Sampling Scheme for Multiple Asset Management Using Locality-Sensitive Hashing (LSH) Algorithm

Zhuo Chen, Xiaoyue Cathy Liu
University Of Utah
Presentation Outline

- Project Overview
- Research Objectives
- Background
- Methodology
- Data Collection and Result Analysis
- Summary
Overview: General Information

- Infrastructure management:
  - Inspection
  - Maintenance
  - Rehabilitation

- Most of current sampling methods consider **randomness** in sampling process for **only one type** of infrastructures.
**Background: Existing Sampling Methods**

Most of current sampling methods consider *randomness* in sampling process for *only one type* of infrastructure. (Schmitt et.al 2006)

<table>
<thead>
<tr>
<th>State</th>
<th>Sampling Selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>California</td>
<td>Randomly selected sample</td>
</tr>
<tr>
<td>Florida</td>
<td>Random Number Generator</td>
</tr>
<tr>
<td>Indiana</td>
<td>Random Number Generator</td>
</tr>
<tr>
<td>New York</td>
<td>Selected by field personnel</td>
</tr>
</tbody>
</table>

**Utah**
Research Objectives

- Develop **sampling scheme** for MMQA to effectively measure **multiple types** of infrastructure conditions and return Level of Maintenance (LOM) grades

- Develop methodological framework to analyze the measured samples in order to evaluate the **accuracy of samples** representing all assets
Methodology

◊ Clustering sampling with high-dimensional data

Population (size N)

Cluster 1 → Cluster 2 → Cluster 3 → ... → Cluster n

Sample 1 Sample 2 Sample 3 Sample 4 Sample 5 Sample 6 Sample 7 Sample 8 ...

◊ Curse of Dimensionality
Locality-Sensitive Hashing (LSH)

- **Locality-Sensitive Hashing (LSH) algorithm**
  - Searching similarity for high-dimensional data
  - Calculate the similarity based on the probability of collision
  - **Hash Function**: Map data of arbitrary size to data of fixed size

- Collision is defined as two vectors assigned to the same bucket or neighboring buckets after applied to the same hash function
Probability of Collision

- Segment Conditions with Letter Grade
  - Segment 1: A+ B+ A- ...
  - Segment 2: A- A+ C ...
  - Segment 3: A+ C A+ ...
  - Segment 4: C A B- ...
  - Segment N: A- C- F ...

- Segment Conditions with Number Grade
  - Segment 1: 16 13 14 ...
  - Segment 2: 14 16 9 ...
  - Segment 3: 16 9 16 ...
  - Segment 4: 9 15 11 ...
  - Segment N: 14 8 2 ...

- Hash Function
  - Bucket 1: 0 1 1 0 0 1 1 1
  - Bucket 2: 1 1 1 0 0 1 1 1
  - Bucket 3: 1 0 1 0 1 0 0 1
  - Bucket k: 1 0 0 0 1 1 0 0

- Estimate the probability by repeating applying different hash functions
Clustering Sampling

◊ Hash Functions:
  ◊ Map the vectors in a M-dimension space
  ◊ M-bit binary string based on the projected value in each dimension

◊ Similarity matrix:

|     | Seg₁ | Seg₂  | Seg₃  | Seg₄  | ...
|-----|------|-------|-------|-------|------
| Seg₁| 1    | 0.215 | 0.781 | 0.241 |      
| Seg₂| 0.215| 1     | 0.11  | 0.762 |      
| Seg₃| 0.781| 0.11  | 1     | 0.02  |      
| Seg₄| 0.24 | 0.762 | 0.02  | 1     |      
|     |      |       |       |       | ...  

◊ Clustering Methods: Spectral Clustering
Sampling Result Evaluation

- Average grade Vs. Grade Distribution

- RMSE (Root-Mean-Square Error): \[ RMSE_i = \sqrt{\frac{\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2}{n}} \]

- Root-Square Error
Data Collection

✧ Inspection records provided by Maintenance Management Quality Assurance (MMQA) program

✧ Utah highway network

✧ 17 measurement activities: snow and ice, litter pickup, vegetation, etc

✧ Maintenance performance expressed as 15 different letter grade (A+ to F-)

✧ Sampling Process: 50 times
Result Analysis

- Using simple random sampling and clustering sampling based on one type of infrastructure as comparison groups
Result Analysis
Result Analysis

[Bar chart with categories like Shoulder Work, Curb and Gutter, Litter Pickup, Weed Control, Grade and Clean Ditches, Maintain Inlets, Repair and Replace Signs, Repair and Replace Delineation, Guardrail Maintenance, Sweeping, Vegetation Control, Fence Maintenance, etc., with two bars for Simp_Rand and LSH, showing normalized percentage values between 0.00 and 0.08.]
Result Analysis

[Bar chart showing comparisons between Simp_Rand and LSH for various tasks including Shoulder Work, Curb and Gutter, Litter Pickup, Weed Control, Grade and Clean Ditches, Maintain Inlets, Repair and Replace Delineation, Guardrail Maintenance, Sweeping, Vegetation Control, and Fence Maintenance.]
Result Analysis

[Graph showing bar charts and line graphs with various categories on the x-axis and normalized percentage on the y-axis.]
Summary

- Sampling scheme for selecting samples which represent all types of infrastructures
- For most types of infrastructure sampling, LSH method is better than simple random sampling and clustering sampling based on one type of infrastructure
- LSH method is less sensitive with fluctuation in grade distribution
- Future work:
  - Grade prediction using Latent Markov Decision Process
  - Combine LSH sampling method with other methods (Simple Random, Spatial-balance, etc)
Thanks for your attention!