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

Zhuo Chen, Xiaoyue Cathy Liu
University Of Utah

Utah Traffic Lab

UDOT Keeping Utah Moving

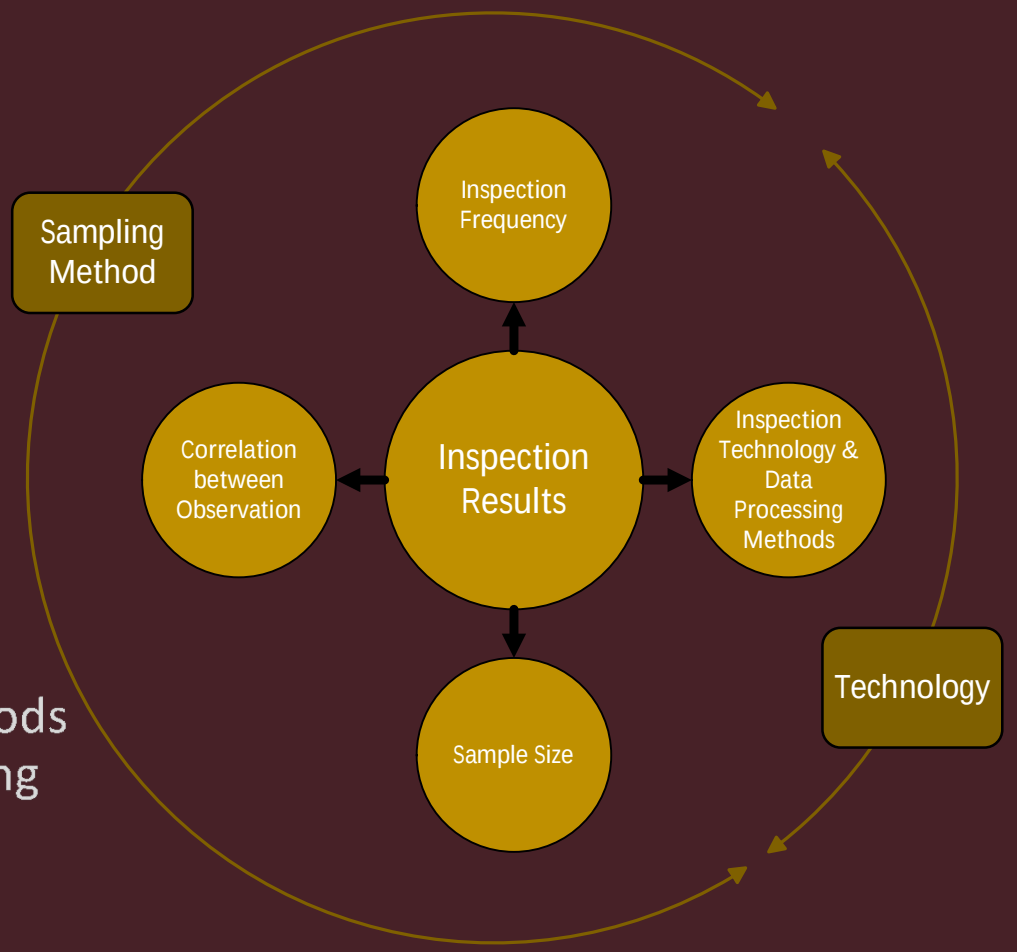
MPC MOUNTAIN-PLAINS CONSORTIUM

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)

State	Sampling Selection
California	Randomly selected sample
Florida	Random Number Generator
Indiana	Random Number Generator
New York	Selected by field personnel

