VTRC
Virginia Transportation Research Council

We Bring Innovation to Transportation

Reliability of Nondestructive Evaluation of Concrete Bridge Decks

https://vtrc.virginia.gov/media/vtrc/vtrc-pdf/vtrc-pdf/26-R25.pdf

SOUNDAR S.G. BALAKUMARAN, Ph.D., P.E.
Associate Director for Structures

AMIR BEHRAVAN, Ph.D., P.E.
Research Scientist

Final Report VTRC 26-R25

VIRGINIA TRANSPORTATION RESEARCH COUNCIL
530 Edgemont Road, Charlottesville, VA 22903-2454
vtrc.virginia.gov

Standard Title Page - Report on Federally Funded Project

1. Report No.: FHWA/VTRC 26-R25	2. Government Accession No.:	3. Recipient's Catalog No.:

4. Title and Subtitle: Reliability of Nondestructive Evaluation of Concrete Bridge Decks	5. Report Date: December 2025
	6. Performing Organization Code: VTRC
7. Author(s): Soundar Balakumaran, Ph.D., P.E. and Amir Behravan, Ph.D., P.E.	8. Performing Organization Report No.: VTRC 26-R25

9. Performing Organization and Address: Virginia Transportation Research Council, 530 Edgemont Road, Charlottesville, VA 22903	10. Work Unit No. (TRAIS):
	11. Contract or Grant No.: 115560
12. Sponsoring Agencies' Name and Address: Virginia Department of Transportation, 1221 E. Broad Street, Richmond, VA 23219; Federal Highway Administration, 400 North 8th Street, Room 750, Richmond, VA 23219-4825	13. Type of Report and Period Covered: Final
	14. Sponsoring Agency Code:

15. Supplementary Notes:
This is an SPR-B report

16. Abstract:

Although visual inspections remain the primary method for bridge deck evaluation, they fail to detect hidden deterioration. Rapid-scanning nondestructive evaluation (NDE) techniques are gaining popularity, but their reliability compared with traditional methods is unclear. This study aimed to determine the reliability of multiple NDE methods for detecting deterioration in concrete bridge decks. This study focused on an interstate bridge for which both in-depth and rapid-scanning NDE evaluations were conducted over 6 years. Inspection data from consultants employing various methods, including manual sounding, infrared thermography, automated sounding, and ground penetrating radar, were compared with ground truth coring results.

The study revealed significant inconsistencies in results across all methods and consultants. Manual chain drag sounding offered reliable results, followed by automated sounding with potential for improvement. Infrared thermography exhibited uniformly poor performance in detecting delamination. Limited evaluation of the emerging time-lapse infrared thermography looked promising, but further improvement is needed.

Matching researchers' expectations, the literature shows that although this study employed the most robust ground truth comparison in a field study compared with the existing literature, limitations still hinder the formation of definitive conclusions. These indefinite conclusions highlight the challenge in field studies where getting numerous ground truth cores for statistical significance is impractical. Although no technology was identified as ideal to serve as the control, this study discusses meaningful ways to improve the reliability of some promising NDE methods.

This study recommended developing a qualifying mockup testing program for NDE methods, operators, and equipment before contracts can be awarded. The study also recommended developing specifications and engineering guidance for NDE methods.

17 Key Words: nondestructive testing, reliability, bridge deck inspection, Federal ID: 1784	18. Distribution Statement: No restrictions. This document is available to the public through NTIS, Springfield, VA 22161.

19. Security Classif. (of this report): Unclassified	20. Security Classif. (of this page): Unclassified	21. No. of Pages: 21	22. Price:

FINAL REPORT

RELIABILITY OF NONDESTRUCTIVE EVALUATION OF CONCRETE BRIDGE DECKS

Soundar S.G. Balakumaran, Ph.D., P.E.
Associate Director
Virginia Transportation Research Council

Amir Behravan, Ph.D., P.E.
Research Scientist
Virginia Transportation Research Council

In Cooperation with the U.S. Department of Transportation
Federal Highway Administration

Virginia Transportation Research Council
(A partnership of the Virginia Department of Transportation
and the University of Virginia since 1948)

Charlottesville, Virginia

December 2025
VTRC 26-R25

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the Virginia Department of Transportation, the Commonwealth Transportation Board, or the Federal Highway Administration. This report does not constitute a standard, specification, or regulation. Any inclusion of manufacturer names, trade names, or trademarks is for identification purposes only and is not to be considered an endorsement.

ABSTRACT

Although visual inspections remain the primary method for bridge deck evaluation, they fail to detect hidden deterioration. Rapid-scanning nondestructive evaluation (NDE) techniques are gaining popularity, but their reliability compared with traditional methods is unclear. This study aimed to determine the reliability of multiple NDE methods for detecting deterioration in concrete bridge decks. This study focused on an interstate bridge for which both indepth and rapid-scanning NDE evaluations were conducted during 6 years. Inspection data from consultants employing various methods, including manual sounding, infrared thermography, automated sounding, and ground penetrating radar, were compared with ground truth coring results.

