Report No. UT-17.12

HOTSPOT AND SAMPLING ANALYSIS FOR EFFECTIVE MAINTENANCE AND PERFORMANCE MONITORING

Prepared For:

Utah Department of Transportation
Research Division

Submitted By:

University of Utah
Department of Civil & Environmental Engineering

Authored By:

Xiaoyue Cathy Liu
Zhuo Chen

Final Report
May 2017

UDOT RESEARCH
Keeping Utah Moving
Utah Department of Transportation - Research Division
4501 South 2700 West - P.O. Box 148410 - SLC,UT 84114-8410

DISCLAIMER

The authors alone are responsible for the preparation and accuracy of the information, data, analysis, discussions, recommendations, and conclusions presented herein. The contents do not necessarily reflect the views, opinions, endorsements, or policies of the Utah Department of Transportation or the U.S. Department of Transportation. The Utah Department of Transportation makes no representation or warranty of any kind, and assumes no liability therefore.

ACKNOWLEDGMENTS

The authors acknowledge the Utah Department of Transportation (UDOT) for funding this research, and the following individuals from UDOT on the Technical Advisory Committee for helping to guide the research:

- Jason Richins
- Kevin Griffin
- Lloyd Neeley
- Mark Marz
- Rukhsana Lindsey
- Sherrie Floyd
- Tim Ularich
- Tammy Misarasi
- Todd Richins

TECHNICAL REPORT ABSTRACT

1. Report No. UT- 17.12	2. Government Accession No. N/A	3. Recipient's Catalog No. N/A
4. Title and Subtitle HOTSPOT AND SAMPLING ANALYSIS FOR EFFECTIVE MAINTENANCE AND PERFORMANCE MONITORING		5. Report Date May 2017 6. Performing Organization Code
7. Author(s) Xiaoyue Cathy Liu and Zhuo Chen		8. Performing Organization Report No.
9. Performing Organization Name and Address University of Utah Department of Civil & Environmental Engineering 110 Central Campus Drive, Suite 2000 Salt Lake City, UT 84112		10. Work Unit No. 5H07622H 11. Contract or Grant No. 16-8506
12. Sponsoring Agency Name and Address Utah Department of Transportation 4501 South 2700 West P.O. Box 148410 Salt Lake City, UT 84114-8410		13. Type of Report & Period Covered Final May 2016 to May 2017 14. Sponsoring Agency Code PIC No. UT15.208
15. Supplementary Notes Prepared in cooperation with the Utah Department of Transportation and the U.S. Department of Transportation, Federal Highway Administration		

16. Abstract

In this project, we propose two sampling methods addressing "how much and where" the agencies need to collect infrastrauctures condition data for accurate Level-of-Maintenance (LOM) estimation in maintenance network with single type or multiple types of infrastructures. The method for single type infrastructure integrates Fisher information with spatial sampling technique that can be customized based on local agencies' requirements, such as station-balanced, spatially-balanced, or others. For infrastructure condition inspection in a network with multiple types of infrastructures, a high-dimensional clustering-based sampling method is proposed. The method is based on the fact that inspection activities are carried out on the roadway segment basis, and selects sample segments that contain multiple types of infrastructures for the accurate estimation of their respective LOMs. The sampling process consists of two components: current condition estimation and high-dimensional cluster analysis. The methods are implemented using the infrastructure inspection records in the State of Utah from September, 2014 to March 2016. The sampling results indicate that both methods outperform simple random sampling method which is widely used across agencies.

17. Key Words		18. Distribution Statement		23. Registrant's Seal
Infrastructure maintenance management, Fisher information, spatial sampling, high-dimensional clustering		Not restricted. Available through: UDOT Research Division 4501 South 2700 West P.O. Box 148410 Salt Lake City, UT 84114-8410 www.udot.utah.gov/go/research		N/A
19. Security Classification (of this report) Unclassified	20. Security Classification (of this page) Unclassified	21. No. of Pages 56	22. Price N/A	

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

UNIT CONVERSION FACTORS

SI* (MODERN METRIC) CONVERSION FACTORS				
APPROXIMATE CONVERSIONS TO SI UNITS				
Symbol	**When You Know**	**Multiply By**	**To Find**	**Symbol**
		LENGTH		
in	inches	25.4	millimeters	mm
ft	feet	0.305	meters	m
yd	yards	0.914	meters	m
mi	miles	1.61	kilometers	km
		AREA		
in^2	square inches	645.2	square millimeters	mm^2
ft^2	square feet	0.093	square meters	m^2
yd^2	square yard	0.836	square meters	m^2
ac	acres	0.405	hectares	ha
mi^2	square miles	2.59	square kilometers	km^2
		VOLUME		
fl oz	fluid ounces	29.57	milliliters	mL
gal	gallons	3.785	liters	L
ft^3	cubic feet	0.028	cubic meters	m^3
yd^3	cubic yards	0.765	cubic meters	m^3
		NOTE: volumes greater than 1000 L shall be shown in m^3		
		MASS		
oz	ounces	28.35	grams	g
lb	pounds	0.454	kilograms	kg
T	short tons (2000 lb)	0.907	megagrams (or "metric ton")	Mg (or "t")
		TEMPERATURE (exact degrees)		
°F	Fahrenheit	5 (F-32)/9 or (F-32)/1.8	Celsius	°C
		ILLUMINATION		
fc	foot-candles	10.76	lux	lx
fl	foot-Lamberts	3.426	candela/m^2	cd/m^2
		FORCE and PRESSURE or STRESS		
lbf	poundforce	4.45	newtons	N
lbf/in^2	poundforce per square inch	6.89	kilopascals	kPa
APPROXIMATE CONVERSIONS FROM SI UNITS				
Symbol	**When You Know**	**Multiply By**	**To Find**	**Symbol**
		LENGTH		
mm	millimeters	0.039	inches	in
m	meters	3.28	feet	ft
m	meters	1.09	yards	yd
km	kilometers	0.621	miles	mi
		AREA		
mm^2	square millimeters	0.0016	square inches	in^2
m^2	square meters	10.764	square feet	ft^2
m^2	square meters	1.195	square yards	yd^2
ha	hectares	2.47	acres	ac
km^2	square kilometers	0.386	square miles	mi^2
		VOLUME		
mL	milliliters	0.034	fluid ounces	fl oz
L	liters	0.264	gallons	gal
m^3	cubic meters	35.314	cubic feet	ft^3
m^3	cubic meters	1.307	cubic yards	yd^3
		MASS		
g	grams	0.035	ounces	oz
kg	kilograms	2.202	pounds	lb
Mg (or "t")	megagrams (or "metric ton")	1.103	short tons (2000 lb)	T
		TEMPERATURE (exact degrees)		
°C	Celsius	1.8C+32	Fahrenheit	°F
		ILLUMINATION		
lx	lux	0.0929	foot-candles	fc
cd/m^2	candela/m^2	0.2919	foot-Lamberts	fl
		FORCE and PRESSURE or STRESS		
N	newtons	0.225	poundforce	lbf
kPa	kilopascals	0.145	poundforce per square inch	lbf/in^2

*SI is the symbol for the International System of Units. (Adapted from FHWA report template, Revised March 2003)

LIST OF ACRONYMS

FHWA	Federal Highway Administration
HDCSS	High-Dimensional Clustering-based Stratified Sampling
LMDP	Latent Markov Decision Process
LOM	Level of Maintenance
LSH	Locality-Sensitive Hashing
M&R	Maintenance and Rehabilitation
MMQA	Maintenance Management Quality Assurance
RMSE	Root-Mean-Square Error
SRS	Simple Random Sampling
UDOT	Utah Department of Transportation

EXECUTIVE SUMMARY

Maintenance management has been relying heavily on collecting asset condition information to plan for maintenance activities and budget allocation. Data collection is often conducted on a sampling basis due to resource constraints. There is thus a perceived need for the development of effective sampling framework that can determine statistically representative samples, reflect the true Level of Maintenance (LOM) at state/region/station levels, and accommodate agencies' requirements. We advance existing knowledge by presenting systemic approaches for sampling schemes development to assist maintenance activity planning. The two proposed methods address "how much and where" the agencies need to collect infrastraucture condition data for accurate LOM estimation in maintenance network with single type or multiple types of infrastructures. The method for single type infrastructure integrates Fisher information with spatial sampling technique that can be customized based on local agencies' requirements, such as station-balanced, spatially-balanced, or others. The framework is showcased via an example application of the Signage Repair & Replace database maintained by the Utah Department of Transportation (UDOT). Four sampling methods that might be tempered to various needs are implemented. Sampling results are presented and compared against historical full asset inventory via similarity analysis. The framework lays a strong theoretical foundation for maintenance asset sampling and is effective for estimating LOM at state/region/station levels to assist with budget allocation.

For infrastructure condition inspection in a network with multiple types of infrastructures, a high-dimensional clustering-based stratified sampling method is proposed. The method complements existing literature by carrying out the inspection activities on the roadway segment basis (which is consistent with inspection process), and selecting sample segments that contain multiple types of infrastructures for the accurate estimation of their respective LOMs. The sampling process consists of two components: current condition estimation and high-dimensional cluster analysis. Current condition estimation predicts infrastructures' "current condition" by considering historical inspection records. High-dimensional cluster analysis represents the core piece of the sampling framework, which employs Locality-Sensitive Hashing algorithm and spectral sampling. Locality-Sensitive Hashing algorithm defines the similarity between segments, and spectral sampling assigns segments into clusters based on similarity matrix. The proposed method outperforms simple random sampling which is widely used in practice, especially under the circumstances where LOM varies greatly within infrastructures. The highlights of the proposed method include: 1. the selection of samples that can fulfill the sampling requirements for collecting multiple infrastructures' information simultaneously; 2. it reduces the sampling rate without compromise in accuracy compared with simple random sampling method.

The methods are implemented using the infrastructure inspection records in the State of Utah from September 2014 to March 2016. They are a potentially useful tool for agencies to effectively conduct infrastructure inspection and can be easily adopted for choosing samples from any infrastructure network.

1.0 INTRODUCTION

1.1 Problem Statement

A sustainable transportation system relies heavily on the preservation and maintenance of infrastructures to ensure and improve its functionality. The adoption of performance-based transportation management has been gaining more and more popularity as the key feature in the Moving Ahead for Progress in the 21st Century Act (MAP-21). It further motivates the need for a streamlined process to make transportation investment decisions on the basis of infrastructure performance. Accurate reporting of infrastructure condition is critical to maintenance planning as it helps identify "where and when" infrastructures need to be reconstructed or replaced and consequently drives budget allocation and project prioritization. The impact of maintenance activities on such has led to the design and implementation of numerous initiatives to improve maintenance quality and establishment of quality assurance programs nationwide (Yurek et al., 2012). Over the years, these programs have evolved to focus more on effectively reporting maintenance outcome and reaching the targeted level-of-maintenance (LOM).

The key question getting great attention is the ability to report maintenance conditions with satisfying accuracy and efficiency, which in large is determined by data availability. Of further interest, data reduction is an indispensable component of today's transportation management. Inspecting infrastructures can be very demanding in terms of resources, personnel, and time required. It is thus desirable to conduct inspection on a sampling basis rather than on the entire infrastructure inventory, yet with a loss of fidelity that is negligible in determining the true infrastructure conditions. Given that, developing proper sampling techniques to manage a region's infrastructures and accurately infer the LOM based on statistically representative dataset has become an intriguing topic over the past decades. Most state agencies, such as Florida Department of Transportation (DOT), Indiana DOT, Colorado DOT, just to name a few, use simple random sampling (SRS) method in their maintenance quality assurance programs to introduce randomness into the sampling process, yet lacking the consideration on infrastructures' spatial distribution and justification on the representativeness of the sampled data (Schmitt et al. 2006). The challenges lie in the inherently missing approach that is theoretically sound for accurately choosing infrastructure samples reflecting the LOM for decision making and budget allocation. Biased sampling can further cause well-intentioned policies to produce unintended consequences.

Besides the accuracy and efficiency, another concern about reporting maintenance conditions is the recurrent inspection costs, especially when there are more than one type of infrastructures on freeway segment. The SRS method is unbiased and able to generate samples that represent all types of infrastructures simultaneously. However, it always requires large sampling rate to justify the representativeness of samples. Another widely used sampling method is stratified random sampling, which divides the population into strata and selects sample from each stratum. It applies relatively small sample rates, but the selected sample is only confined to a single type of infrastructure. It is time consuming and operationally inefficient for field personnel if the samples of different infrastructures are widely distributed across all segments. The ideal sampling method is to select the group of highway segments for inspection, in which the sampled infrastructures are representative to reflect their respective LOMs within the entire network. Such sampling method, allowing to conduct once-for-all inspection instead of once-for-each-infrastructure-type, will significantly reduce the inspection costs.

1.2 Objectives

This research project focuses on developing systematic approaches to assess "where and to what extent" to collect infrastructure conditions with maximum information retained for LOM estimation. The challenges in developing such approaches vary as the types of infrastructures change. When there is a single type of infrastructure in the network, the accuracy and efficiency of sampling method are closely associated with the sample size and the spatial correlation between the sampled segments. In a network with multiple types of infrastructures, the success of a sampling method lies in whether it allows to conduct once-for-all sampling instead of once-for-each-infrastructure-type, which significantly reduces the inspection costs.