- Utah

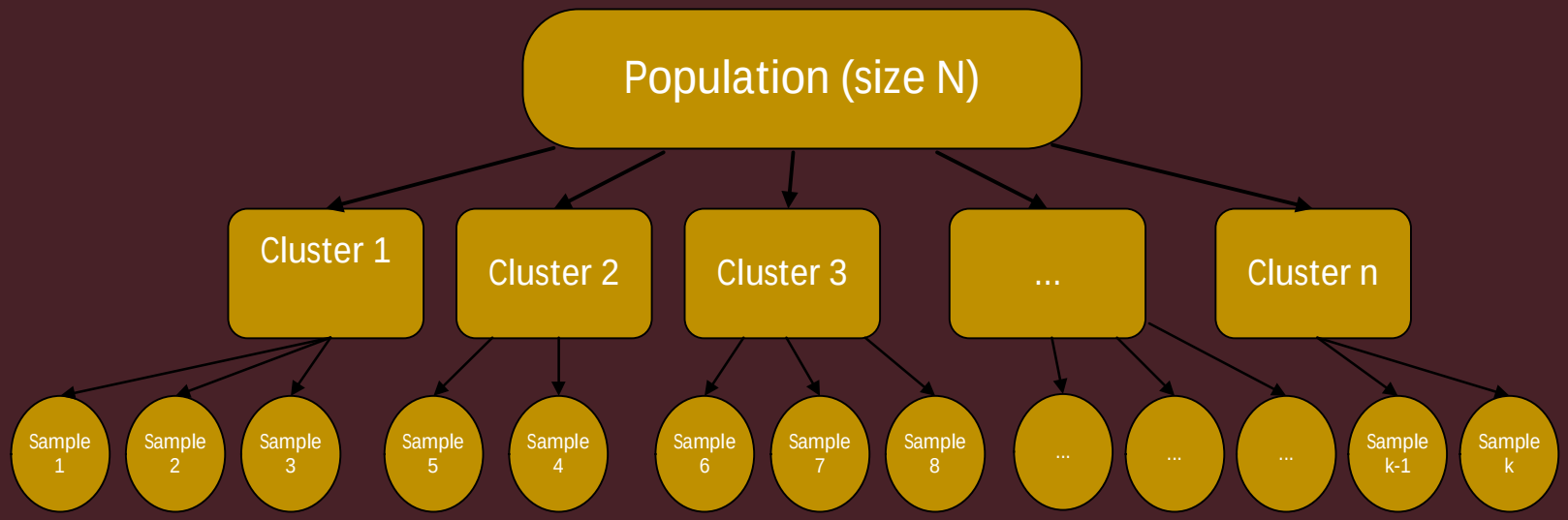
Research Objectives

- 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

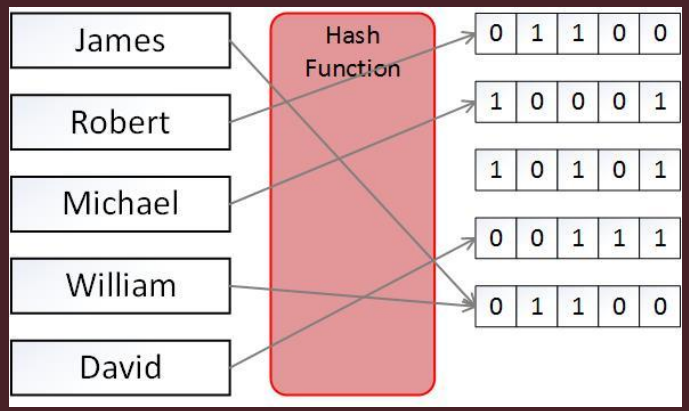
- Clustering sampling with high-dimensional data



- Curse of Dimensionality

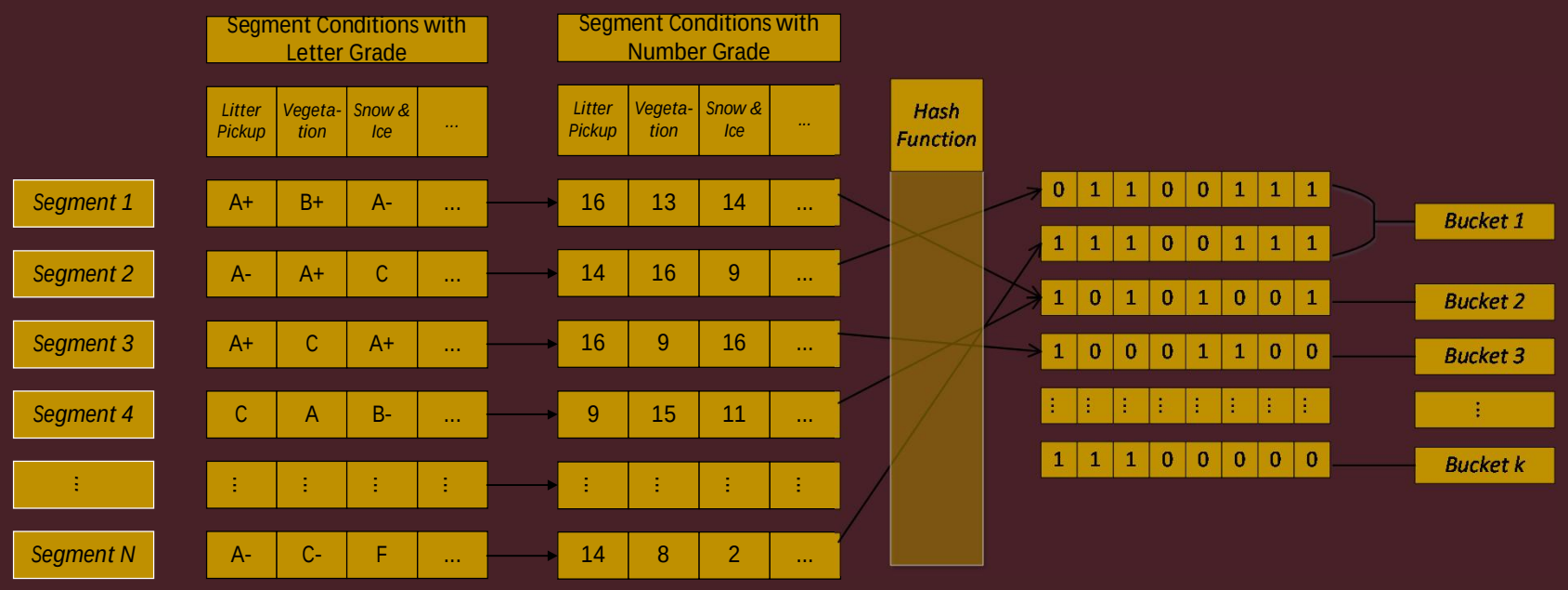
Locality-Sensitive Hashing (LSH)