The study revealed significant inconsistencies in results across all methods and consultants. Manual chain drag sounding offered reliable results, followed by automated sounding with potential for improvement. Infrared thermography exhibited uniformly poor performance in detecting delamination. Limited evaluation of the emerging time-lapse infrared thermography looked promising, but further improvement is needed.

Matching researchers' expectations, the literature shows that although this study employed the most robust ground truth comparison in a field study compared with the existing literature, limitations still hinder the formation of definitive conclusions. These indefinite conclusions highlight the challenge in field studies where getting numerous ground truth cores for statistical significance is impractical. Although no technology was identified as ideal to serve as the control, this study discusses meaningful ways to improve the reliability of some promising NDE methods.

This study recommended developing a qualifying mockup testing program for NDE methods, operators, and equipment before contracts can be awarded. The study also recommended developing specifications and engineering guidance for NDE methods.

FINAL REPORT

RELIABILITY OF NONDESTRUCTIVE EVALUATION OF CONCRETE BRIDGE DECKS

Soundar S.G. Balakumaran, Ph.D., P.E.
Associate Director
Virginia Transportation Research Council

Amir Behravan, Ph.D., P.E.
Research Scientist
Virginia Transportation Research Council

INTRODUCTION

Virginia is a deicing-salt-using state. Bridges in Virginia often start to deteriorate from the top components, except in coastal areas. Being the most exposed to salt, traffic, and elements, bridge decks in noncoastal areas deteriorate the quickest. Proper inspection of concrete bridge decks is key to forming maintenance strategies and allocating appropriate repair funds to extend their service life. However, the mandatory routine inspection is mainly visual in nature. This method can detect problems evident on the surface, such as cracks, spalls, and previous patches. However, internal flaws such as delaminations and rust formation on the reinforcement cannot be perceived from visual inspections alone.

Several nondestructive evaluation (NDE) techniques have been developed and fine-tuned recently for bridge inspection purposes, resulting from improved data processing speed and the availability of affordable electronic components. State departments of transportation (DOTs) implement these techniques frequently. Among them, mature nondestructive techniques in terms of experience and research effort, such as ground penetrating radar (GPR) and infrared thermography (IRT), are expected to provide useful information regarding the condition of concrete bridge decks for maintenance needs. For this reason, the Virginia Department of Transportation (VDOT) was one of the first states to try these nondestructive technologies.

Further evolution of these NDE technologies has led to rapid-scanning versions of GPR and IRT at near-highway speeds, which cause fewer or no traffic disruptions. This reason has made the technologies popular with several state DOTs. However, the loss of data resolution is inevitable when collecting data at higher speeds, even at higher scan rates. It is challenging to balance the accuracy of information gained and the traffic disruption it causes. These rapid-scanning techniques are often expected to serve as screening tools to filter deteriorating structures for further indepth inspection. Time-lapse infrared thermography (TLIRT), a recently evolved NDE technique (Chase and Anderson, 2020), involves capturing several infrared images of a target area at predetermined intervals. The temperature data from the images are used to form a thermal inertia map, which might reveal delaminations more accurately. This method requires intermittent traffic control only during the equipment's setup and removal because the system will be mounted at the parapet to monitor the bridge deck.

If reliable, using rapid-scanning NDE techniques as screening tools and TLIRT for indepth evaluation methods can provide timely warnings about hidden deterioration in the concrete decks while saving traffic control costs and delays to the traveling public. Advanced information about impending damages is valuable in allowing VDOT to strategically form rehabilitative measures to help avoid irreversible damage to the bridges, resulting in a safe and reliable transportation infrastructure.

SHRP2 R06A Implementation Assistance Program

Under the Federal Highway Administration's (FHWA) Strategic Highway Research Program 2 (SHRP2) initiative, VDOT conducted a statewide deployment of rapid-scanning GPR and IRT. Under this initiative, three consultants scanned and analyzed 25 bridge decks in two phases (Figure 1).