The first objective of this project is to develop an infrastructure sampling method for network with single type infrastructure. The sampling method determines sample size based on data-driven analytics rather than intuitively. Fisher information is calculated to estimate the minimum sampling rate sufficient to capture the infrastructure conditions throughout the network. To select spatially well-distributed sample, rules combining Generalized Random-Tessellation Stratified Design (GRTS) and hierarchical randomization are applied in the sampling framework.

Another objective of this project is to develop sampling method for networks with multiple types of infrastructures. The sampling method selects samples that can accurately reflect LOMs of all infrastructures throughout the network. It allows transportation agencies (e.g. DOTs) to customize the parameters such as sample size, inspection frequency, and infrastructures of interest. The proposed method integrates infrastructure deterioration prediction, high-dimensional cluster analysis, and Locality-Sensitive Hashing (LSH). The sampling method can incorporate different features, such as infrastructure condition, geographic information, traffic condition, and geometric design, as the information based on which sample is selected. The method is adaptable to any infrastructure changes, since sampling process is constantly updated with previous inspection results and maintenance records.

1.3 Scope

This research project breaks down into two parts: sampling method for single type infrastructure, and sampling method for multiple types of infrastructures.

Tasks in developing sampling method for single type infrastructure include:

- Construct Fisher Information with infrastructure conditions to estimate the minimum sampling rate which is sufficient to capture the infrastructure LOM in the network; and
- Develop an algorithm combing GRTS and hierarchical randomization to select spatially-balanced sample.

Tasks in developing sampling method for multiple types of infrastructures include:

- Develop an algorithm to estimate the infrastructure deterioration process with historical infrastructure maintenance records;
- Use LSH to construct the similarity matrix between roadway segments (sample population);

- Implement spectral clustering analysis to construct strata for stratified sampling; and
- Develop performance measurement index for non-parametric sampling results.

To conduct all the above mentioned tasks, it requires the support of mass amount of historical data from multiple sources and jurisdictions. The infrastructure inspection records are provided by the Utah Maintenance Management Quality Assurance (MMQA). The MMQA program was established by the UDOT in 1997 for evaluating and reporting the effectiveness of its maintenance activities. The program has evolved ever since then to provide systematic guidance on feature condition thresholds, funding projection and allocation, and LOM measurements. MMQA offers guidelines on a total of 17 measurement activities such as snow and ice, litter pickup, vegetation control, etc. It further refines specifications on the criteria of desired/deficient conditions of each activity. Inspectors are required to be familiar with the procedure and methodologies described for each maintenance activity before going into the field. The graphical description in MMQA helps them confidently describe the condition of any particular feature. Maintenance performance is measured and reported in the form of LOM, expressed as 15 different letter grades (A+ to F-). The entire statewide highway system is divided by 76 maintenance stations. Each station further divides each of its routes into one or more segments (2,048 segments in total). The personnel conduct inspections for each route segment, and record both the total number of features to be maintained on that segment and the total number of deficient features.

1.4 Outline of Report

The rest of the report is structured as follows. Chapter 2 summarized literature on state-of-the-art in infrastructure sampling, Fisher information, spatial-balanced sampling method, and high-dimensional clustering. Sampling methods developed in this project for both single type and multiple types of infrastructures are presented in Chapter 3. Chapter 4 describes the data sources used for testing the sampling methods and Chapter 5 presents the sampling results. We conclude the study and give recommendations for future research in Chapter 6.

2.0 LITERATURE REVIEW

2.1 Overview

Sampling methods in infrastructure maintenance management have been intensively studied in recent years. This chapter presents a literature summary on four major components in this project: state-of-the-art infrastructure maintenance sampling, Fisher information, spatially-balanced sampling, and high-dimensional clustering.

2.2 State-of-the-Art in Infrastructure Sampling

Many state DOTs have developed MQA program guidelines, most of which adopt certain forms of simple random sampling technique for asset data collection (Schmitt et al., 2006). Simple random sampling chooses segments randomly by applying a fixed sampling rate. The probability of each segment being chosen is the same. With simple random sampling, network segmentation directly affects sampling efficiency. For a given network, the sample population is determined by the length of sample segment as maintenance activities are carried out segment-by-segment. Long segment leads to small sample population and consequently increases sampling rate in order to meet the requirement of minimum sample size. On the contrary, short segment leads to large sample population. It also increases the probability of scattered segments to be sampled, which correspondingly increases labor hours of collecting the data since the maintenance personnel might need to drive through more unsampled segments between sampled segments. The selection of segment length is an empirical process, a decision made by maintenance operators. For example, California Department of Transportation (Caltrans) and New York Department of Transportation (NYDOT) use 1 mile as sampling unit; North Carolina uses 0.2 mile, while Florida, Indiana, Texas, Virginia, Washington, and Wisconsin use 0.1 mile (Schmitt et al., 2006). Once the sampling segment unit is determined, question is directed to the selection of sample size.

Generally, three methods have been widely used in previous studies to determine sample size: fixed percentage of population (Templeton and Lytton, 1984), statistical method (De la Garza et al., 2008; McCullouch and Sinha, 2003; Medina et al., 2009; Schmitt et al., 2006;

Selezneva et al., 2004), and optimization method (Gharaibeh et al., 2010; Mishalani and Gong, 2009a, 2009b). Among the three, fixed percentage of population is easy to implement, yet accuracy is compromised due to its empirical nature and a lack of scientific validation. The proportion of samples needed from the entire population varies by the type of Maintenance and Rehabilitation (M&R) activities and needs. Templeton and Lytton (1984) believed that a sample size of 30% to 35% is needed to predict the cost to repair segments below certain condition threshold. Among the surveyed transportation agencies that have MQA programs, maintenance sampling range varied from 1.5% to 20% (Yurek et al., 2012). Statistical methods are based on approximated sampling distributions, appearing to be more statistically valid comparing with fixed percentage of population. Schmitt et al. (2006) summarized a series of applications of standard statistical methods in maintenance sampling, such as using confidence interval of normal population, number of observations for t test of mean, etc. Selezneva et al. (2004) tested sample sizes on different randomly picked reliability levels until the corresponding sample meet the requirement of quality assurance criteria. One novel method in determining sample size is by optimizing the maintenance plan. Most studies utilizing optimization techniques apply Latent Markov Decision Process (LMDP). LMDP is a classic approach to solve long-term network-level M&R policy optimization problem. The purpose of LMDP is to maximize performance (e.g. higher LOM) with a given budget or to minimize costs with required performance. For example, in pavement maintenance, LMDP can determine how to assign routine maintenance, resurfacing, and inspection activities to the network (Mishalani and Gong 2009b). Mishalani and Gong (2009a) considered sample size as a decision variable in the LMDP optimization framework. Research on LMDP in terms of sampling rate is very limited by far. Compared with the two aforementioned methods, LMDP is used specifically for maintenance activities and flexible for different types of assets. Gharaibeh et al. (2010) optimized sample size by minimizing the costs of performing sampling and the equivalent cost of inconvenience caused by poor-quality materials and construction. Due to the complexity of optimization, simplified assumptions are often made for the probability function in optimization methods, compromising the model fidelity. To implement optimization methods, it also requires good historical database to ensure accurate construction of transition matrix.

Once sample size is determined, a sampling plan needs to be designed to obtain features of interest. Several sampling design schemes have been widely used, including simple random sampling, sampling with replacement, sampling without replacement, stratified random sampling, etc. For MQA, simple random sampling and stratified random sampling are the most popular methods (Schmitt et al. 2006). However, as pointed out by previous researchers, the accuracy of true population condition estimate not only depends on the quality of measurements and sample size, but also depends on the correlation among asset conditions at different locations (Mishalani et al., 2011). Such spatial correlation exists in maintenance sampling. As Mishalani and Goel (2011) mentioned, smaller spatial correlation leads to more accurate estimation of asset conditions. Simple random sampling and stratified random sampling do not take this into account.

2.3 Fisher Information

Fisher information is a measure of information that is expected within a trial X about the parameter θ. It can be defined as the derivative of the log-likelihood function with respect to θ (Ly et al., 2014):

$$\mathrm{I}(\theta) = \mathrm{Var}(\frac{d}{d\theta}\log f(X|\theta)) = -\mathrm{E}(\frac{d^2}{d\theta^2} log f(X|\theta)) \tag{1}$$

Fisher information has been applied within a variety of statistical paradigms to answer different substantive questions. Liu and Yu (2009) utilized Fisher information to determine optimal geolocation data compression ratio for transportation target identification. Towsley et al. (2006) applied Fisher information metric to determine flow size distribution from packet sampling for network monitoring. At its minimum, numerous other studies have used it to either define a prior default parameter, determine sample size, or measure model complexity (Lee and Wagenmakers, 2013; Myung, 2003; Rissanen, 1996; Stevens, 1957). It plays a pivot role in statistical modeling.

Within the context of the maintenance asset sampling scheme, Fisher information is a measurement of the maximum likelihood that the sampled maintenance inspection outcome represents the true asset condition. It can be utilized to determine the appropriate sample size. Taking signage inventory as an example. The inspection can be treated as a Bernoulli process, where the sign's condition is either desired ("1") or deficient ("0"), or vice versa. The probability density function (pdf) of a Bernoulli model can then be expressed as:

$$f(x|\theta) = P(X = x) = \theta^{x}(1-\theta)^{1-x}, \text{ where x=0 or x=1} \tag{2}$$

where θ is the probability that the sign's condition takes on the value of 1.

The Fisher information of sample from Bernoulli model can be thus calculated by plugging Equation (2) into Equation (1), which yields:

$$I(\theta) = -\sum_{x=0}^{1} \frac{d^2}{d\theta^2} logP(X = x)P(X = x) = \frac{1}{\theta(1-\theta)} \tag{3}$$

As shown in Figure 1, Fisher information demonstrates the sensitivity of a Bernolli model with respect to parameter θ. As Fisher information increases, the sample becomes more accurate in describing the real condition of the population. When θ reaches 0 or 1, the expected Fisher information goes to infinity. Namely, when the conditions of signs are all "desired" or "defect", any sample can perfectly reflect the real condition of all segments.

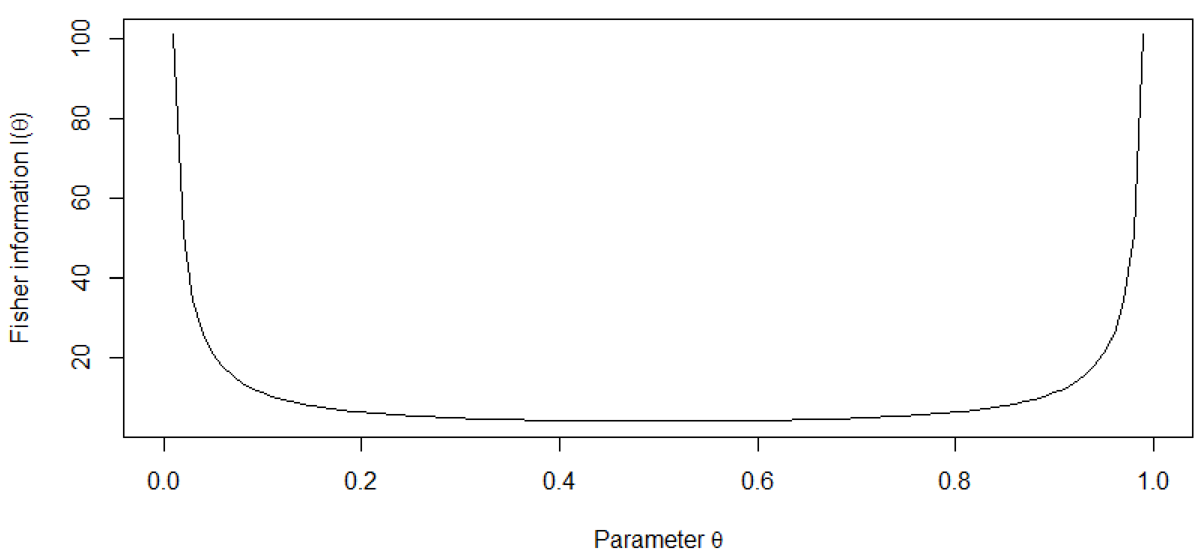


Figure 1 Fisher information as a function of θ in Bernoulli model.

2.4 Spatial Sampling

Putting into the context of spatial sampling, samples are collected typically in 1-, 2- or 3-dimensional space. Generic situations arise when the resource population is represented as collections of points, lines, or areas over spatial extents. Spatial sampling can be conducted using the traditional sampling methods mentioned in the foregoing cited studies, such as simple random sampling, systematic sampling, stratified random sampling, just to name a few. It can also take into account the unique spatial features resided within population, such as spatial autocorrelation and spatial heterogeneity (Goodchild et al., 1992; Ripley, 2005; Wang et al., 2012). Previous studies have applied a variety of spatial sampling techniques that appear to perform reasonably well in different sampling applications for getting a spatially-balanced sample. Yet still, numerous difficulties present themselves in multiple occasions. For example, when applying stratified sampling on one- or two-dimensional populations, it is difficult to split the entire population into spatially contiguous strata especially when variable probability or substantial variations in spatial density exist (Stevens and Olsen, 2004). Spatial stratification has a wide range of applications due to the fact that heterogeneity can be reduced within stratum and the easiness for collecting samples that are highly representative (Wang et al., 2010). Among them, the Generalized Random-Tessellation Stratified Design (GRTS) provides a flexible means for selecting spatially well-distributed samples (Stevens & Olsen, 2000). By combining the GRTS with hierarchical randomization, it maps data from 2-dimentionsal space into a 1-dimensional linear structure for sampling, which can eventually result in a spatially well-balanced random sample. The method is well-suited to be employed in the sampling procedure for maintenance activities, especially given the needs of transportation agencies in terms of maximized spatial coverage when collecting data.