- 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



- Estimate the probability by repeating applying different hash functions

Clustering Sampling

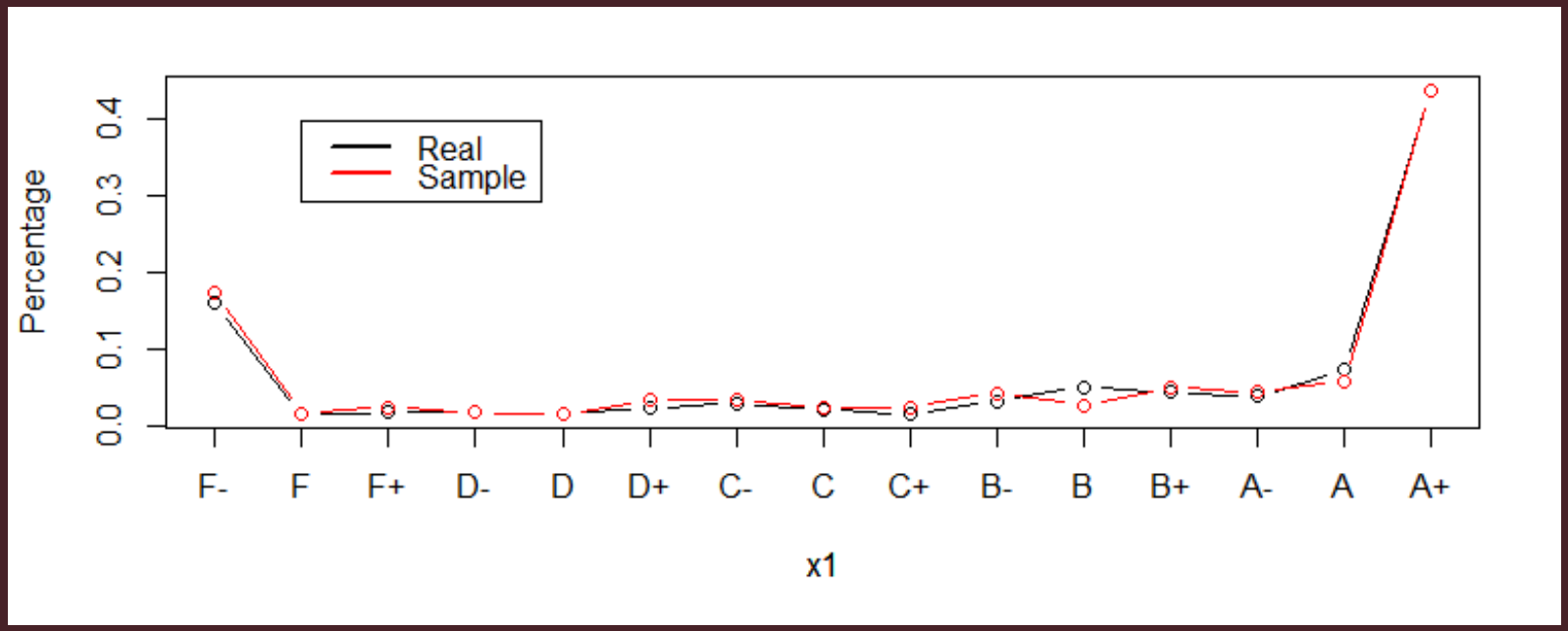
- Hash Functions:
  - Map the vectors in a M-dimension space
  - M-bit binary string based on the projected value in each dimension
- Similarity matrix:

	Seg_1	Seg_2	Seg_3	Seg_4	…
Seg_1	1	0.215	0.781	0.241	
Seg_2	0.215	1	0.11	0.762	
Seg_3	0.781	0.11	1	0.02	
Seg_4	0.24	0.762	0.02	1	
…					

- Clustering Methods: Spectral Clustering

Sampling Result Evaluation

- Average grade Vs. Grade Distribution



- RMSE (Root-Mean-Square Error): $RMSE_i = \sqrt{\frac{\Sigma_{j=1}^{n}\left(\overline{x_{ij}} - X_{ij}\right)^2}{n}}$