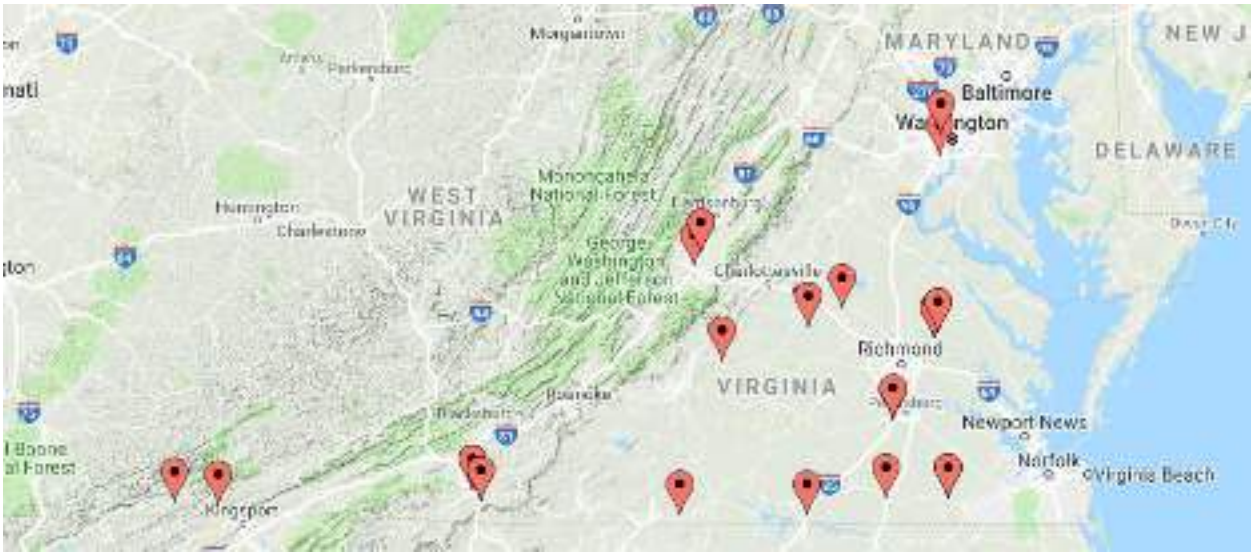


Figure 1. SecondStrategic Highway Research Program 2 R06A Implementation Project Bridge Locations

Table 1 shows the project results regarding the percentage of deteriorated bridge decks, the corresponding element conditions states, and the combined condition states 3 (poor) and 4 (severe). The rapid-scanning GPR results indicated higher damage than rapid-scanning IRT results, sometimes as much as 34 times more. This observation can be expected because GPR indicates the initiation of concrete deterioration by moisture and ion accumulation in the deck. Alternatively, IRT is expected to reveal actual delaminations in the deck. However, a comparison of NDE data and deck surfaces under combined condition states 3 and 4 from recent inspection data showed no apparent correlation. Nonetheless, no definite conclusions can be made regarding the reliability of either NDE surveys or routine inspections from these data because the SHRP2 R06A implementation study did not include any ground truth verification.

Table 1. Deck Deteriorated Area from Nondestructive Evaluation and Routine Inspection

Bridge No.	GPR Deterioration	IRT Delaminations	Total Condition States 3 and 4, % Area
1	8.4%	0.9%	0.0%
2	9.0%	1.1%	0.0%
3	9.0%	1.3%	0.0%
4	10.4%	2.7%	20.0%
5	6.3%	1.7%	0.0%
6	10.2%	6.4%	100.0%
7	30.2%	4.7%	8.9%
8	36.6%	7.9%	0.1%
9	21.1%	0.0%	0.2%
10	9.6%	0.0%	0.4%
11	1.4%	0.8%	2.2%
12	0.4%	0.5%	2.3%
13	3.1%	0.0%	0.0%
14	27.3%	4.3%	0.0%
15	31.3%	0.5%	19.4%
16	55.1%	2.2%	0.2%
17	41.2%	1.2%	0.3%
18	35.7%	2.8%	5.5%
19	10.5%	0.6%	3.7%
20	16.7%	0.6%	0.2%
21	3.0%	0.1%	0.0%
22	3.6%	0.5%	0.1%
23	7.6%	2.6%	0.1%
24	29.4%	11.4%	0.0%
25	30.0%	2.0%	1.8%

GPR = ground penetrating radar; IRT = infrared thermography.

Regarding reliability, the question remains as to how the rapid-scanning NDE techniques compare with indepth and traditional evaluation methods of concrete bridge decks. If the rapid-scanning NDE techniques offer reliability in detecting deterioration, the time and cost spent on indepth evaluations can be safely reduced. If these techniques are unreliable, the ways to improve the NDE methods should be understood.

An opportunity appeared when a VDOT District Bridge Engineer volunteered to study two bridges with a history of multiple NDE inspections conducted by multiple consultants. As a result, this project was initiated to compile and analyze the accumulated inspection results.

PURPOSE AND SCOPE

The purpose of the study is to determine the reliability of multiple modern NDE methods for detecting deterioration in concrete bridge decks. The scope includes an interstate bridge in the VDOT Staunton District on which both multiple indepth and rapid-scanning NDE evaluations were performed as part of district maintenance activities during the past 6 years. This study compares the information collected and interpreted from these evaluations.

METHODS

Task 1: Literature Review

Researchers evaluated the performance and practicality of nondestructive testing methods for bridge decks in the past few decades. Because the technologies and processing techniques have evolved significantly in recent years, this literature review was limited only to the major studies conducted since 2010.