2.5 High-Dimensional Clustering

In the sampling method for multiple types of infrastructures, stratification is implemented via high-dimensional cluster analysis. Since each highway segment often contains multiple infrastructures, we consider a segment as a high-dimensional vector and each type of infrastructure as one dimension of that vector. By applying high-dimensional cluster analysis, we divide all segments into several clusters based on their infrastructures' conditions. The challenge in dealing with high-dimensional data lies in "Curse of Dimensionality". The concept is originally defined by (Bellman, 1961), referring to the difficulty of optimizing a multi-variable function within the multi-dimensional context. In cluster analysis, as dimensionality increases, the number of data points within each dimension becomes increasingly "sparse" (Steinbach et al., 2004). As illustrated in Figure 2, a dataset with 10 points is randomly distributed from 0 to 1 in one-dimensional space. The points are in close vicinity of each other. There are 4 points within the range [0, 0.5]. But when the dataset is expanded to two-dimension, if we still use 0.5 as the discretization unit in each dimension, there are then only 3 points in the range of [0, 0.5] in each dimension. When we further expand the dataset to three-dimension, there are only 2 points within the same unit. So for high-dimensional data, distance may no longer be effective to distinguish points and most cluster techniques applicable to low-dimension data (e.g. centroid-based clustering, density-based clustering) render meaningless.

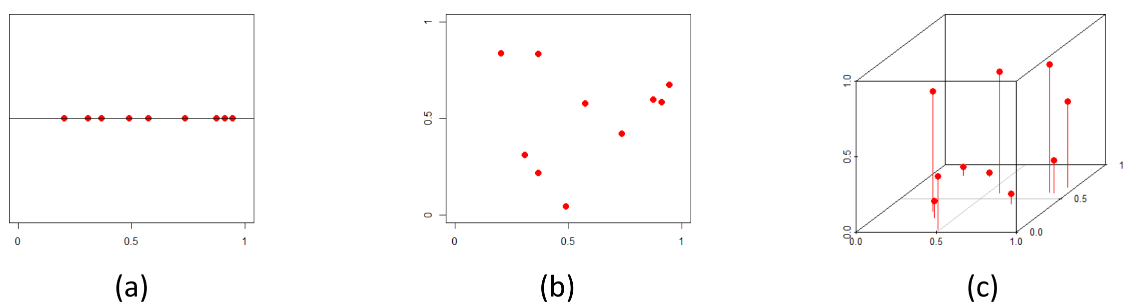


Figure 2 Illustration of sparsely distributed data points due to Curse of Dimensionality.

During the past decades, much effort has been devoted to avoid the *Curse of Dimensionality*. One approach to high-dimensional clustering is to develop new measurements for distance or similarity across clusters, including grid (Hinneburg and Keim, 1999), sum of similarities along dimensions (Aggarwal, 2001), and approximate similarity (Li et al., 2002). (Li et al., 2002) also suggested a practical similarity measurement called Locality-Sensitive Hashing (LSH). LSH is a widely-used algorithm to search similarity between high-dimensional data for fast indexing and database searching. LSH maps high-dimensional data points to a low-dimensional space by applying hash functions. As mentioned in (Datar et al., 2004), a hash function family $H = \{h_1, h_2, h_3, \dots, h_i, \dots\}$ is called (d_1, d_2, p_1, p_2)-sensitive for any two high-dimensional vectors q and v:

- if $\mathrm{D}(q, v) \leq \mathrm{d}_1$, then $P_H[\mathrm{h_i}(q) = \mathrm{h_i}(v)] \geq \mathrm{p}_1$

- if $\mathrm{D}(q, v) > \mathrm{d}_2$, then $P_H[h_i(q) = h_i(v)] \leq p_2$

where d_1 and d_2 are the critical distances to determine if q and v are similar, p_1 and p_2 are the critical probabilities, and D is the distance measurement in the low-dimensional space. If the distance between the mapped values is less than d_1, then the probability that q and v are similar is greater than p_1. On the contrary, if the distance between mapped values is greater than d_2, then the probability of q and v are similar is less than p_2. Based on such definition, researchers proposed different function schemes and validated their reliability in capturing the underlying similarity, including inner product (Charikar, 2002), learned Mahalanobis distance (Jain et al., 2008), and normalized kernel function (Kulis and Grauman, 2012).

2.6 Summary

This chapter summarized the key finds from the literature research for this study. Three major issues of focus with regard to infrastructure sampling in this project include sample size, uneven spatially distributed sample, and sample with high-dimensional information. To address these issues, we apply Fish information, spatial-balanced sampling, and high-dimensional clustering in the proposed sampling methods. Previous studies applied these methods/algorithms mainly in statistics, electrical engineering, and computer science. In the following chapters, we demonstrate how these methods can be used to determine the infrastructure sampling for optimal maintenance management.

3.0 RESEARCH METHODS

3.1 Overview

This chapter describes two sampling methods for single type infrastructures and multiple types of infrastructures, respectively. The sampling method for single type infrastructure uses Fisher information to estimate the proper sample size and GRTS combined with hierarchical randomization to select proper inspection segments. The sampling method for multiple types of infrastructures consists of infrastructure deterioration estimation, high-dimensional clustering, and stratified sampling.

3.2 Sampling Method for Single Type Infrastructure

The proposed method integrates Fisher information with spatial sampling techniques and is able to accommodate local agency's needs (e.g. sample on various functional classes, stations, maximum spatial coverage, etc.). It is also flexible for potential integration with spatial optimization to set certain resource constraints. The proposed framework lays a strong theoretical foundation for maintenance asset sampling and is effective in estimating LOM at state/region/station levels for budget allocation.

3.2.1 Fisher Information

In reality, we only observe a single outcome $\vec{\mathrm{x}}$ (the maintenance inspection at certain time) of size n (sample size) and have to infer θ instead. The goal for maintenance management is to provide a reasonable guess of the true value θ^*, such that the true LOM is unveiled. Fisher information can be used to determine the asymptotically least number of trials n that needs to be collected such that an estimator $\vec{\mathrm{X}}$ yields estimates $\vec{\mathrm{x}}$ at a certain level of accuracy. For a complete derivation on measuring the performance of an estimator of a Bernoulli Model, interested readers are welcome to refer to (Ly et al., 2014). The general consensus is that as the number of trails n increases, more information is extracted about θ. With i.i.d. (independent and identically distributed) assumption for $\vec{\mathrm{x}}$, variance of the estimator $\vec{\mathrm{X}}$ is given by: $\mathrm{Var}(\mathrm{X}) = \theta(1-\theta)$. Thus the goal is to tame the variance such that the largest variation is minimized, which occurs at $\theta = 0.5$ (as shown in Figure 1).

To determine the sample size n for the signage inspection $\vec{X}$ with $X \sim \mathrm{Ber}(\theta)$, the problem can be formulated as the chance of obtaining an estimate that is more than α distance away from the true value is no larger than β, which can be expressed as:

$$P(\bar{X} \in [0.5 \pm \alpha]) \gg 1 - \beta \tag{4}$$

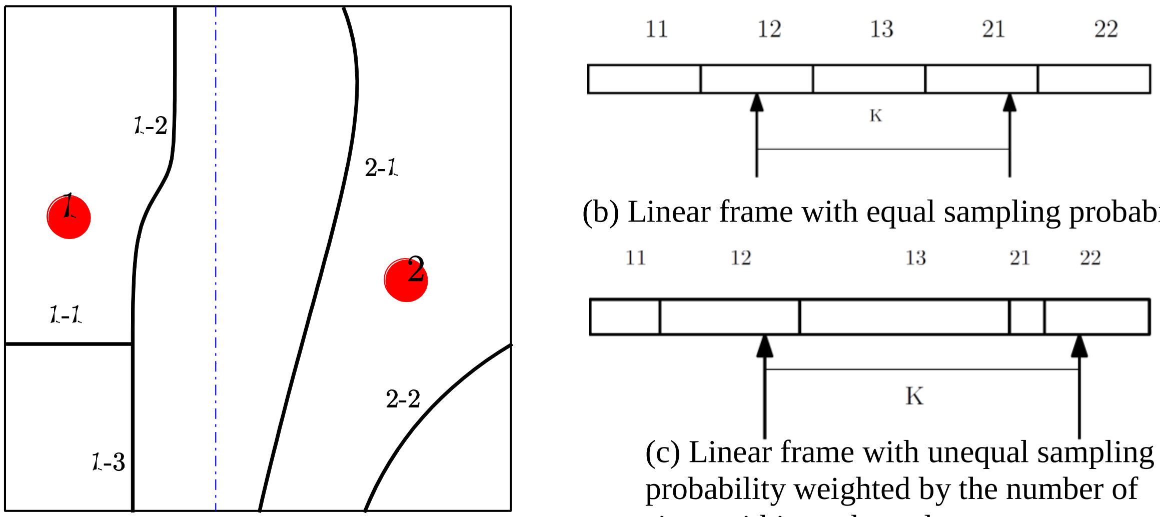
3.2.2 GRTS for Maintenance Management

For maintenance management purposes, sampling process should be able to accommodate varying spatial sample intensity, and spread the sample points evenly and regularly over the domain. Most agencies are using simple random sampling for maintenance management, yet it tends to exhibit uneven spatial patterns. Signage population, as an example, exists in spatial matrix. Although systematic sampling might compensate on the spatial feature, it has limited flexibility in changing sample point density or accommodate various inclusion probability. GRTS design combines simple random and systematic characteristics, and guarantees all possible samples are distributed across the resource. The basic idea of GRTS method is to create a quadrant-recursive function that maps two-dimensional space into a one-dimensional one, thereby defining an ordered spatial address for the population. Unequal probability sampling can be achieved by giving each point a length proportional to its inclusion probability.

In maintenance application, the target population is the roadway segments partitioned by stations. To illustrate the GRTS sampling scheme, Figure 3 shows an example region where five roadway segments are under the maintenance jurisdictions of two stations (circled). The segments are randomly labeled according to the station ID. Different from the classic quadrant partitioning of a region, the maintenance application already labeled (or spatially partitioned) each segment by stations. Therefore, when mapping the 2-dimensional space into an ordered 1-dimensional linear structure, the features in Figure 3 (a) would be transformed into Figure 3 (b). Note that in Figure 3 (b), each segment is assigned equal probability, yet unequal probability can be tempered by the allowance of unequal length for each unit as shown in Figure 3 (c). The sampling scheme can be expressed as:

$$d + (i - 1) \times k \qquad for\ i = 1,2, \ldots., n \tag{5}$$

where d is a random start within the 1-dimensional space along $[0,k)$, L is the total length of line, n is the sample size, and $k=L/n$.



(a) Example region for maintenance activities

(b) Linear frame with equal sampling probability

(c) Linear frame with unequal sampling probability weighted by the number of signs within each roadway segment

Figure 3 Example of GRTS algorithm for maintenance activity sampling.

In Figure 3(b)'s example, two out of five of the segments are sampled (n=2). If agencies desire segments that contain more signs to be sampled, then the length of the line can be entertained to represent signage amount in within, as shown in Figure 3 (c), and the sampling result will vary correspondingly. Properly selected k can ensure that each station has at least one segment being sampled. This will be discussed in length in the example application.

<u>3.2.3 Spatially-Balanced GRTS Sampling for Optimal Maintenance Management</u>

The aforementioned method ensures the segments be ordered in the sequence of randomly labeled stations. And with properly chosen k, each station will have one or more segments selected. However, some agencies would prefer to have a spatially-balanced sample rather than a station-balanced sample when reporting LOM. To fulfill such spatially-balanced sampling feature, a hierarchical randomization can be applied to randomly order the generated addresses based on the quadrant-recursive function in GRTS.

In the classic GRTS design, a grid is divided into four cells, each of which is further divided into four subcells, and so on. The quadrant-recursive function is defined by the limit of the successive intensification within the grid. The recursion is carried through division and each division will pair the point with an address based on the order that the division was performed, where each digit of the address represents a step in the subdivision. A spatially referenced address can be constructed following the pattern of such partitioning. Random permutation that defines the hierarchical randomization is performed. Such recursive partitioning generates a nested hierarchy of grid, and puts the sampling process within the entire spatial context. Note that the order corresponds to the ranking obtained by reversing the sequence of the base-4 digits and treating the reversed sequence as a base-4 fraction. The reverse hierarchical order then gives a spatially well-balanced sample (Stevens and Olsen, 2004).

Different from the classic GRTS design that follows quadrant-recursive function with the resulting address appears as digits in a base-4 fraction, in a maintenance management setting, the segments are already partitioned within each station with varying sizes, leaving the difficulty of creating the address with a consistent base-N fraction. We remedy this with the following approach. Assume the maximum possible number of segments that a station has within the entire region is *N*, and the number of segments contained within station *i* is *n*, then the digit assigned to the ordered segment *j* can be expressed as:

$$T_{ij} = \begin{cases} 1, & when\ j = 1 \\ \frac{N}{j}, & otherwise \\ N, & when\ j = n \end{cases} \tag{6}$$

Reverse hierarchical order can again be applied to this base-N fraction via reversing the digits and sorting. The generated sequence will be reduced to the linear frame discussed in the previous section and is available for sampling with equal/unequal probability. The result will be a contiguous set of sample segments that are spatially-balanced, where the samples are well spread out over the population domain. Figure 4 demonstrates the process following the example shown in Figure 3.