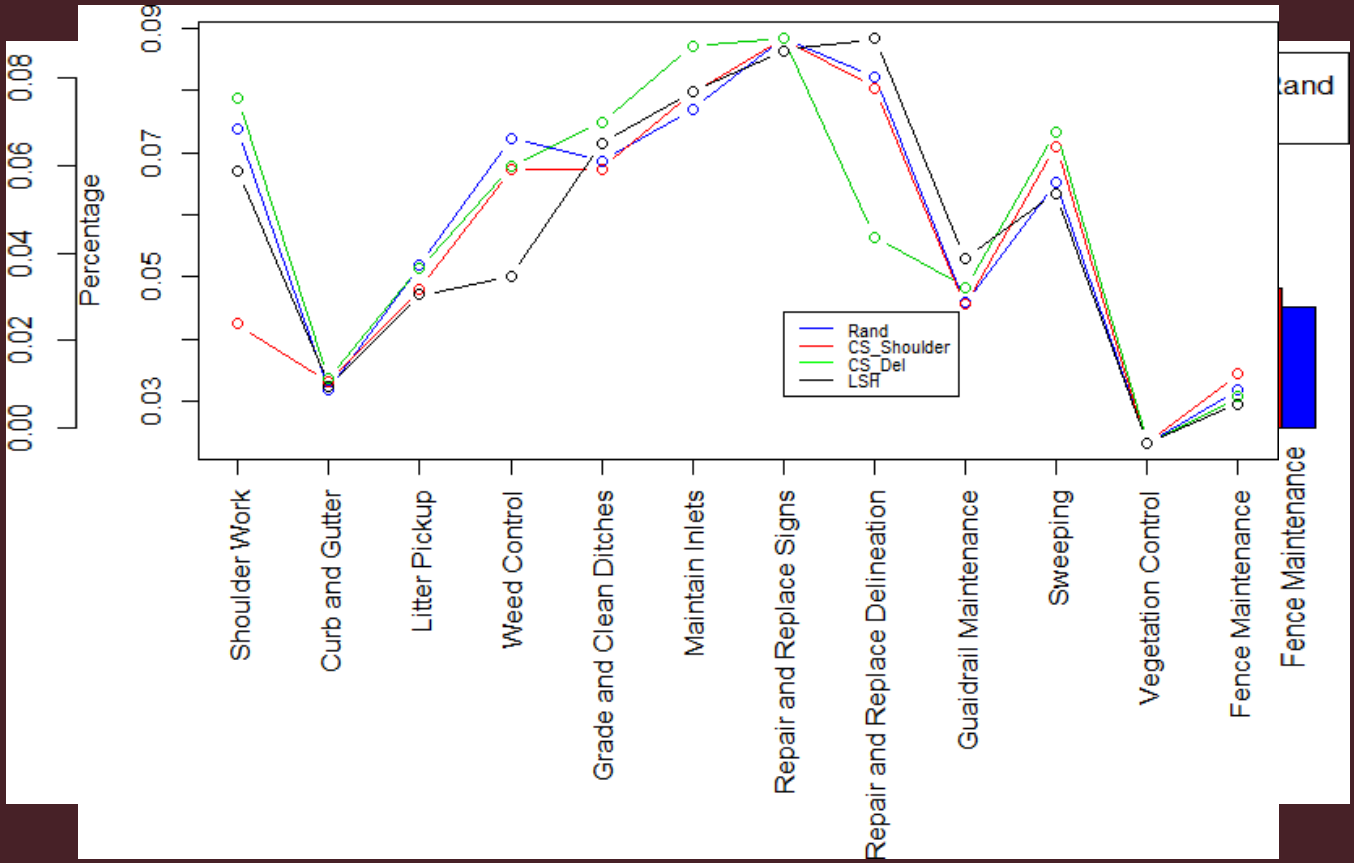
- Root-Square Error

Data Collection

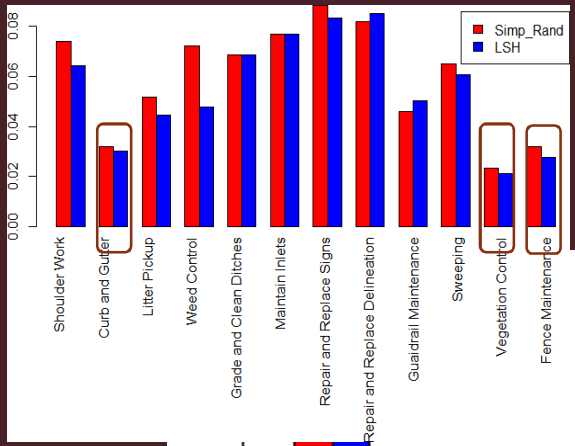
- Inspection records provided by Maintenance Management Quality Assurance (MMQA) program
  - Utah highway network
  - 17 measurement activities: snow and ice, litter pickup, vegetation, etc
  - Maintenance performance expressed as 15 different letter grade (A+ to F-)
- Sampling Process: 50 times