Task 2: Collection of Raw Inspection Data

A bridge carrying Interstate 81 (I-81) southbound traffic over the Middle River (Federal ID 1784) in the Staunton District was selected for analyzing the raw inspection data based on the vast number of valuable NDE inspection records. Researchers gathered detailed inspection records spanning 6 years from multiple consultants and conducted NDE surveys and ground truth coring for comparison purposes. The I-81 northbound bridge was also under consideration. However, because the bridge had to be overlaid with rapid-setting latex-modified concrete overlay in 2021, ground truth coring could not be completed in time for comparison. Therefore, the northbound bridge was eliminated from the scope. Because the consultants used different methods to present the inspection results, an enormous effort was put into standardizing all the NDE results and projecting them on the bridge plan view.

Task 3: Data Analysis

The standardized data were compared with the ground truth coring, and statistical analysis was performed to rate the NDE technologies.

RESULTS AND DISCUSSION

Task 1: Literature Review

A study under SHRP2 compared multiple NDE methods and operating teams for evaluating bridge decks (Gucunski et al., 2012). Ten teams from academia and industry, along with FHWA's Turner Fairbanks Highway Research Center, assessed a bridge in Northern Virginia, focusing on different technologies. Researchers stated that the primary aim was not to evaluate the accuracy of NDT technologies because of the limited ground truth performed but to evaluate the ease of use, speed, cost, and repeatability of the methods used among the teams. Only eight cores were removed from the deck for comparison. Regarding delaminations, five teams used three technologies: impact echo, IRT, and chain drag sounding. Ground truth matches with the eight cores varied across the teams. Therefore, strong conclusions regarding the accuracy of the NDE methods could not be made from this study. However, some technologies were identified to be fair to good for detecting delaminations, corrosion, and concrete deterioration.

A study by Sultan and Washer (2018) looked at the reliability of GPR and IRT for evaluating bridge decks. The report focused on the difficulty of making strong conclusions based on limited ground truth coring. It employed the receiver operator characteristics method to evaluate the true positive rate of the NDE methods. The receiver operator characteristics curve is plotted between true positive rates and false positive rates and uses the area under the curve to evaluate statistical models. The study concluded that IRT indicated higher area under the curve and, thus, higher detection rates when compared with GPR for delaminations. Unfortunately, the usefulness of this receiver operator characteristics method also depends on having a minimum sample size, similar to every other evaluation method (Blume, 2009). The author implied that only limited destructive coring was conducted in this study. Therefore, strong evidence could not support the study's results.

The Indiana DOT funded a large-scale study to determine the reliability of several NDE methods, including GPR, IRT, manual chain drag sounding, automated sounding, and impact echo (Jia et al., 2022). Researchers used data from multiple bridges collected by nine consultant teams and a team from the Indiana DOT. The study concluded that some technologies like impact echo showed promise, but the results from different technologies and operators did not often agree. However, no ground truth comparison was planned for the study, so the operators' conclusions could not be supported by any degree of statistical significance.

In the evaluation of NDE technologies for bridge decks, the approach mirrors principles often applied in medical research, for which funding constraints and skepticism frequently limit full-scale randomized controlled trials. Similarly, the limited number of core samples that can be extracted without compromising the integrity of the structure constrains field validation of NDE tools. However, targeted field trials, even with minimal coring, are valuable for demonstrating the potential effectiveness of NDE methods in identifying subsurface flaws. These small-scale studies allow researchers and inspectors to refine testing procedures, build technical credibility, and generate data that can support future investment and broader adoption. Furthermore, the natural variability in concrete properties across bridge decks presents an opportunity to assess the sensitivity and repeatability of different NDE approaches, ultimately helping to optimize inspection strategies in real-world conditions (White and Ernst, 2001).

Ground Truth Sample Size

When comparing ground truth results with the NDE predictions, the results can be categorized as one of the four following outcomes:

- True positive. The case in which NDE correctly identified true flaws. The consequence of poor true positive detection is the diminished effectiveness and accuracy and, therefore, reduced reliability of the NDE method.
- False positive. The case in which NDE incorrectly identified sound concrete as flawed. The consequence of too many false positives is that DOTs will waste funds to address sound concrete, which takes away critical funds that can be allocated for truly deteriorated decks.
- True negative. The case of sound concrete that NDE correctly predicted as sound. The consequence of poor true negative detection is similar to that of too many false positives.