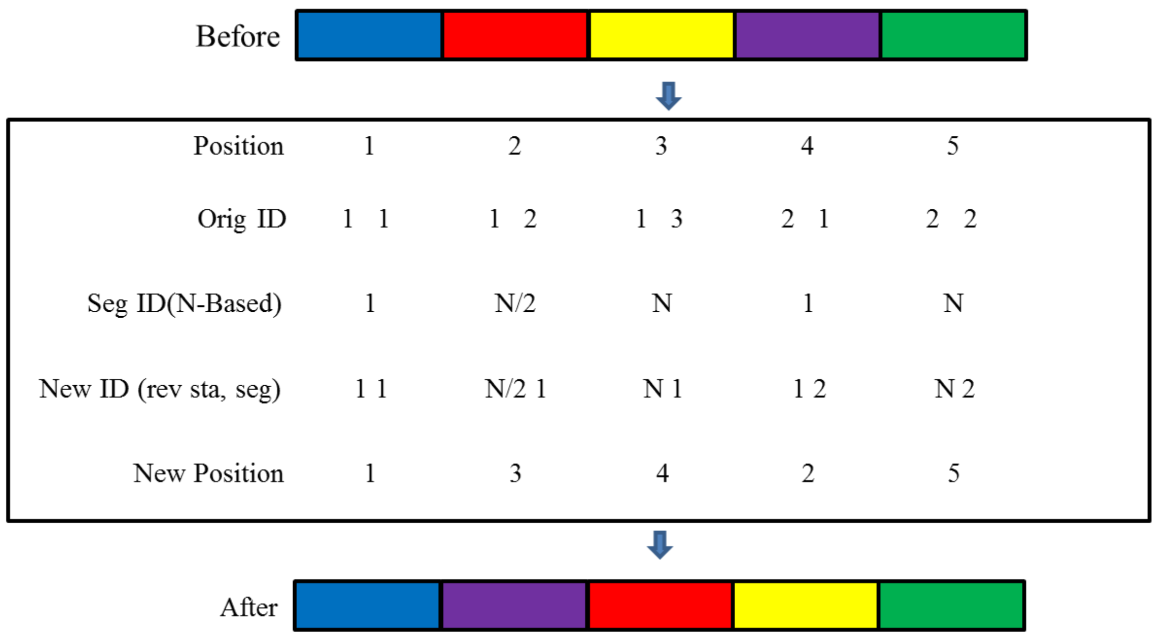


Figure 4 Reversed hierarchical order illustration following the example shown in Figure 3.

3.3 Sampling Method for Multiple Types of Infrastructures

The stratification for the multiple types of infrastructures sampling, as illustrated in Figure 5, consists of two major components: current condition estimation and high-dimensional cluster analysis. Current condition estimation "predicts" the infrastructure condition (e.g. in the form of LOM) based on historical records. This is to ensure that for the next round of inspection, sampling is conducted based on previous inspection results and deterioration rate of the infrastructure. High-dimensional cluster analysis then divides segments into clusters and selects representative segments as samples. Segments within each cluster share similar pattern with regard to infrastructure conditions. Thus by selecting segments across clusters, we select representative samples across all patterns. The sample size is a fixed percentage of segments in the network, constrained by labor or budget limits. Segments within each cluster are chosen randomly. Once the sampled segments are inspected, M&R activities can be further conducted accordingly on those segments whose performance is below certain threshold. The M&R records and inspection results will be applied to the next round of sampling process for inspection.

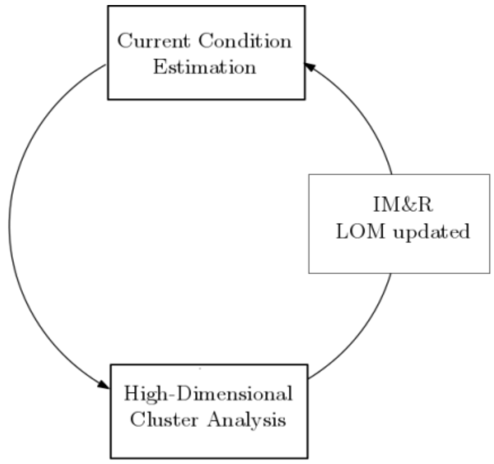


Figure 5 Illustration of stratification in the proposed method.

3.3.1 Current Condition Estimation

As the sampling unit for maintenance activities, segment possesses multiple features: infrastructure facilities (shoulder work, litter, weed, sweeping, etc.), geometric characteristics (number of lanes, segment length, etc.), and traffic information (AADT, peak hour volume, etc.), just to name a few. Each segment can therefore be described as a high-dimensional vector:

$$S_n = \{a_{shoulder\ work}, a_{litter\ pickup}, a_{weed}, \ldots;\ g_{length}, g_{lane\_num}, \ldots; t_{AADT}, t_{Peak\_Vol}, \ldots\}$$

where S_n refers to the segment n, $1 \leq n \leq N$, N is the number of segments in the network, and a, g, and t refer to the features associated with infrastructure, geometric, and traffic, separately. In this study, we will only consider infrastructure type and condition as segment features as they are the focus for sample selection.

The current condition estimation starts with translating the deterioration process of infrastructure on the segments into a deterioration matrix. The infrastructure conditions are described using 15 letter scores from A+ to F-. A+ represents the best condition and F- the worst. In previous studies, infrastructure deterioration has been considered as linear (Brint, 2006) or non-linear (Prozzi and Hong 2008) process. We assume that infrastructure deterioration is a linear process, yet the rate may vary across different types of infrastructures or different segments. For example, on Segment *1*, the time during which segment's shoulder condition deteriorates from A to A- is the same as the time from A- to B+. Yet in the meantime, the condition of littering might deteriorate from A to C. And on Segment *2,* while the shoulder deteriorates from A to A- on Segment *1*, the shoulder condition might deteriorate from A to B+.

The deterioration matrix is constructed based on the paired consecutive inspection records without any intervention (e.g. M&R) in between. In this study, we filtered out all the consecutive inspection records whose latter result was better than the former one. Yet exceptions might occur when the asset might still deteriorate to a worse condition even after repair or maintenance, in which case this method might underestimate deterioration rates of the infrastructures.

To simplify calculation, the 15 letter grades of infrastructures from A+ to F- are converted to numerical scores of 15 to 1. The deterioration process thus can be considered as the score is decreasing as time goes by. The deterioration rate of a segment is calculated as the score difference divided by time duration between two inspections.

The historical records in our study span years during which segments may be inspected multiple times. Therefore some segments might have more than one pair of consecutive records producing different historical deterioration rates. In such case, average of historical deterioration rates is employed as the deteriorate rate of the segment. For segments without such prior records, deterioration rate is replaced with the network-averaged value. For example, if no consecutive record for shoulder work is available on one segment, its deterioration rate of that segment is replaced with average shoulder work deterioration rate of all segments. Deterioration matrix is constructed as:

$$D = \begin{bmatrix} d_{seg1\_ShoulderWork}, & d_{seg1\_LitterPickup,} & d_{seg1\_IceSnow}, & \cdots \\ d_{seg2\_ShoulderWork}, & d_{seg2\_LitterPickup}, & d_{seg2\_IceSnow}, & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ d_{segN\_ShoulderWork}, & d_{segN\_LitrerPickup}, & d_{segN\_IceSnow}, & \cdots \end{bmatrix} \quad (7)$$

D can always be updated with latest inspection results and maintenance activities.

With the deterioration matrix constructed, we can estimate current conditions of infrastructures on each segment. Previous condition of infrastructure network is expressed as:

$$M_{Previous} = (S_1, S_2, \ldots, S_N)^T \quad (8)$$

where M represents the previous network infrastructure inspection conditions, S represents the infrastructure conditions within segment, with:

$$S_i = (s_{i\_ShoulderWork}, s_{i\_LitterPickup}, s_{i\_IceSnow}, \ldots) \quad (9)$$

The current conditions of infrastructure are estimated by considering previous conditions and inspection frequency, which is expressed as:

$$M_{Current} = M_{Previous} + tD \quad (10)$$

where $M_{Current}$ is the estimated current network infrastructure conditions, and t is time duration between previous and current inspections.

3.3.2 High-Dimensional Cluster Analysis

The key of high-dimensional cluster analysis is to jointly analyze all the infrastructure conditions on a segment rather than examining them individually. With the current infrastructure conditions estimated, LSH is implemented to define the similarity between segments. All segments are then divided into clusters based on the similarity matrix via spectral clustering. A fixed percentage of the segments can then be randomly chosen from each cluster. Figure 6 illustrates the LSH process in detail.

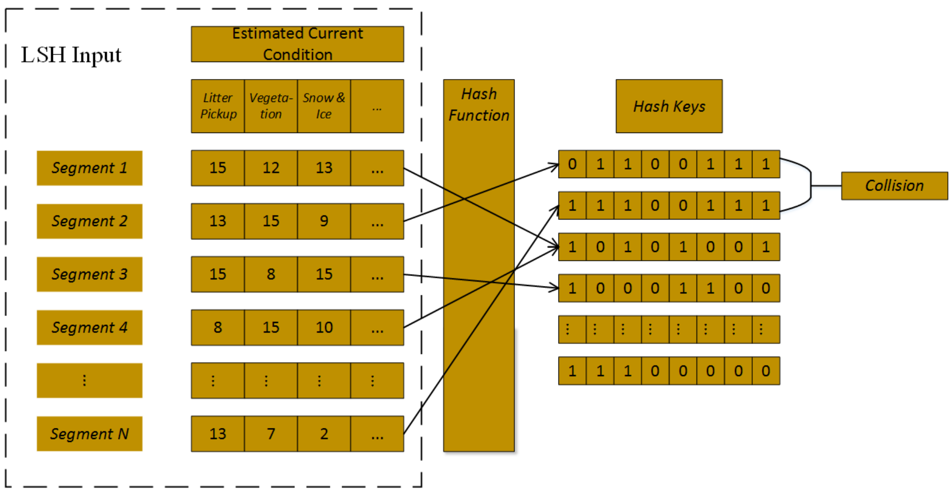


Figure 6 Illustration of LSH process.

The input to high-dimensional cluster analysis is the estimated current infrastructure conditions, including:

$$M_{current} = (S_1^*, S_2^*, \dots, S_N^*)^T \tag{11}$$

$$S_i^* = \{s_{i\_ShoulderWork}^*, s_{i\_LitterPickup}^*, s_{i\_IceSnow}^*, \dots\} \tag{12}$$

where S_i^* represents the estimated current infrastructure conditions on Segment i.

The first step in LSH is to define hash functions. In this study, we use inner product hash functions proposed by Charikar (2002). Hash function transforms a k-dimensional segment into a binary string. For example, in Figure 6, each segment is transformed into 8-digit binary strings. To determine the first digit of a binary string, we pick a k-dimensional vector $\boldsymbol{r} = (r_1, r_2, r_3, \dots, r_k)$. Each dimension ($r_1$,$r_2$, …) in vector $\boldsymbol{r}$ is randomly generated following Gaussian distribution. Then we calculate the inner product between the segment and $\boldsymbol{r}$ as:

$$h = \boldsymbol{r} \cdot S^* = r_1 s^*_{ShoulderWork} + r_2 s^*_{LitterPickup} + r_3 s^*_{IceSnow} + \cdots \tag{13}$$

where h is the inner product. When h is greater than or equal to 0, the first digit of the binary string is 1, and 0 otherwise. By repeating the process for eight times, 8-digit binary string is constructed. The binary strings are the hash keys of a hash function family. The same process applies to all segments, with each segment assigned a hash key. Note that some segments may have the same hash keys.

Since the hash keys are binary strings, we use Hamming distance as the difference measurement to compare them (Kulis and Grauman, 2009). For two strings with equal length, Hamming distance is defined as the number of digits at which the corresponding symbols are different. As illustrated in Figure 7, Hamming distance between the two strings is 1.

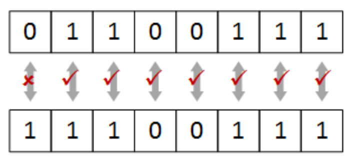


Figure 7 Illustration of Hamming distance.

When the difference between two segments' hash keys is less than a certain threshold, it is referred to as "collision" between the two segments and they are considered similar. As illustrated in Figure 6, hash key of Segment *1* is 01100111, and hash key of Segment *N* is 11100111. Hamming distance between the two hash keys is 1. If the threshold to define collision is set as 2, the difference between Segment *1* and *N*'s hash keys fulfills the requirement and thus the two segments are deemed similar.

Until this step, the LSH algorithm is fully implemented. Yet the algorithm can only determine whether two segments are similar or not rather than quantifying such similarity. Considering that hash function utilizes randomly generated vectors, using different vectors would lead to different hash keys. In Figure 6's example, Segment 1 and Segment N are considered similar. But if we generate another eight vectors, there is a probability that Segment 1 and Segment N are no longer similar. To remedy this, we perform LSH algorithm multiple times (i.e., 300 runs), and define similarity as the probability that two segments are similar across all the runs. For example, if Segment 1 and Segment N are identified as similar for 240 times out of 300 times, the similarity between them is $240/300 = 0.8$. With the similarity between each pair of segments in the network quantified, a matrix of similarity $\mathbb{S} = [\mathrm{Sim}_{\mathrm{ij}}]$is constructed, where $\mathrm{Sim}_{\mathrm{ij}}$ represents the similarity between segments i and j.