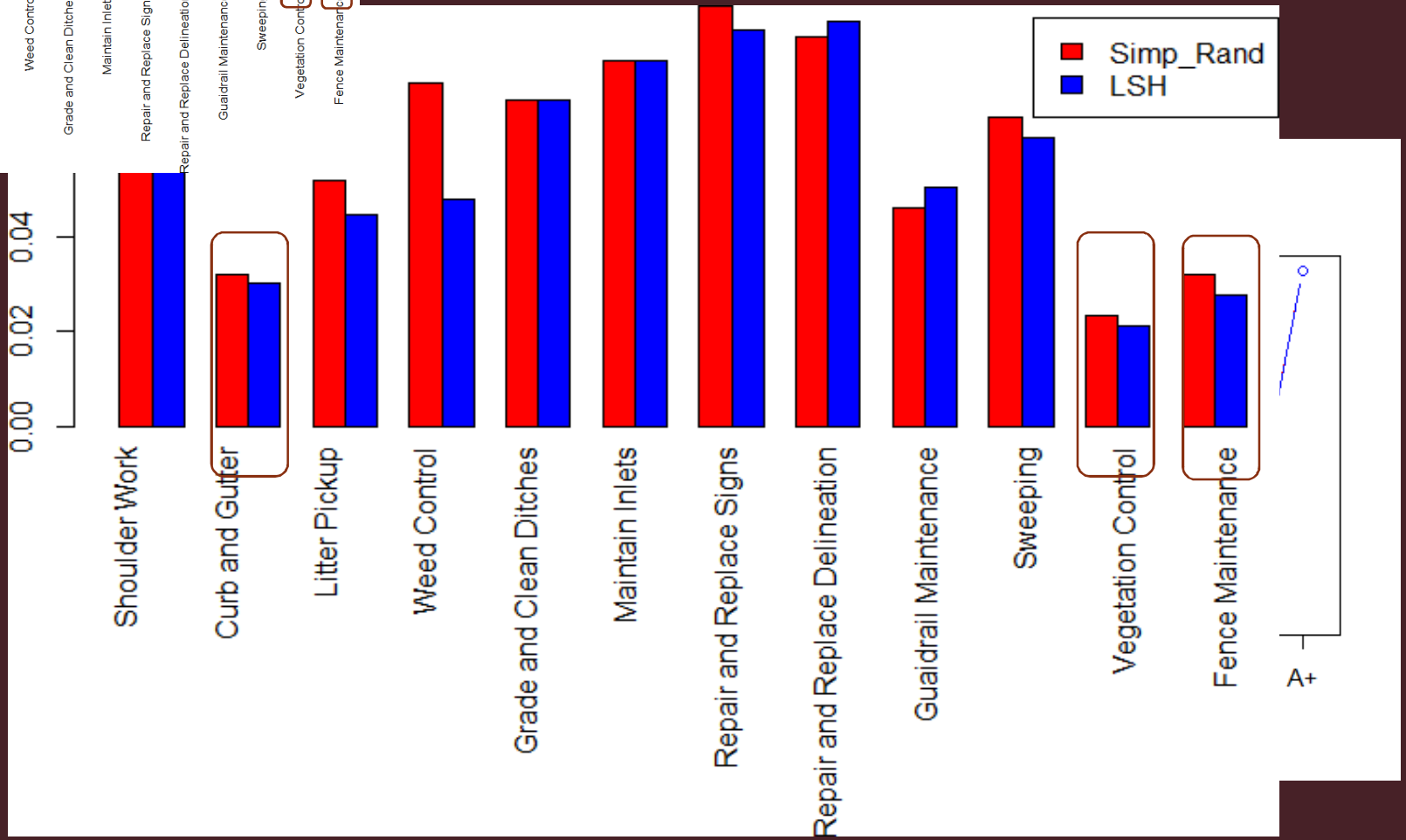
Result Analysis

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

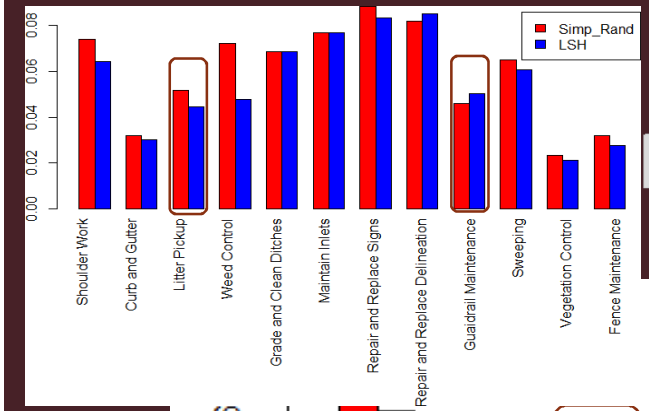