- False negative. The case of true flaws that NDE missed. The consequence of too many false negatives would be reduced safety for the traveling public and rapidly worsening deck areas that could have been addressed earlier in a relatively cost-effective manner. For evaluating NDE methods, the terms "sensitivity" (Equation 1) and "specificity" (Equation 2) are used.

$$Sensitivity = \frac{\text{True Positivites}}{\text{True Positivites} + \text{False Negatives}} \quad \text{Equation 1}$$

$$Specificity = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} \quad \text{Equation 2}$$

Researchers face the major challenge of determining appropriate sample sizes for comparison studies. In the context of bridge deck evaluation, sensitivity is the proportion of actual flaws that a NDE method correctly identifies, and specificity is the proportion of sound concrete that a NDE method correctly identifies as free of flaws. The higher the sensitivity and specificity, the better the detection quality. These terms need different minimum sample sizes depending on the prevalence of flaws.

For the context of bridge owners and maintenance personnel, this study considered sensitivity to be a more critical need than specificity. The rationale is that the purpose of using NDE is to catch flaws early in their lifecycle so that they do not propagate and cause safety issues to the traveling public. Specificity might be valuable information for the bridge owners to assess the effectiveness of the work performed by a consultant. Bridge owners often aim to address a larger area around the predicted flaws in sound condition to limit future maintenance needs and maintenance frequency. Such actions will reduce traffic disruptions and improve the safety of field workers. This study will primarily address the sensitivity of NDE methods.

When determining the number of ground truth cores for the statistical robustness of the evaluation, the prevalence of a specific flaw can be a significant challenge. For example, a relatively new bridge deck of less than 10 years of service would be expected to have very few flaws, about 2 to 3% by surface area. However, such a deck evaluation will need a large number of samples to confirm the sensitivity (presence of flaws). A bridge deck of more than 40 years in a corrosive environment would be expected to have a higher number of flaws, about 15 to 20% by surface area. In this case, the number of samples needed to confirm the sensitivity will be lower because it is easier to end up with cores in flawed areas of the deck. A seminal paper written by Nancy Buderer (1996) addressed this important need, and the study included the adjusted statistical analysis. Equation 3 and Equation 4 show the minimum sample size required for sensitivity and specificity, respectively, considering the prevalence of flaws.

$$n1 = \frac{Z^2_{\propto/2} \cdot \frac{SN(1-SN)}{E^2}}{P} \quad \text{Equation 3}$$

$$n2 = \frac{Z^2_{\propto/2} \cdot \frac{SP(1-SP)}{E^2}}{P} \quad \text{Equation 4}$$

Where:
n1 = sample size required for sensitivity.
n2 = sample size required for specificity.
$Z_{\alpha/2}$ = standard normal table value for a confidence level.
SN = expected sensitivity value.
SP = expected specificity value.
E = margin of error.
P = historical or estimated prevalence of flaws in a bridge deck.

Beyond comparing the sensitivity of different methods, Cohen's Kappa score, also known as the Kappa coefficient, is a statistic used to evaluate how well binary classifiers compare. The Kappa score illustrates the reliability within and across the ratings for each category. The score accounts for the chance that the agreement between two results happened by coincidence. Therefore, it is considered to be more useful to rank the different methods rather than a simple agreement on sensitivity.

Task 2: Raw Data Collection

Figure 2 shows the timeline of the various inspections and the technologies employed during those times.

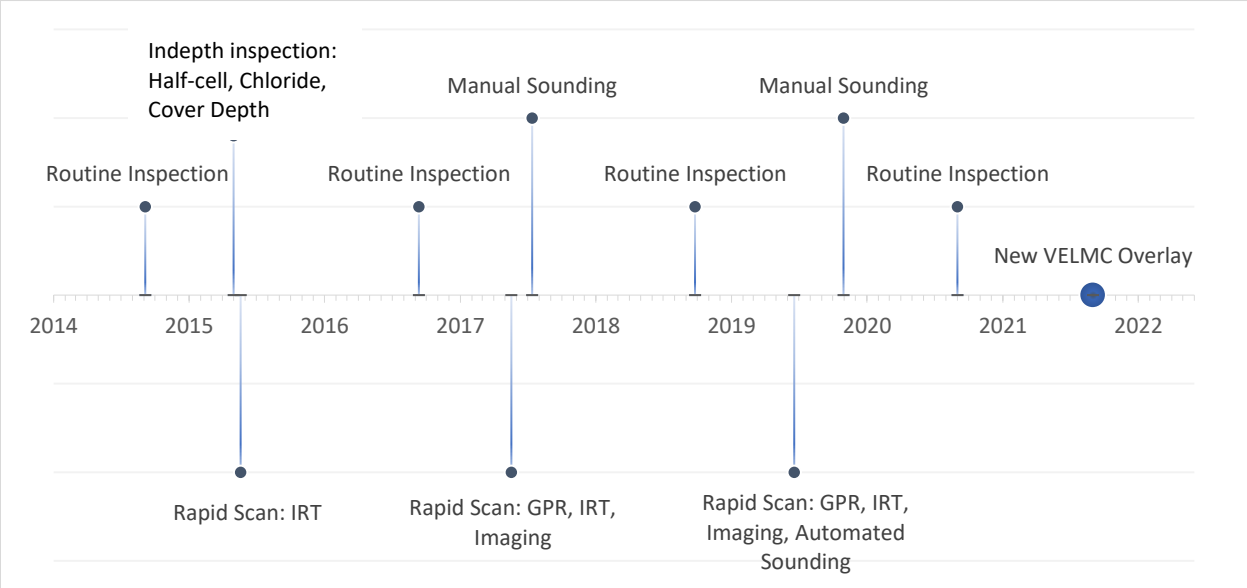


Figure 2. Timeline of Inspections on Interstate 81 Bridge. GPR = ground penetrating radar; IRT = infrared thermography.