Then we apply spectral clustering, which is one of the most popular clustering algorithms due to its simplicity and efficiency (Piao et al., 2016; Von Luxburg, 2007). It is originated from partitioning clustering, which gives weights to links between data points and divides clusters by removing the least weighted links between clusters. Spectral clustering combines partitioning clustering with graph Laplacian matrices. The calculation is based on the spectrum of similarity matrix. The detailed computation is available in Appendix A.

3.4 Summary

In this chapter, the sampling methods for single and multiple types of infrastructures are presented. The method for single type infrastructure focuses on sample size determination and spatial-balanced sampling. Fisher information is applied to estimate the minimum sample size to obtain the infrastructure LOM of the network. Samples are selected with algorithm combining GRTS and hierarchical randomization. The method for multiple types of infrastructures is a high-dimension clustering-based stratified sampling method. It includes infrastructure deterioration estimation, similarity matrix construction, and stratified sampling. In the following chapters, the methods developed will be tested with the maintenance records collected by the MMQA program.

4.0 DATA COLLECTION

The proposed sampling methods are tested with highway infrastructure inspection records provided by the Utah MMQA Program. Previously, MMQA performed full inventory inspection for infrastructure maintenance. The maintenance personnel recorded total numbers of infrastructures to be maintained and deficient infrastructures on each segment. Then inspection records were entered into the MMQA+ software to calculate the LOM (letter grade). One motivation to develop an infrastructure sampling method is to reduce costs of infrastructure inspection by estimating the overall network LOM on a sample basis. For the State of Utah, the entire highway network is divided into 489 segments. Inspection was performed semi-annually from September 2014 to March 2016, with several segments inspected multiple times within one inspection period. The inspection record archives overall infrastructure condition, as well as segment id, infrastructure type, inspection date, and deficiency locations. There are more than 7,000 records in the database.

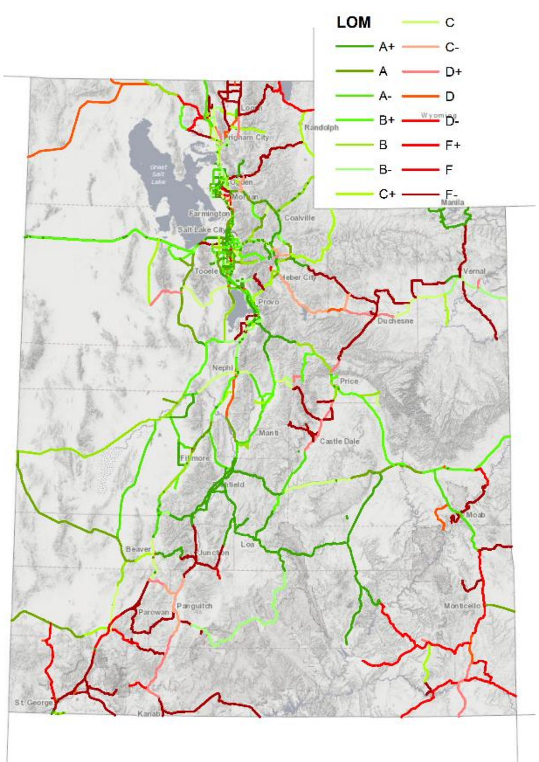
In Fall 2014, MMQA team launched MMQA Mobile, an ipad application that the inspection personnel use to record the defect assets. The MMQA mobile combines field inventory with integrated GPS/GIS mapping for maintenance data collection. Traditional data collection method only reports a total number of defects within a segment, which might result in significant bias due to human factor that cannot be validated. The MMQA Mobile platform, on the other hand, by enabling geotagging and description of defects on the ipad application, provides detailed geographic information of each asset deficiency as well as the asset condition. It adds another layer of credibility to the data, by allowing back-end post-processing to validate this data set collected from the crew for determining the LOM.

To test the sampling method for single type infrastructure, *Signage Repair and Replace* database is used. The grading scale for signage is shown in Table 1:

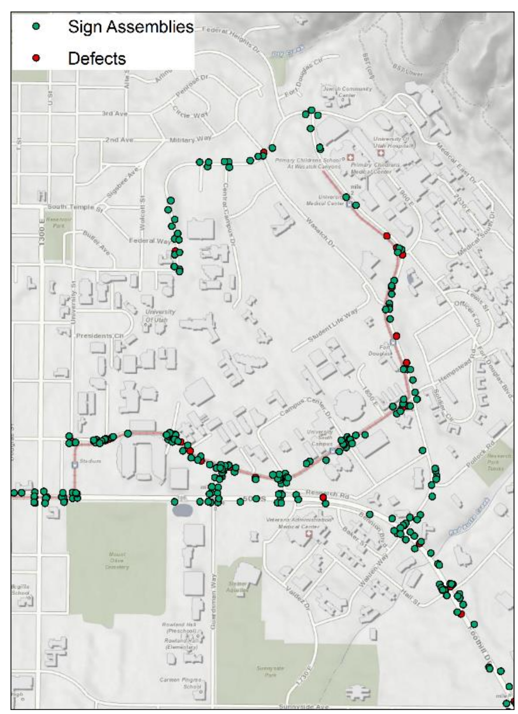
Table 1 LOM Grading Scale for Signage in MMQA

Percent Deficient	Grade	Percent Deficient	Grade
0.00-1.71	A+	13.41-14.99	C-
1.72-3.41	A	15.00-16.69	D+
3.42-5.00	A-	16.70-18.39	D
5.01-6.70	B+	18.40-19.99	D-
6.71-8.40	B	20.00-21.69	F+
8.41-10.00	B-	21.70-23.39	F
10.01-11.70	C+	23.40-100.00	F-
11.71-13.40	C		

The signage data used in this study were collected from September 2014 to March 2015 through MMQA Mobile at 100% signage coverage. There are a total of 67,259 sign assemblies statewide. More than 8,500 defect observations were recorded in the database. Figure 8 illustrates the maintenance network with segments color-coded to represent LOM during this data collection effort. A snapshot which is a sample zoom-in inspection on the signs in desire/deficient conditions is also shown.



(a)



(b)

Figure 8 Location of defects and color-coded roadway segment LOM: (a) LOM of the Utah roadway network (b) A sample snapshot of the signage inspection result

To test the sampling method for multiple types of infrastructure, 14 types of infrastructures are used in our study, including Shoulder Work (SW), Curb & Gutter (CG), Litter Pickup (LP), Weed Control (WC), Grade & Clean Ditches (GCD), Maintain Inlets (MI), Erosion Repair (ER), Pavement Markings (PM), Repair & Replace Signs (RRS), Repair & Replace Delineation (RRD), Guardrail Maintenance (GM), Sweeping (SP), Vegetation Control (VC), and Fence Maintenance (FM).

Table 2**Error! Reference source not found.** shows the average deterioration rate for each infrastructure. Note that the table only shows the aggregated (averaged) deterioration rates of all infrastructures. For example, the average deterioration rate of CG is 0.0996. It means that the conditions of CG deteriorate by 1 level (from A+ to A, or from A to A-) in approximately10 months on average. Yet on individual segments, the rate can be different. For example, deteriorate rate of CG on some segment can be 0.2, indicating that it takes 5 months for CG to deteriorate from A+ to A.

Table 2 Average Deterioration Rates of Infrastructures (per Month)

Infras	SW	CG	LP	WC	GCD	MI	ER
Det_rate	0.0756	0.0996	0.0821	0.0151	0.0394	0.0698	0.0813
Infras	PM	RRS	RRD	GM	SP	VC	FM
Det_rate	0.0181	0.0988	0.0244	0.0752	0.0022	0.0207	0.0825

In high dimensional cluster analysis, we use 14-digit binary string as hash key. The “collision” threshold is set as 2, indicating that when Hamming distance between the hash keys of two segments is less than 2, those two segments are similar. To avoid too many or too few segments in each cluster, all segments were divided into 10 clusters. For comparison purposes, SRS is also conducted. In the following section, both methods have been performed 50 times for sensitivity analysis.

5.0 DATA EVALUATION

5.1 Overview

In this chapter, we will apply the sampling methods presented in Chapter 3 to the maintenance records collected by the MMQA program. The analysis shows performance evaluation of the proposed sampling methods. Compared with SRS, the proposed methods show great improvements in terms of accuracy and efficiency, implying the potential application of these methods in infrastructure maintenance management.

5.2 Sampling Method for Single Type Infrastructure

Figure 9 shows a sensitivity analysis of α, β used in Equation (4) with regards to sample size (x axis in logarithmic scale). Note that although the signage sampling can be considered a Bernoulli process (that takes on value of either 0 or 1), in maintenance management, inspection is carried out on a segment basis. Since the signage condition on a segment can be considered as “desired” or “defect”, we adopt Fisher information for Bernoulli model to segment sampling. Correspondingly, segment condition is characterized as “desired” or “defect”. With N=2,048 (number of segments) for the State of Utah, Figure 9 demonstrates the optimal sample size that would yield the minimum Fisher information. It shows in the figure, for example, when sample size is equal or greater than 109, the probability of sample segments being significant at 90% level is 90 percent.

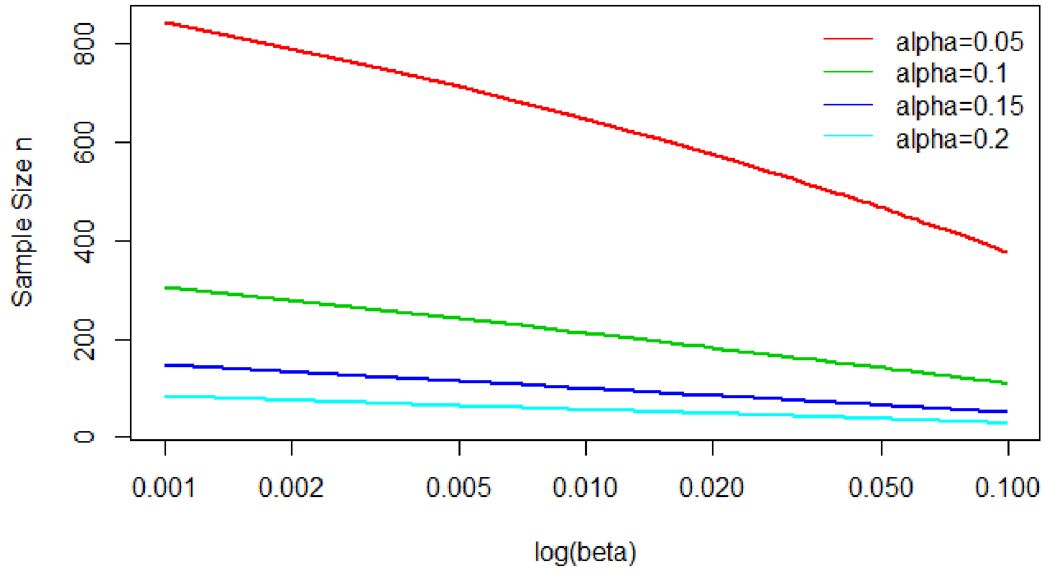


Figure 9 Sensitivity analysis of α, β with regard to segment sample size (x-axis in logarithmic scale).

To accommodate agencies' various sampling needs, four multi-density sampling schemes on the basis of GRTS are implemented in this study. Note that k is chosen as 500. This is selected due to the fact that the minimum number of segments that a station contains is 230. This threshold ensures that the majority of stations are sampled and avoids oversampling. Four sampling methods are explored as described in the previous section, and they are *GRTS Sampling with Equal Segment Weight*, *GRTS Sampling Weighted by Signage*, *Spatially-Balanced Sampling with Equal Segment Weight*, and *Spatially-Balanced Sampling Weighted by Signage*. Note that the unequal probability is implemented to assign segment length in the 1-dimenstional linear structure (Figure 3(c)) based on the number of signs each segment contains. Figure 10 presents sampling results using the foregoing four methods in the spatial context of the maintenance network. The sampled segments are highlighted in red. Note that due to the variation in segment length, the visualization might not reflect the actual sample size (with several short segments unrecognizable in the figure). The sampled number of segments for the four methods are 136, 133, 136, 134, separately, which all meet the requirements for maximum Fisher information. The latter two methods exhibit a spatially-balanced coverage.

