\[ \Rightarrow y \in C | \max \{\Pi\} = \pi_C \]
Statistical Structural Attribute Modeling

**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

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
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%
Statistical Structural Attribute Modeling

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

\[ \varepsilon = 1 - \frac{\text{sum}[\text{diag}(T)]}{\text{sum}[T]} \]
Statistical Structural Attribute Modeling

Sample Multinomial Logistic Model from Dataset

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

\[
\begin{align*}
\ln\left(\frac{\pi_{C/S}}{\pi_C}\right) &= -2.216 - 0.324x_1 - 0.572x_2 - 0.000013x_3 \\
\ln\left(\frac{\pi_{RM}}{\pi_C}\right) &= -0.574 + 0.337x_1 - 0.355x_2 - 0.000012x_3 \\
\ln\left(\frac{\pi_S}{\pi_C}\right) &= -1.154 - 0.130x_1 - 1.502x_2 + 0.000084x_3 \\
\ln\left(\frac{\pi_{URM}}{\pi_C}\right) &= -0.401 + 0.382x_1 - 0.388x_2 - 0.000017x_3 \\
\ln\left(\frac{\pi_W}{\pi_C}\right) &= -8.906 + 11408x_1 + 9.099x_2 - 0.000049x_3
\end{align*}
\]

<table>
<thead>
<tr>
<th>Explanatory Variable</th>
<th>$x_1$</th>
<th>$x_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height (High-rise)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Height (Medium-rise)</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Height (Low-rise)</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Statistical Structural Attribute Modeling

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

\[
\begin{align*}
\frac{\pi_{C/S}}{\pi_C} &= e^{-2.817401} = 0.05976103 \\
\frac{\pi_{RM}}{\pi_C} &= e^{-0.9549136} = 0.3848454 \\
\frac{\pi_{URM}}{\pi_C} &= e^{-0.8268078} = 0.4374435 \\
\frac{\pi_{W}}{\pi_C} &= e^{+0.8956142} = 2.4488395
\end{align*}
\]

\[
\Pi = \{0.2272, 0.01357, 0.0874, 0.0163, 0.0994, 0.5562\}
\]

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

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

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

\[
\Pi = \\{\pi_C, \pi_{C/S}, \pi_{RM}, \pi_S, \pi_{URM}, \pi_W\}
\]
Statistical Structural Attribute Modeling

Summary of Classification Accuracy Using Multinomial Logistic Regression and Dataset 2
Statistical Structural Attribute Modeling

Application in probabilistic seismic hazard analysis (PSHA)

\[ v_L = P[L > l \mid i \in (0,1)] \]

\[ = \sum_{i} v_i \left[ \sum_{\text{all Occ. all Str. classes}} \sum_{\text{all the events}} \int \int \int P[L > l \mid D, OCC] P[OCC = occ, STR = str \mid X_{o} ] \times f_{D\mid IM, STR}(d\mid im, str).f_{IM\mid R,M}(im \mid r, m).f_{R\mid M}(r \mid m).f_{M}(m).dd .dim .dr .dm \right] \]

- \( v_L \) = Annual rate of exceeding a certain level of loss
- \( L \) = Loss (or other risk metric)
- \( t \) = Time (years)
- \( v_i \) = 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