Seven consultants conducted surveys on this bridge for 6 years. Table 2 provides the details.

Table 2. Consultants and Survey Details Between 2015 and 2021

Consultant ID	Survey Details	Notes	Mode of Survey	Year
Consultant A	Visual, Cover Depth, Half-cell potentials, Chloride Content, Coring, Petrography, Compressive Strength	Sounding was implied but not conducted	Manual survey with traffic control	2015
Consultant B	Optical Imaging, Infrared Thermography	Demonstration of technology; no ground truth	Vehicle-mounted rapid-scanning	2015
Consultant C	Optical Imaging, Ground Penetrating Radar	Demonstration of technology; only partially post-processed results shared with VDOT; not used for analysis in this research	Vehicle-mounted rapid-scanning	2015
Consultant D	Optical Imaging, Ground Penetrating Radar, Infrared Thermography	--	Vehicle-mounted rapid-scanning	2017
Consultant E	Chain Drag Sounding, Visual	--	Manual survey with traffic control	2017
Consultant F	Optical Imaging, Ground Penetrating Radar, Infrared Thermography, Automated Sounding	--	Vehicle-mounted rapid-scanning except	2019
Consultant G	Manual Chain Drag Sounding, Visual	--	Manual survey with traffic control	2019
Consultant H	Time-Lapse Infrared Thermography	Demonstration of technology; post-processed results shared with VDOT; limited used for analysis in this research	Barrier-mounted camera inspection	2019

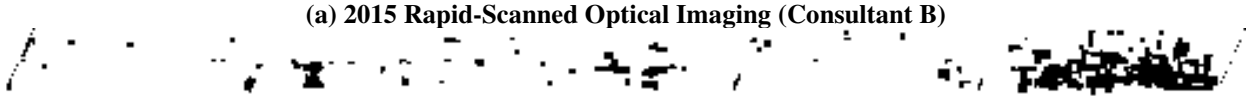
Task 3: Data Analysis

Recent NDE data are more appropriate for a reasonable comparison. For this reason, the two bridge decks were re-scanned using rapid-scanning NDE methods through contract work.

Patches

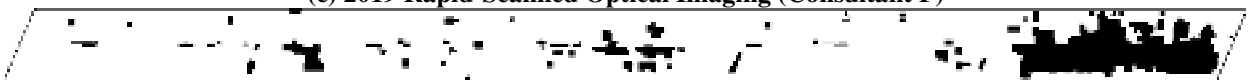
Patches on a bridge deck indicate the defects identified and repaired during its service life. These patches are easily identified visually from the surface-color difference of the relatively newer concrete opposed to the older original deck concrete. Figure 3 shows four sets of data on the patches for the I-81 southbound bridge deck, available for 3 separate years.

(a) 2015 Rapid-Scanned Optical Imaging (Consultant B)



(b) 2017 Manual Visual Survey (Consultant E)

(c) 2019 Rapid-Scanned Optical Imaging (Consultant F)



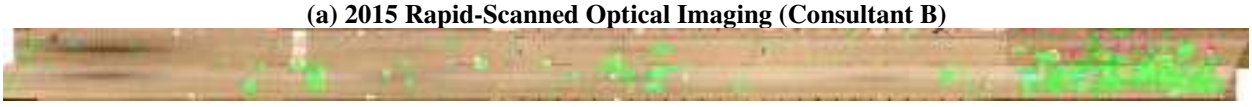
(d) 2019 Manual Visual Survey (Consultant G)

Figure 3. Deck Patches as Reported by Consultants

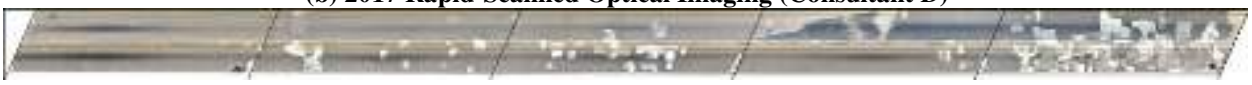
The results from Consultants B, E, and F (Figure 3a, Figure 3b, and Figure 3c, respectively)—one hand drawn manually and two digital images—clearly show similar patches during 4 years (2015 to 2019). A few more patches were added to the bridge deck over the years. However, the 2019 manual chain drag survey by Consultant G (Figure 3d) shows a different case for patches. Although a few similarities can be seen in the overall location of patches, the patches have different shapes, densities, and sizes. Because it was a manual survey, no photographs were taken to verify the patches. However, the 2019 survey by Consultant G (
Figure 4d) during a rapid-screening NDE survey helped compare the patches' ground truth.