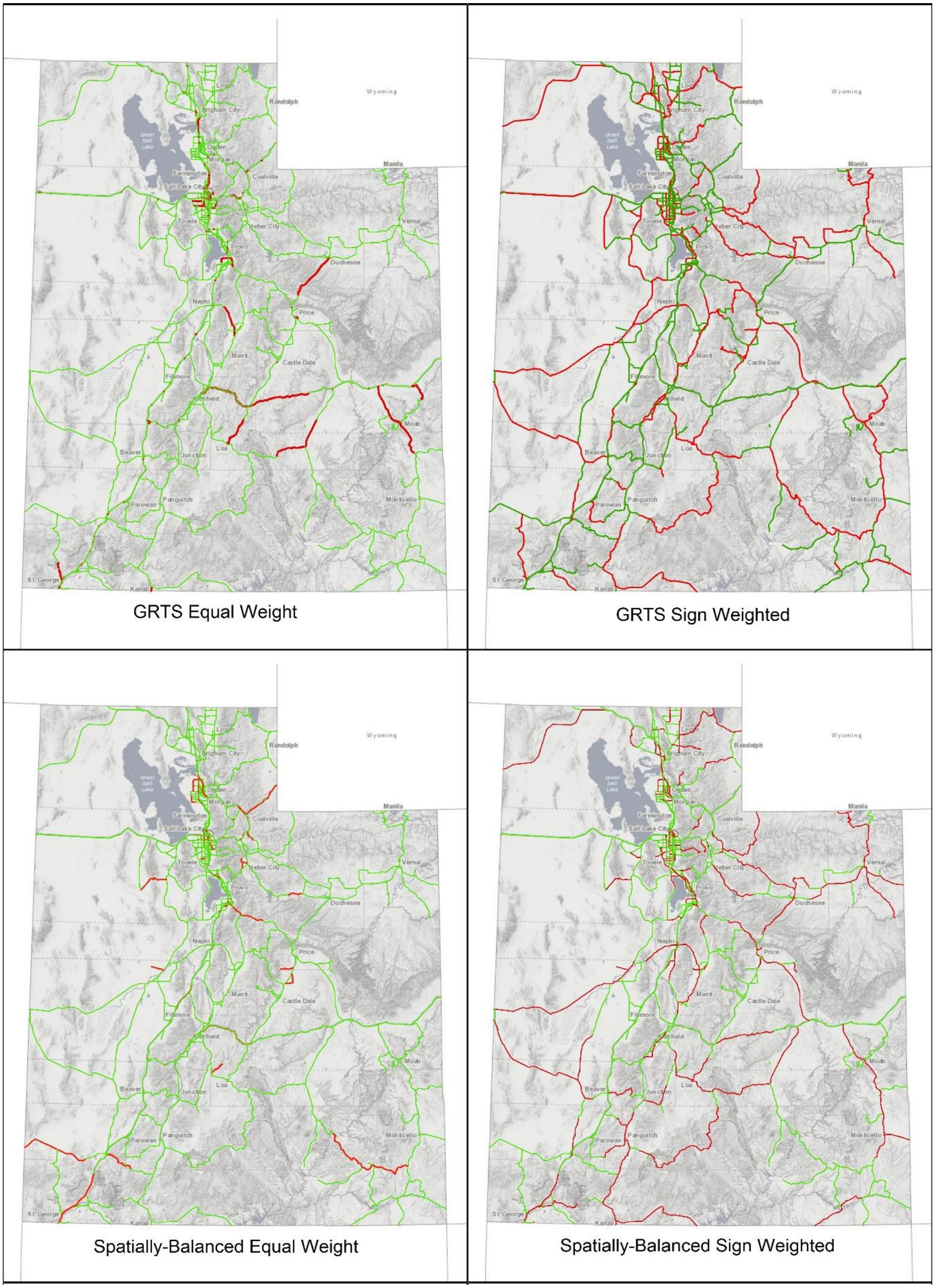


Figure 10 Maintenance sample segments by four sampling methods.

The purpose of sampling method development for MMQA is to provide accurate estimation of LOM at state/region/station levels for effective budget or resource allocation. Thus, LOM can be used as an index for assessing the effectiveness of different sampling schemes when compared against ground truth statewide inventory. Note that the MMQA Mobile data collected at 100% coverage is used as the ground truth data. Figure 11 shows the scaled LOM histogram for the sampled segments based on the four different sampling schemes compared against the ground truth data (red solid line). It is visually shown that *GRTS with Equal Weight* and *Spatially-Balanced with Equal Weight* match the statewide LOM pattern better than the rest.

Figure 11 LOM histogram for the sampled segments compared against the statewide inventory.

To quantitatively measure the sampling effectiveness, similarity analysis is conducted between statewide inventory and samples. The statewide asset LOM distribution can be represented as:

$$\mathrm{S}^{\mathrm{T}} = (P_{A+}^{T}, P_{A}^{T}, \ldots, P_{i}^{T}, \ldots, P_{F-}^{T}) \tag{14}$$

where $\mathrm{P}_{\mathrm{i}}^{T}$ is the true percentage of LOM i in the full inventory.

The expected LOS distribution from sample set $\mathfrak{R}$ is referred to as:

$$S_{\mathfrak{R}} = \left(P_{\mathfrak{R},1}, \dots, P_{\mathfrak{R},i}, \dots, P_{\mathfrak{R},n}\right) \tag{15}$$

where $P_{\mathfrak{R},i}$ is the percentage of LOM i in sample set $\mathfrak{R}$.

The similarity measure (*d*) between sample data and ground truth data is measured using Euclidean Distance:

$$d = \sqrt{\Sigma_{i=1}^{n}\left(P_{\mathfrak{R},i} - P_i^T\right)^2} \tag{16}$$

The higher the *d*, the lower the similarity between sample set and ground truth data. We performed 30 iterative sampling runs with the four proposed methods and simple random sampling, which is currently widely used by transportation agencies. The similarity analysis is shown in Table 3. It is noted that among the four proposed methods, *Spatially-Balanced with Equal Weight* yields the best result and matches the ground truth data most closely by giving the lowest similarity score and standard deviation. Both *GRTS Sampling with Equal Segment Weight* and *Spatially-Balanced with Equal Segment Weight* methods outperform the current simple random method with much lower average similarity score. Depending on the priority or specific goals set forth by the agencies (e.g. reflect statewide LOM, station-balanced, or spatially-balanced), the appropriate sampling method can be chosen accordingly.

Table 3 Similarity Analysis Result: Comparing the Ground Truth Inventory with Sampling Results

	Weighted by sign		Equal weighted		
	GRTS	Spatially-Balanced	GRTS	Spatially-Balanced	Simple Random
Average	0.35588	0.35308	0.03356	0.03213	0.062627
Std. Deviation	0.02546	0.02199	0.02067	0.01493	0.017191

5.3 Sampling Method for Multiple Types of Infrastructures

The purpose of infrastructure inspection is to assess infrastructure conditions and report LOMs of overall highway network for investment decisions. Ideally, infrastructures' grade (LOM) distribution, measured from samples, can reflect both overall condition and condition variation. To assess the effectiveness of our sampling method, the difference between condition estimated by samples and full inventory is computed with Root-Mean-Squared-Error (RMSE). For any infrastructure, the letter grade distribution is expressed as $(X_{A+}, X_A, X_{A-}, \dots, X_{F-})$, where X_i is the actual percentage of grade i in the full inventory (all segments). The grade distribution estimated from sample is expressed as $(x_{A+}, x_A, x_{A-}, \dots, x_{F-})$, where x_i is the estimated percentage of grade i among all the sampled segments.

The RMSE between estimated (from sample) and ground-truth grade distributions is then calculated as:

$$\text{RMSE} = \sqrt{\frac{\Sigma\,(x_i - X_i)^2}{15}} \tag{17}$$

RMSE reflects the error induced during the sampling process. As the value increases, estimated condition deviates from the ground-truth. To compare the performance of our proposed sampling method and SRS method, we conducted experiments of estimation accuracy between the two methods with the full inventory data collected by UDOT. We choose the most recent inspection records of each segment as the infrastructure conditions to be sampled and inspected, and the second most recent inspection records as the historical record based on which to estimate the current infrastructure conditions. To validate the robustness of the sampling method, particularly, its sensitivity to different data dimensionalities, a series of sensitivity tests are performed. Using the same highway network, sampling is conducted with 6, 8, 10 and 14 different types of infrastructures, separately. The types of infrastructures are randomly selected when the number of types is less than 14. Figure 12 shows the average RMSEs when sampling is conducted based on sample rates ranging from 5% to 30% of the entire segment inventory. The sampling rate for the proposed sampling method refers to the percentage of sample chosen from each stratum. Since the number of sample in each stratum is rounded to the nearest integer, the number of sample is always less than the same rate of the entire population. For example, when the sample rate is 10%, there are about 49 segments chosen as sample from the network with 489 segments. But in reality, the total number of segments chosen is less than 49. To make sure that the sampling methods are compared based on the same sample rate, the sample size of SRS method is same as the number of segments chosen by the stratified sampling method. It is noted that for low dimensional (less than 10 types of infrastructures) data, the average RMSEs show no significant difference when dimensionality changes. But when the dataset becomes high dimensional (more than 10 types of infrastructures), the average RMSEs start to demonstrate improvements. It further validates the effectiveness and suitability of our proposed sampling method for high-dimensional clustering analysis. The LSH algorithm is designed for data from high-dimensional space where the Euclidean distance is no longer valid as similarity measurements. As dimensionality increases, the proposed method tends to provide more accurate LOM estimation of the overall infrastructure condition.

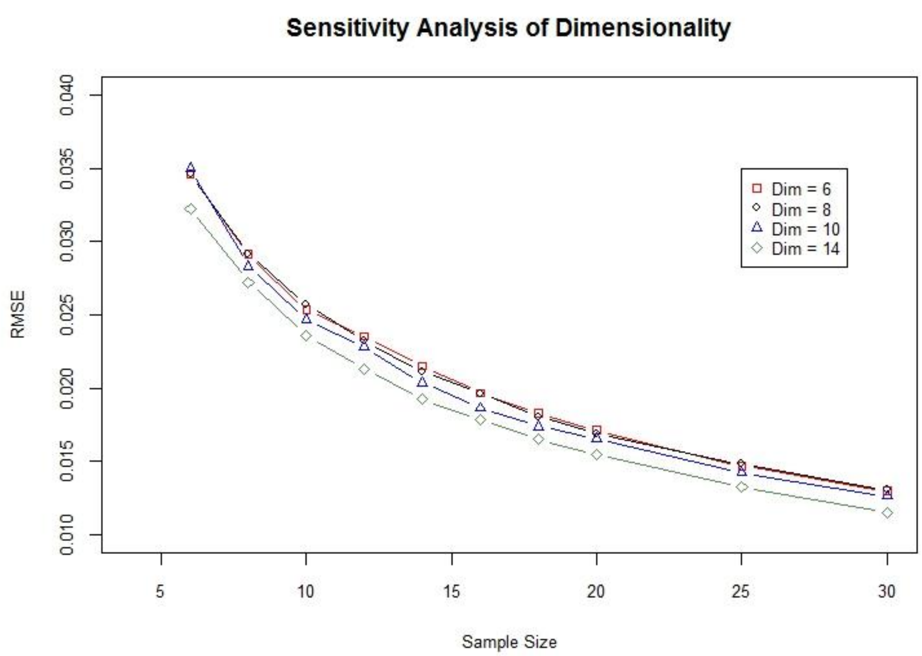
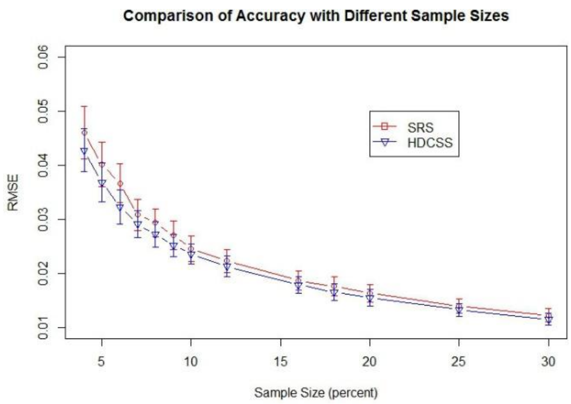
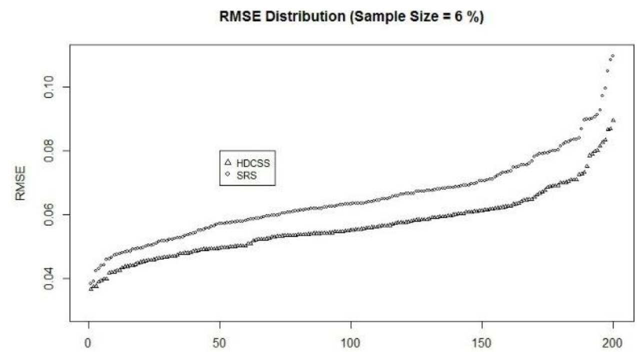


Figure 12 Sensitivity analysis of dimensionality (types of infrastructures) with different sample sizes.

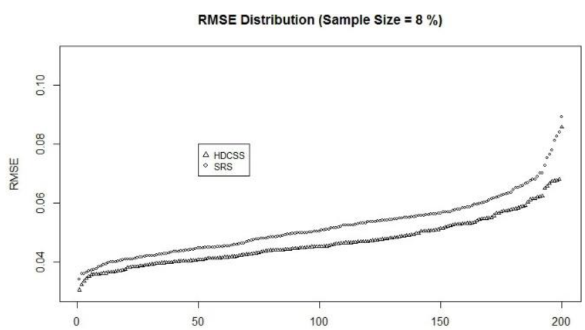
Figure 13(a) shows sensitivity analysis of sample size. Note that the RMSE is averaged out both across grades and across infrastructures. As shown in Figure 13, there is a trade-off between accuracy and sampling rate. Both average RMSE and standard deviations of RMSE decrease as the sampling rate increases. We observe a clear cut-off point at around 20% sampling rate, where the RMSE drops significantly as the sampling rate increases until 20%. After that, the trend becomes mild. The HDCSS method constantly outperforms SRS by providing lower average RMSE. Figure 13(b-d) show the RMSE distributions with sample size of 6%, 8%, and 10%. When sample size is less than 10 percent, there is a distinct difference between the performances of two sampling methods.