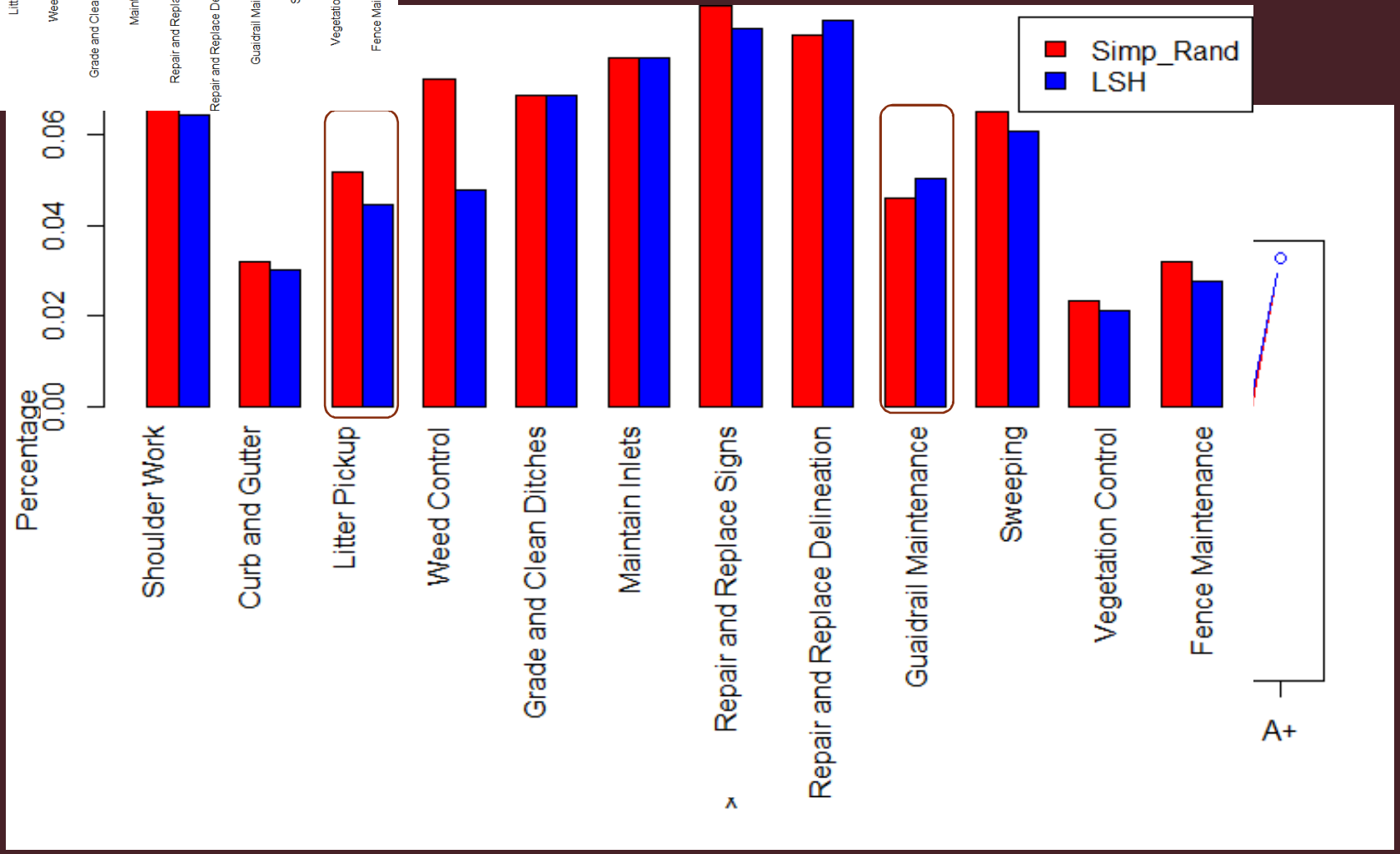
Result Analysis



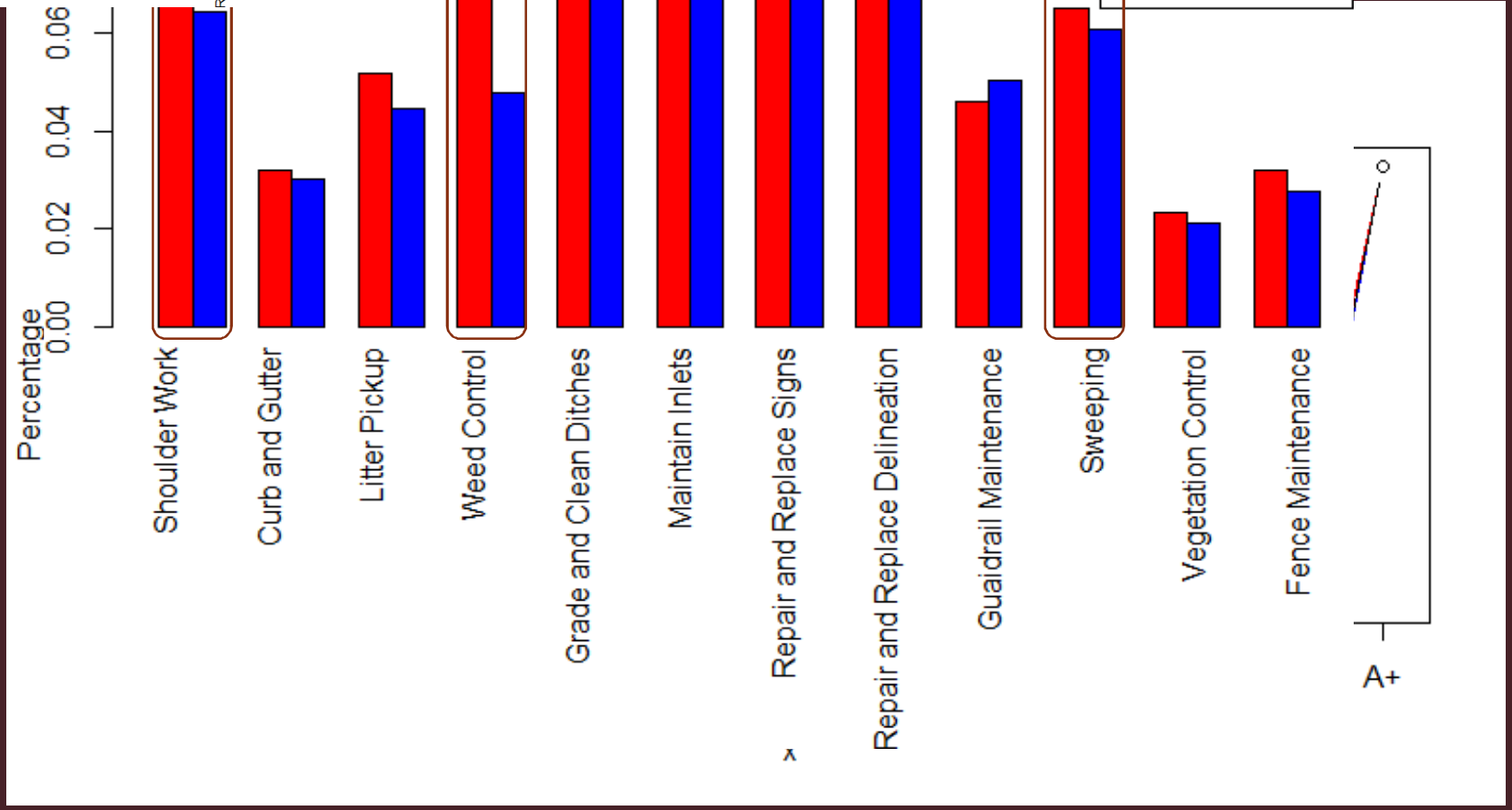


Result Analysis

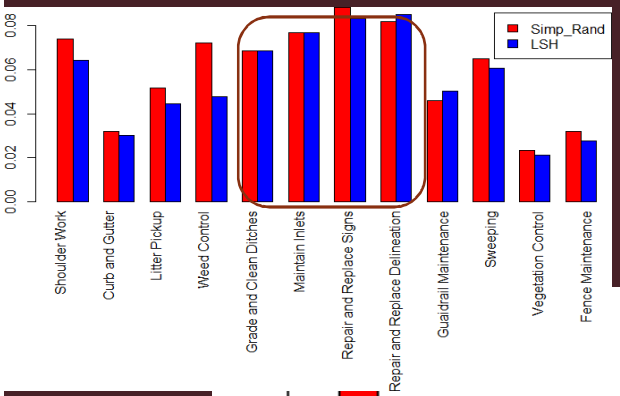