Figure 4 shows the digital images captured during 2015, 2017, and 2019. The patches appear similar to the first three maps (Figure 3a, Figure 3b, and Figure 3c). The fourth map is a gross deviation from the actual bridge deck condition.

(a) 2015 Rapid-Scanned Optical Imaging (Consultant B)



(b) 2017 Rapid-Scanned Optical Imaging (Consultant D)



(c) 2019 Rapid-Scanned Optical Imaging (Consultant F)

Figure 4. Optical Imaging of the Decks by Consultants

The difference in the clarity of digital images is evident. The time of day and the availability of sunlight during the surveys significantly affect image clarity. However, gathering useful visual information from the surface with various image processing techniques is still possible.

In the image Consultant D captured (
Figure 4b), flaws appear on the passing lane in the fourth span from the north. This appearance could be due to a disturbance in the vehicle ride that might have caused the blur. Repetition of imaging surveys will be necessary in such a case. In contrast, a drone-mounted camera could avoid such issues.

Spalls

Very few spalls appeared on the bridge deck (Figure 5). This observation was expected because VDOT districts quickly address such spalls to avoid any safety issues affecting traffic in these areas. For this reason, spalls will not be discussed further in the context of this study.

Figure 5. Manually Drawn Spall Map by Consultant G. The dots in this figure correspond to the size and location of the spalls and appear small for this reason.

Ground Truth

Before the initiation of this study, the researchers assumed some degree of agreement between traditional and modern NDE methods based on experience in field structures. The researchers expected that manual chain drag sounding, the most commonly employed NDE method for the longest period of time, could serve as the control or the "gold standard" compared with the modern NDE methods. However, some significant concerns from the manual chain drag sounding results from two consultants eliminated this method from being considered the control. Therefore, this study unintendedly became an uncontrolled experiment. As the name suggests, no control method can be identified in an uncontrolled study. As a result, the researchers decided that the ground truth coring would serve as the control, although this method could only provide accurate condition data at individual spots, unlike sounding results.

In addition to the limitation of cores used as ground truth imposed by sampling size, the following scenario is noted. Although a bridge deck core used as ground truth might indicate an impending delamination interface, no NDE method, including manual sounding, may capture that early condition if separation has not occurred.

A statistical experiment design was performed to create a starting point to decide the number of ground truth cores. The initial assumptions are as follows:

- Expected prevalence of flaws: 20%.
  - Given the advanced age of the bridge deck at 53 years (as of 2020) and the total quantity of delaminations, patches, and spalls predicted by the manual sounding in 2017 as 22.7% and by another manual sounding in 2019 as 10.1% (with questionable patch mapping), a rough estimate of 20% expected delamination was assumed.
- 95% confidence level.
- Target sensitivity (true positive rate) = 90%.
- Margin of error = ±5%.

This calculation can be inferred as the minimum number of samples needed to predict 85 to 95% of the true delaminations with a 95% confidence level, assuming 20% of expected delaminations in the deck. This estimation resulted in a minimum sample size of 692, which is neither practical nor feasible for a bridge deck in service.

Therefore, the assumptions were relaxed to the other extreme to check for a sample size that would enable the least statistically significant results that could yet result in meaningful conclusions. The following percentages were the revised assumptions:

- Same expected prevalence of flaws: 20%.
- Reduced to 90% confidence level.
- Reduced target sensitivity (true positive rate) = 75%.
- Reduced margin of error = ±24%.

This new assumption can be inferred as the minimum number of samples needed to predict 51 to 99% of the true delaminations with a 90% confidence level, assuming 20% of expected delaminations in the deck. This minimum sample size was 45, or 9 cores per span.

During multiple field trips, researchers gathered 44 cores on this bridge by strategically picking locations to assess the performance of the multiple NDE results from consultants. After removing the cores, 10 locations out of 44 were confirmed to have delaminations by checking the cored wall of the deck concrete.

Delaminations

This study included two sets of manual chain drag sounding, three sets of IRT scans, and one set of automated sounding scans in relation to delaminations. Figure 6 shows the consolidated delamination maps from six combinations of NDE methods and consultant teams.