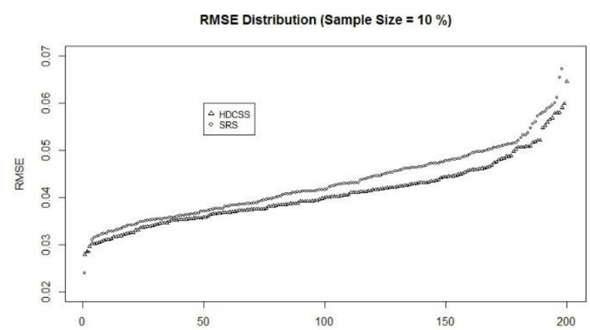
(a)



(b)



(c)



(d)

Figure 13 Sensitivity analysis of sample sizes between SRS and HDCSS methods.

One highlight of our proposed method is that the selected sample segments can accurately reflect the LOMs of all the infrastructures throughout the network. Figure 14 provides a detailed look on the sampling accuracy for each asset using SRS and our proposed method, where RMSE (mean and standard deviation) is shown for all 14 assets with a 20% sampling rate.

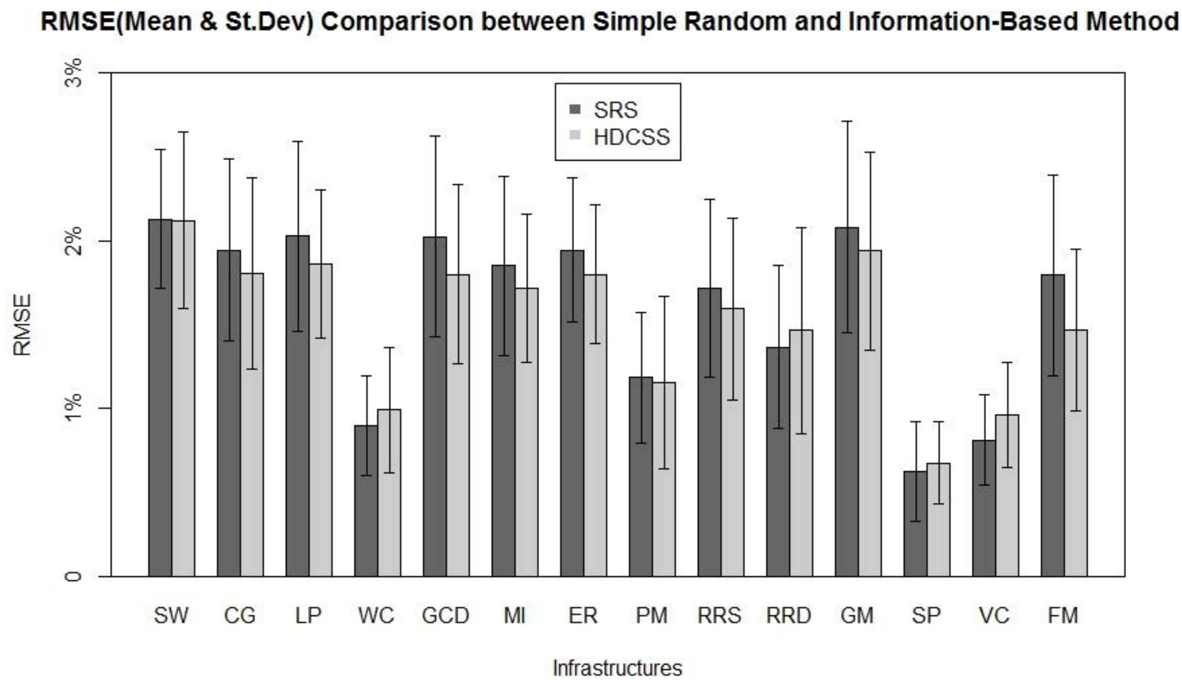


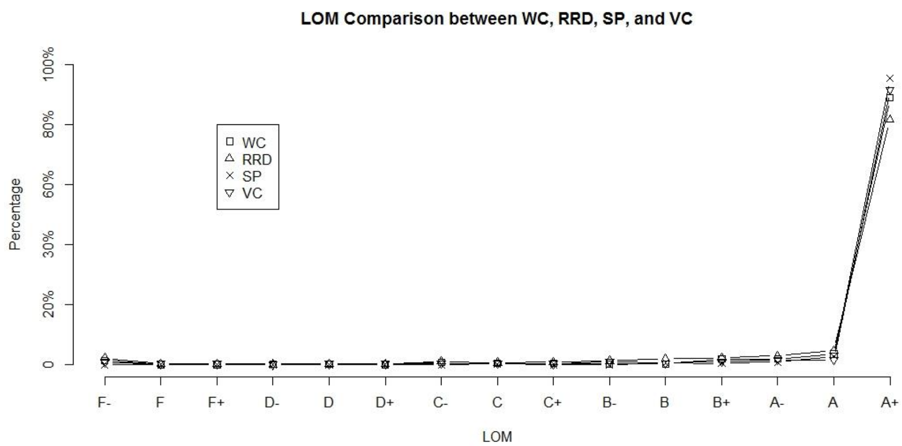
Figure 14 Comparison of RMSE (Mean and Standard Deviation) between SRS Method and HDCSS Method

As seen in Figure 14, the RMSEs are similar between two methods but vary significantly across infrastructures. For most types of infrastructures, SRS has higher RMSE than the proposed method, indicating the superiority of our proposed method. However, also note that for certain infrastructures (WC, RRD, SP, and VC), SRS yields lower RMSE. To further explore the reasons, we compared the LOM distributions between each type of these infrastructures.

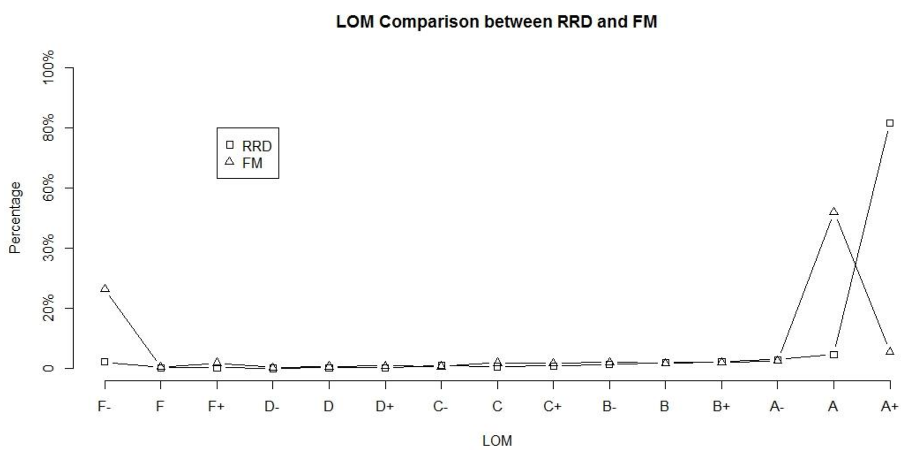
Figure 15(a) shows the ground truth grade distributions of WC, RRD, SP, and VC. For these infrastructures, more than 80% of segments are of A+ grade. Under such circumstance, since the difference between individual samples is insignificant, it is high likely that choosing different samples would not influence the result much. An extreme case in such situation is that if all the segments are of grade A+, then samples selected by any method would yield the same result. For infrastructures with such skewed grade distribution, both methods would estimate the overall conditions with low errors. Yet one unique aspect of high-dimensional cluster analysis is that, when one dimension in the high dimensional vectors lacks of variation, clustering relies more on other dimensions, and the importance (weight) of that dimension thus diminishes.

Correspondingly, the overall condition of that infrastructure (with little variation) is less represented by the sample selected by HDCSS than randomly picked. That explains the underlying reason of the low RMSEs for the four types of infrastructures and the outperformance of SRS for them.

The relation between the LOM distribution and RMSE is reflected in the ranking of RMSE values of these four infrastructures. As shown in Figure 15(a), the four infrastructures, ranked by the percentage of grade A+ for each type in descending order, are SP, VC, WC, and RRD. This sequence is exactly the same as the RMSE ranking using both methods.



(a)



(b)

Figure 15 Grade distribution comparison between infrastructures: (a) WC, PM, RRD, SP, and VC; (b) FM and RRD.

Another interesting phenomenon observed from Figure 14 is that the average RMSEs of two infrastructures, FM and RRD yield same result with our proposed method. However, the values are quite different with SRS. In Figure 15(a), it is shown that for FM, almost 80% segments are either of grade A+ or grade F. Thus, the entire grade distribution is quite dispersed due to the occurrence of two dominant grades. In such case, HDCSS method selects samples from both dominant grades yet such scenario is not guaranteed with SRS. As the distribution shifted to a single peak instead of two (see Figure 15(b) for the comparison), the RMSE of SRS increases from around 1.5% to 1.9%. As shown in Figure 15(b), with the rest grades remain at a very low percentage, two dominant grades have significantly high percentages and those two percentages are comparable.

In infrastructure inspection sampling, one prominent concern is to reduce the sample rate without too much compromise in accuracy, since the sampling rate is directly tie to costs and budget allocation. According to (Schmitt et al., 2006), lead states assume that the infrastructure conditions follow normal distributions, so the sampling rate can be estimated with given confidence interval and accuracy. For example, North Carolina DOT performs sampling based on 90-95% confidence interval and 6% accuracy. Virginia DOT requires the confidence interval of sampling be at 95% and the accuracy of 4%. However, there is lack of evidence to justify the assumption. For comparison purpose, we define an “accuracy rate” for each method, representing the probability of a sample being considered as accurate within certain error threshold. The sampling result is considered “accurate” if and only if the errors between the estimated conditions of all assets and ground truth are within an acceptable range. The error is still quantified via RMSE.

Figure 16 shows the sensitivity analysis of accuracy rate when different sample sizes apply. It is noted that under the same error threshold, when the sampling rate is less than 20 percent, our proposed method always yields higher accuracy rate than SRS. For an accuracy rate of 90 percent with error threshold of 0.06, the required sampling rate is around 8% for our method as opposed to 10% for SRS. To achieve an accuracy rate of 95 percent with error threshold of 0.4, by using the HDCSS method, the sample size can be reduced from 20 percent to 16 percent. Such decreased sample size can bring a significant reduction in inspection costs for infrastructure management, especially for large scale highway network.

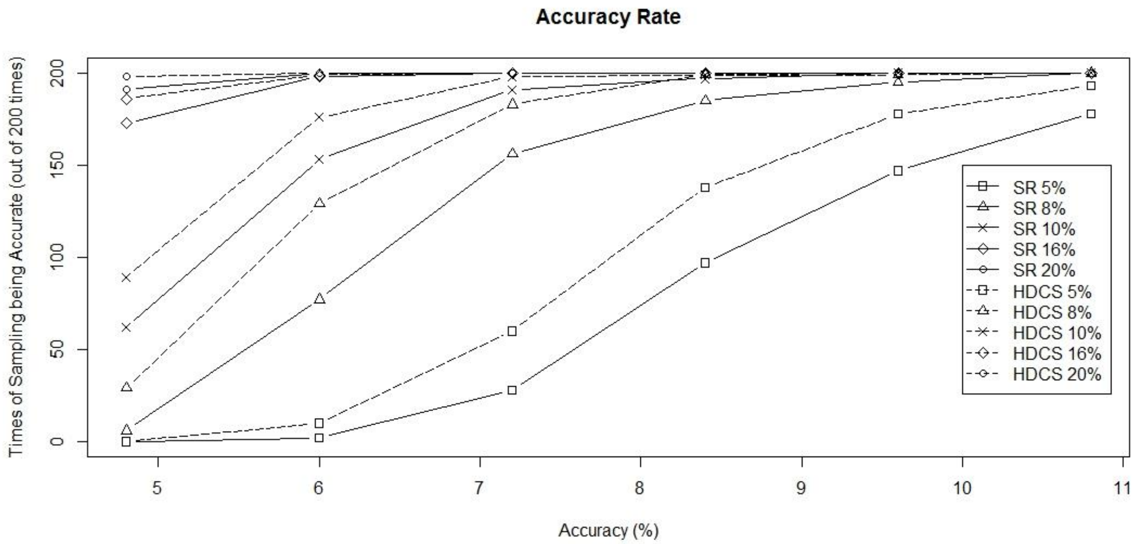


Figure 16 Sensitivity analysis of accuracy rates under different error thresholds and with sample sizes.

To further explore the sample rate reduction quantitatively, we performed one-way ANOVA test to analyze the difference of errors between the two sampling methods with different sample rates. The results of ANOVA tests show how the sample size changes with different sampling methods when there is no significant difference between the sampling results. Table 4 shows that result of ANOVA test between the errors of samples selected by HDCSS method with sample rate of 16 percent and SRS method with sample rate of 18 percent.

Table 4 Result of ANOVA for Errors of Samples Selected by HDCSS (16%) and SRS (18%) Methods

	Degree of Freedom	Sum of Squares	Mean Square	F-Value	P-Value	Significantly Different, Yes/No
Between Features	1	0.0000033	3.314e-06	1.28	0.259	No
Within Features	398	0.0010306	2.589e-06			

The results of ANOVA concludes that there is no significant difference between the errors of estimated infrastructure conditions by using the proposed method with sample rate of

16 percent and SRS with sample rate of 18 percent. That is to say, for any ongoing sampling scheme using SRS method with 18 percent of the population as the sample rate, our method can effectively reduce the sampling rate to 16%. Similar sample rate reduction results have been observed under other precision requirements, as shown in Table 5. It is observed that when the sample rate of SRS is below 15 percent, our proposed method can reduce the sample rate by 1 percent. When the sample rate SRS is above 15 percent, it can reduce the sample rate by 2 percent. And most notably, when the sampling method is applied to large highway networks, these reductions in sample rates can significantly reduce the inspection costs.