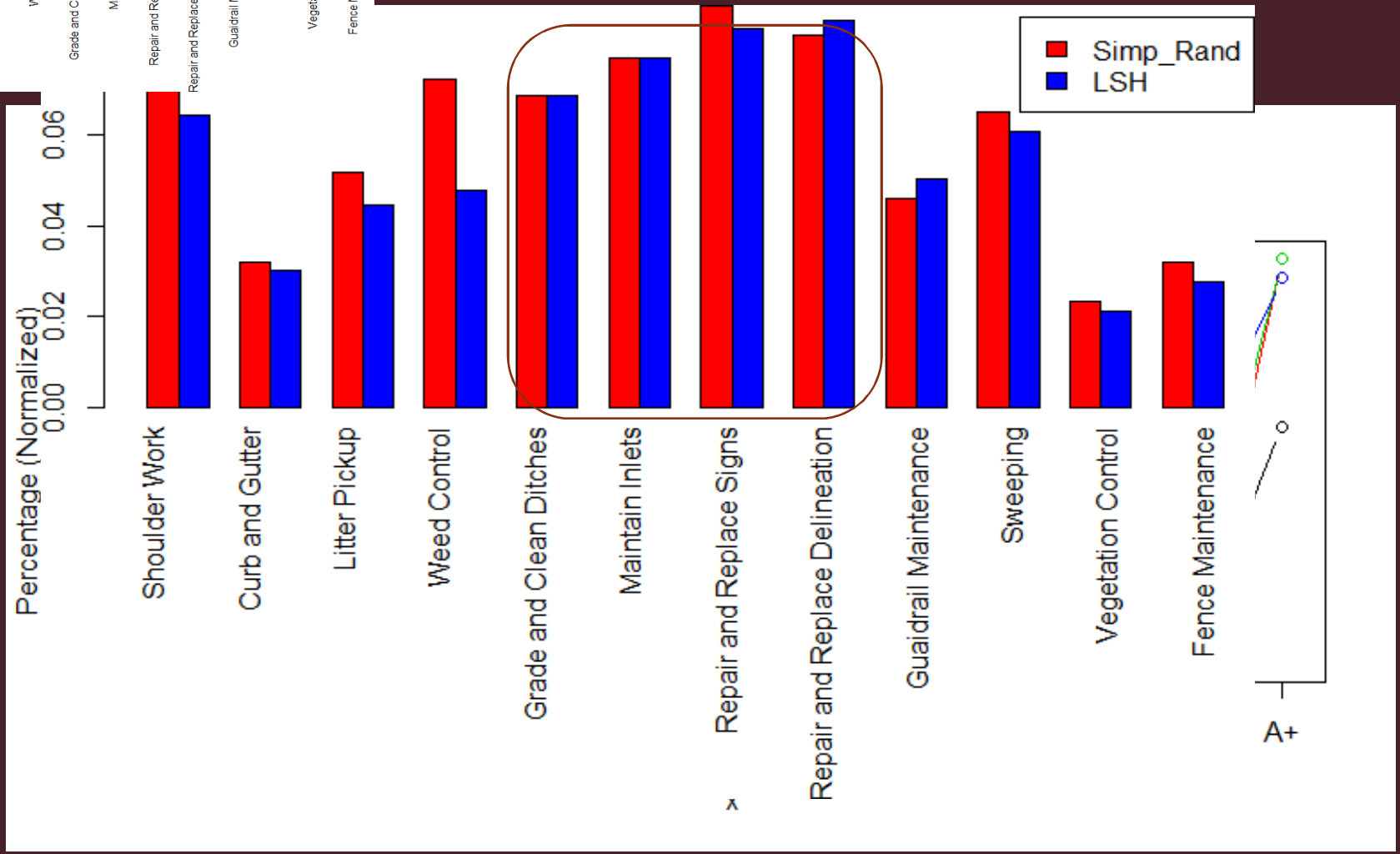


Result Analysis

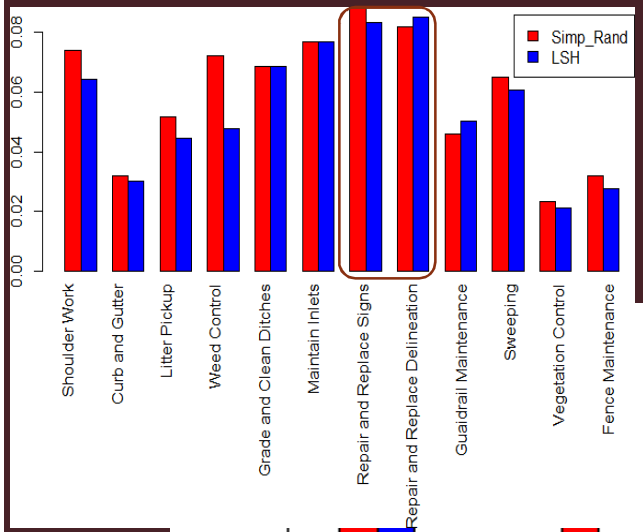


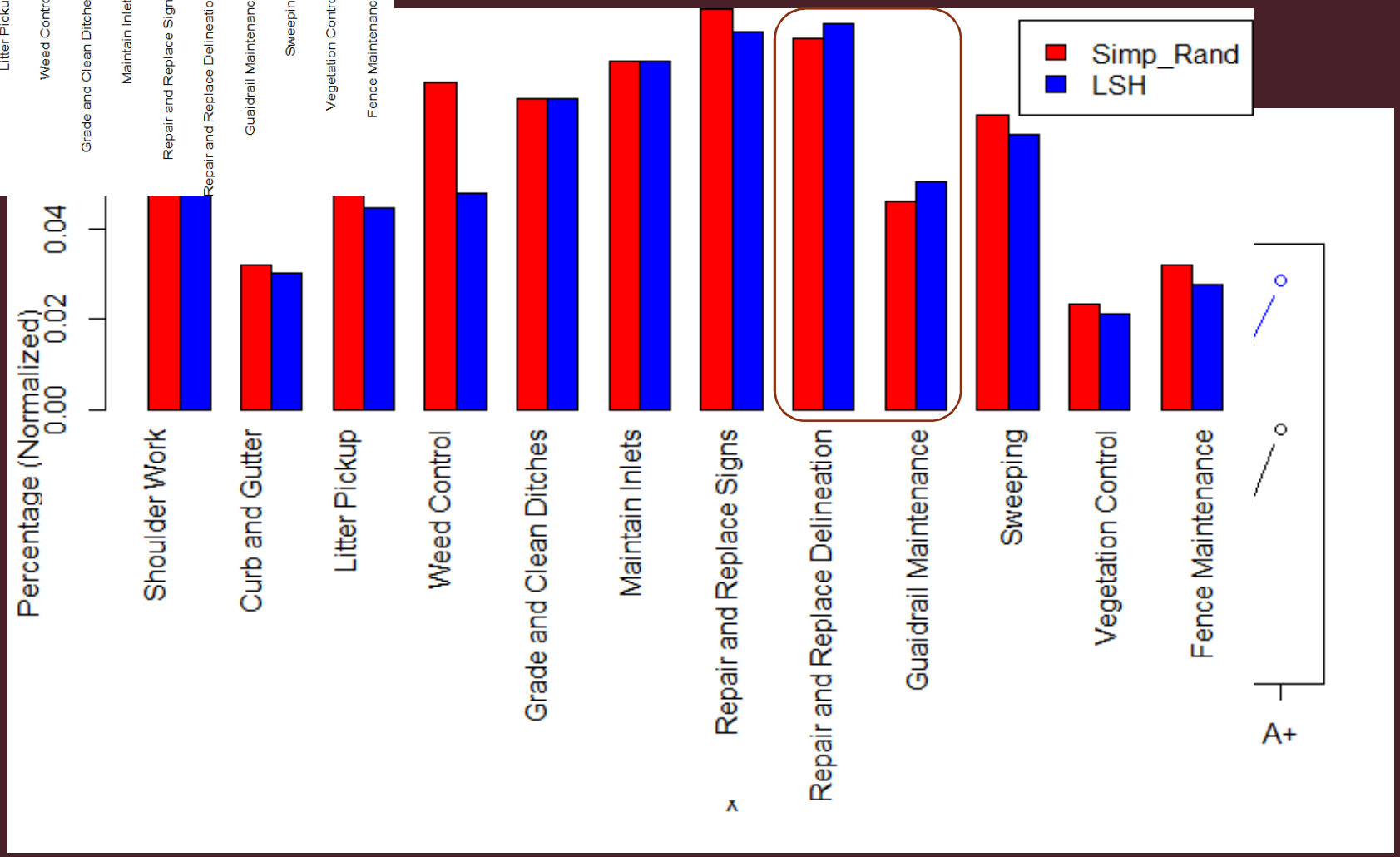
Result Analysis