Table 5 Results of ANOVA Tests

HDCSS Sample Rate (%)	SRS Sample Rate (%)	P-Value
8	9	0.51
10	11	0.73
11	12	0.806
12	13	0.814
16	18	0.119
18	20	0.259
19	21	0.298

5.4 Summary

In this chapter, in-depth analysis of the sampling results from the proposed two methods has been conducted. The sampling methods are tested with infrastructure maintenance management records provided by the MMQA. In the sampling method for single type infrastructure, the minimum sample size estimated with Fisher information varies as the sample population and required accuracy change. The sampling results indicate that both *GRTS with Equal Weight* and *Spatially-Balanced with Equal Weight* outperform SRS method in estimating the infrastructure LOM in the network. In the sampling method for multiple types of infrastructures, the accuracy of the sampling results is closely associated with the number of infrastructure types. The error between estimated LOM and ground truth decreases as the number of dimensions (types of infrastructures) increases. By applying the HDCSS, the sample size can be reduced as much as 2% comparing with SRS.

6.0 CONCLUSIONS

6.1 Summary

This project focuses on developing systematic approaches to assess “where and to what extent” to collect infrastructure conditions with maximum information retained for LOM estimation. When there is only one type of infastructure in the network, the method combining Fisher information and spatial sampling techniques is proposed. It can be customized based on local agencies’ requirements, such as station-balanced, spatially-balanced, or functional class based. These requirements are rooted in the very fundamentals of maintenance management. Fisher information is applied in the study to determine asset sample size, and GRTS based sampling methods are implemented to entertain sampling priorities set forth by the agencies. The basic idea of GRTS method is to create a quadrant-recursive function that maps two-dimensional space into a one-dimensional one, thereby defining an ordered spatial address for the population. Unequal probability sampling can be achieved by giving each point a length proportional to its inclusion probability. Coupled with hierarchical randomization, the method is able to offer a spatially well-balanced sample.

To estimate the infrastructure LOMs in a network with multiple types of infrastructures, HDCSS method is proposed. The sampling segments selected by this method can accurately represent the overall conditions of the full infrastructure inventory. The method consists of two components: current condition estimation and high-dimensional cluster analysis. The *current condition estimation* aims at providing predicated infrastructure condition for cluster analysis based on historical inspection records. In *high-dimensional cluster analysis,* segments with multiple types of infrastructures are considered as high-dimensional vectors. By applying LSH algorithm and spectral clustering, the similarities of segments are measured and segments are assigned to clusters.

6.2 Findings

The findings in this project are summarized as follows:

6.2.1 Sampling Method for Single Type Infrastructure

The sampling method is showcased via an example application of the Signage Repair & Replace database maintained by the UDOT. It is enhanced on the basis of GRTS design by tailoring it to the maintenance setting. Different from the classic GRTS scheme that follows quadrant-recursive function with the resulting address appears as digits in a base-4 fraction, an innovative algorithm is developed to create address for each segment with a base-N fraction given the fact that segments are already partitioned within each station with varying sizes. Four sampling methods that might be tempered to various needs are implemented including *GRTS Sampling with Equal Segment Weight*, *GRTS Sampling Weighted by Signage*, *Spatially-Balanced Sampling with Equal Segment Weight*, and *Spatially-Balanced Sampling Weighted by Signage*. The sampling results are presented and compared against ground truth asset inventory. The method lays a strong theoretical foundation for the maintenance asset sampling based on the customized requirements/needs for local agencies and is effective in estimating LOM at state/region/station levels for budget allocation.

6.2.2 Sampling Method for Multiple Types of Infrastructures

Using the inspection records from the State of Utah, our proposed method outperforms SRS for most types of infrastructures, especially under the circumstances where LOM varies greatly within infrastructures. For the infrastructures when most of the segments are of similar condition, both HDCSS method and SRS yield low errors. HDCSS method can effectively reduce the sample rate without compromise in accuracy compared with the SRS method, leading significant decrease in inspection costs, especially for large scale networks. By using the proposed sampling method, DOTs can obtain accurate estimation of infrastructure LOMs for network with multiple types of infrastructures, saving enormous resources and time for infrastructure inspection.

6.3 Recommendations

Both sampling methods proposed in this study outperform the SRS method currently applied by DOTs. The method for single type infrastructure estimates the sample size based on data-driven analytics, and is able to provide spatially-balanced samples. The method for multiple types of infrastructures reduces the sample size and enables inspection personels to collect the LOMs of all types of infrastructures simultaneously. Based on this study, several intriguing topics emerge. First, from the sampling results, it is observed that the accuracy is heavily influenced by the LOM distribution. It is interesting to test the sampling method for single type infrastructure with maintenance records of different types of infrastructures, e.g., litter pickup, vegetation control. Second, as an important component of the method for multiple types of infrastructures, deterioration matrix construction can significantly influence the accuracy of the sampling method. It is necessary to apply more rigorous data analysis tool to enhance the estimation of deterioration process. Third, it is desired to involve other more efficient high-dimensional cluster analysis methods in the sampling process which can potentially improve the accuracy of the sampling results.

REFERENCES

Aggarwal, C., 2001. Re-Designing Distance Functions and Distance-Based Applications for High Dimensional Data. ACM SIGMOD Rec. 30, 256–266. doi:10.1145/375551.383213

Bellman, R.E., 1961. Adaptive Control Processes: A Guided Tour.

Brint, A.T.T., 2006. Sampling on Successive Occasions to Re-Estimate Future Asset Management Expenditure. Eur. J. Oper. Res. 175, 1210–1223. doi:10.1016/j.ejor.2005.06.031

Charikar, M.S., 2002. Similarity Estimation Techniques from Rounding Algorithms. Proc. thiry-fourth Annu. ACM Symp. Theory Comput. - STOC '02 380–388. doi:10.1145/509907.509965

Datar, M., Immorlica, N., Indyk, P., Mirrokni, V.S., 2004. Locality-Sensitive Hashing Scheme Based on P-Stable Distributions. Proc. Twent. Annu. Symp. Comput. Geom. 253–262. doi:10.1145/997817.997857

De la Garza, J.M., Piñero, J.C., Ozbek, M.E., 2008. Sampling Procedure for Performance-Based Road Maintenance Evaluations. Transp. Res. Rec. J. Transp. Res. Board 2044, 11–18. doi:10.3141/2044-02

Gharaibeh, N.G., Garber, S.I., Liu, L., 2010. Determining Optimum Sample Size for Percent-Within-Limits Specifications. Transp. Res. Rec. J. Transp. Res. Board 2151, 77–83. doi:10.3141/2151-10

Goodchild, M., Haining, R., Wise, S., 1992. Integrating GIS and spatial data analysis: problems and possibilities. Int. J. Geogr. Inf. Syst. 6, 407–423. doi:10.1080/02693799208901923

Hinneburg, A., Keim, D.A., 1999. Optimal Grid-Clustering: Towards Breaking the Curse of Dimensionality in High-Dimensional Clustering 506–517.

Jain, P., Kulis, B., Grauman, K., 2008. Fast Image Search for Learned Metrics. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 1–8. doi:10.1109/CVPR.2008.4587841

Kulis, B., Grauman, K., 2012. Kernelized Locality-Sensitive Hashing. IEEE Trans. Pattern Anal. Mach. Intell. 34, 1092–1104. doi:10.1109/TPAMI.2011.219

Kulis, B., Grauman, K., 2009. Kernelized Locality-Sensitive Hashing for Scalable Image Search. Proc. IEEE Int. Conf. Comput. Vis. 2130–2137. doi:10.1109/ICCV.2009.5459466

Lee, M., Wagenmakers, E., 2013. Bayesian modeling for cognitive science: A practical course. Cambridge UP.

Li, C., Chang, E., Garcia-Molina, H., Wiederhold, G., 2002. Clustering for Approximate Similarity Search High-Dimensional Spaces. IEEE Trans. Knowl. Data Eng. 14, 792–808. doi:10.1109/TKDE.2002.1019214

Liu, X., Yu, L., 2009. Development of Geolocation Data Compression for Transportation Target Identification. Transp. Res. Rec. J. Transp. Res. Board. doi:10.3141/2105-10

Ly, A., Verhagen, J., Grasman, R., 2014. A Tutorial on Fisher Information 1–55.

McCullouch, B., Sinha, K., 2003. Maintenance Quality Assurance Program, JTRP Technical Reports. doi:10.1080/07399018508967771

Medina, R. a., Haghani, A., Harris, N., 2009. Sampling Protocol for Condition Assessment of Selected Assets. J. Transp. Eng. 135, 183–196. doi:10.1061/(ASCE)0733-947X(2009)135:4(183)

Mishalani, R.G., Gong, L., 2009a. Optimal Infrastructure Condition Sampling over Space and Time for Maintenance Decision-Making under Uncertainty. Transp. Res. Part B Methodol. 43, 311–324. doi:10.1016/j.trb.2008.07.003

Mishalani, R.G., Gong, L., 2009b. Optimal Sampling of Infrastructure Condition: Motivation,

Formulation, and Evaluation. J. Infrastruct. Syst. 15, 313–320. doi:10.1061/(ASCE)1076-0342(2009)15:4(313)
Mishalani, R.G.R.R.G., Goel, P.K.P.P.K., NEXTRANS, C., 2011. Optimal Condition Sampling for a Network of Infrastructure Facilities.
Myung, I., 2003. Tutorial on maximum likelihood estimation. J. Math. Psychol.
Ng, A.Y., Jordan, M.I., Weiss, Y., 2001. On Spectral Clustering: Analysis and an Algorithm. Adv. Neural Inf. Process. Syst. 14 14, 849–856. doi:10.1.1.19.8100
Piao, X., Hu, Y., Gao, J., Sun, Y., Lin, Z., 2016. A Submodule Clustering Method for Multi-way Data by Sparse and Low-Rank Representation.
Ripley, B., 2005. Spatial Statistics. John Wiley & Sons.
Rissanen, J., 1996. Fisher information and stochastic complexity. Inf. Theory, IEEE Trans.
Schmitt, R.L.R., Owusu-ababio, S., Weed, R.M.R., Nordheim, E.E. V, Schmitt, R.L.R., 2006. Development of a Guide to Statistics for Maintenance Quality Assurance Programs in Transportation.
Selezneva, O., Mladenovic, G., Speir, R., Amenta, J., Kennedy, J., 2004. National Park Service Road Inventory Program: Quality Assurance Sampling Considerations for Automated Collection and Processing of Distress Data. Transp. Res. Rec. 1889, 106–105. doi:10.3141/1889-12
Steinbach, M., Ertöz, L., Kumar, V., 2004. The Challenges of Clustering High Dimensional Data, in: New Directions in Statistical Physics. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 273–309. doi:10.1007/978-3-662-08968-2_16
Stevens, D.L., Olsen, A.R., 2004. Spatially Balanced Sampling of Natural Resources. J. Am. Stat. Assoc. 99, 262–278. doi:10.1198/016214504000000250
Stevens, S., 1957. On the psychophysical law. Psychol. Rev. 64(3), 153–181.
Templeton, C.J., Lytton, R.L., 1984. Estimating Pavement Condition and Rehabilitation Costs Using Statistical Sampling Techniques.
Towsley, B., Bolot, T., Ribeiro, B., Towsley, D., Ye, T., Bolot, J.C., 2006. Fisher information of sampled packets: an application to flow size estimation, in: Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement. pp. 15–26. doi:10.1145/1177080.1177083
Von Luxburg, U., 2007. A Tutorial on Spectral Clustering. Stat. Comput. 17, 395–416.
Wang, J., Haining, R., Cao, Z., 2010. Sample surveying to estimate the mean of a heterogeneous surface: reducing the error variance through zoning. Int. J. Geogr. Inf. Sci. doi:10.1080/13658810902873512
Wang, J., Stein, A., Gao, B., Ge, Y., 2012. A review of spatial sampling. Spat. Stat. 2, 1–14. doi:10.1016/j.spasta.2012.08.001
Yurek, R., Albright, N., Brandenburg, J., Haubrich, M., Hendrix, L., Hillis, D., Rodriguez, L., Zimmerman, K., 2012. Best Practices In Performance Measurement For Highway Maintenance And Preservation.

APPENDIX A

SPECTRAL CLUSTERING ALGORITHM (Ng et al., 2001):

Given a set of points $V = \{v_1, v_2, \dots, v_n\}$, the similarity matrix $S = \{s_{ij}\}$, where s_{ij} refers to the similarity between v_i and v_j

1. Define D to be the diagonal matrix $D_{ii} = \Sigma_{\mathrm{j}} A_{ij}$, and construct the matrix $L = D^{-1/2} A D^{-1/2}$
2. Find $x_1, x_2, \dots, x_k$, the k largest eigenvectors of L, and form the matrix $X = [x_1, x_2, \dots x_k]$ by stacking the eigenvectors in columns.
3. Form the matrix Y from X by renormalizing each of X's rows to have unit length (i.e. $Y_{ij} = X_{ij} / \left(\Sigma_j X_{ij}^2\right)^{1/2}$).
4. Treating each row of Y as a point in $\Re^k$, cluster them into k clusters via K-means.
5. Finally, assign the original point v_i to cluster j if and only if row i of the matrix Y was assigned to cluster j.