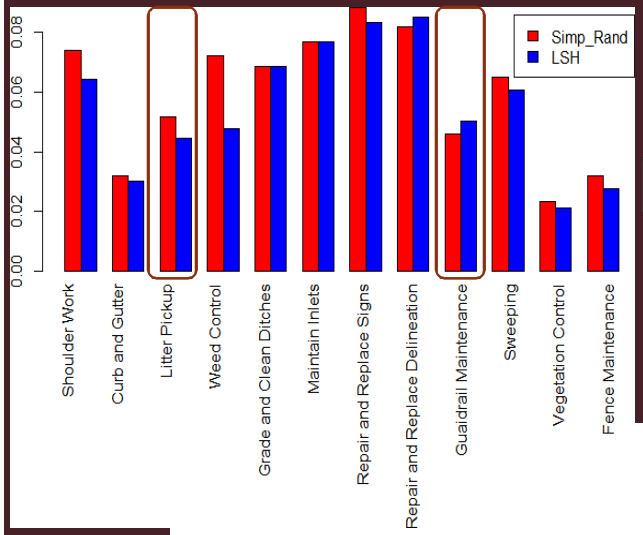




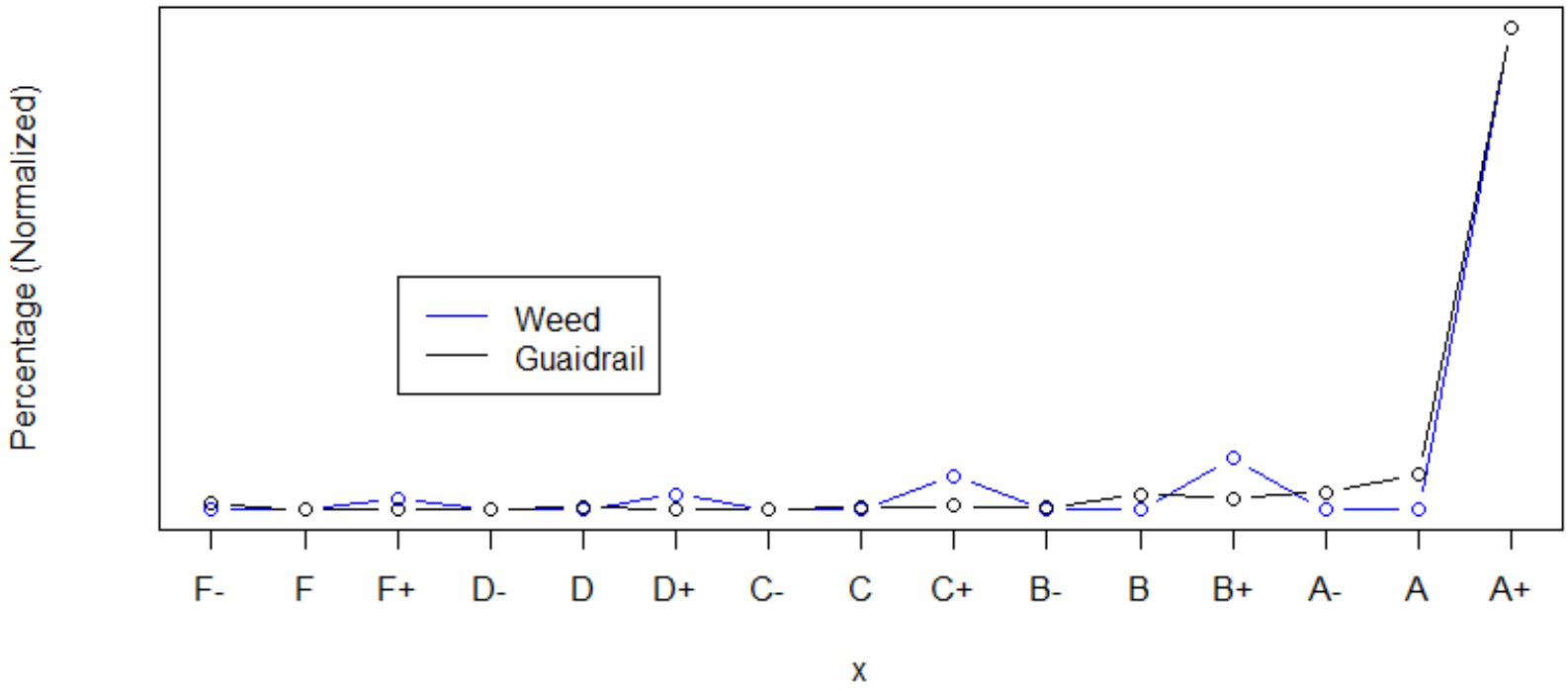
Result Analysis







Result Analysis



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!