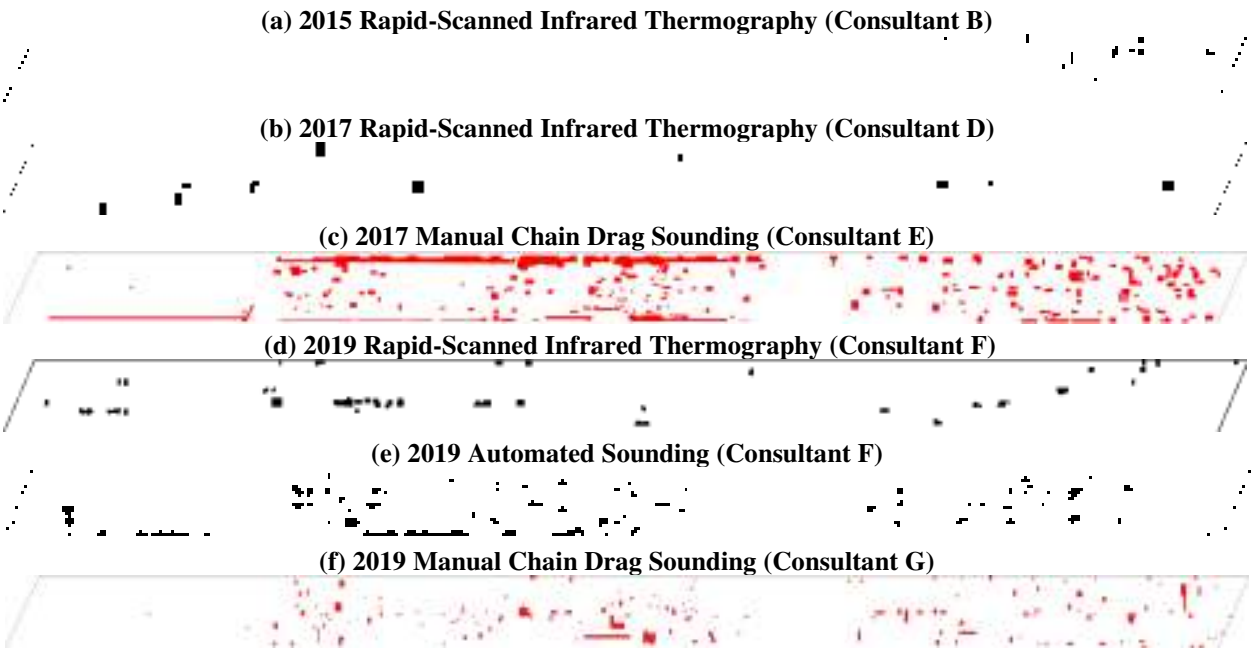


Figure 6. Delamination Survey Results from Consultants

Visually, all six images appear significantly different. Figure 6 shows the difficulty identifying subsurface delaminations consistently across inspection teams and technologies.

Delaminations cannot decrease over time without repairs that would be evident as patches. On Figure 6c showing the 2017 manual sounding results from Consultant E, long, narrow delaminations appear primarily near the curb lines. During a field check, the researcher and the District Bridge Engineer observed that these delaminations were not continuously delaminated by tapping with a hammer. Some of those areas turned out to be fascia girder lines. The sound waves' frequency differed from other sound concrete areas, so a well-trained inspector could have found the difference. In addition, given the noticeably poor results on deck patches by Consultant G (Figure 3d), it raised concerns about the validity of their delamination results. Because the patches did not increase much over time, and from the observations previously discussed, the first survey in 2017 by Consultant E (Figure 6c) likely was slightly more conservative, and the survey in 2019 (Figure 6f) did not properly transfer data from the field to the computer-aided design, or CAD, drawings.

Table 3 compares the delamination quantities reported by all the consultants and technologies. It is disappointing that only marginal overlaps between the technologies were found. Even two technologies employed by the same Consultant F in 2019 showed only a 1.53% overlap in the delamination locations. The highest overlap was found between the manual chain drag sounding in 2017 and the automated sounding in 2019 at 5.69%, which is very low compared with the expectations.

Table 3. Comparison of Technologies and Consultant Teams for Delaminations

Year	**→**		**2015**	**2017**	**2017**	**2019**	**2019**
↓	**Consultant and Technology**	**Total Reported Delamination Area as % of Total Surface Area**	**B-IRT**	**D-IRT**	**E-Sounding**	**F-IRT**	**F-Automated Sounding**
2015	B–IRT	0.28%					
2017	D-IRT	0.92%	0%				
2017	E-Sounding	11.64%	0.81%	0.70%			
2019	F-IRT	1.57%	1.25%	1.54%	1.27%		
2019	F-Automated Sounding	3.23%	0.89%	1.00%	5.69%	1.53%	
2019	G-Sounding	3.03%	0.07%	0.12%	3.77%	1.75%	2.05%

IRT = infrared thermography.

The error of locating and transferring from data collection to a CAD format can add up. Because VDOT would address the areas surrounding the predicted flaws to reduce the propagation of deterioration and to reduce frequent future maintenance needs, a sensitivity comparison was also calculated using a 12-inch buffer from the defect areas for practical purposes (Table 4).

Table 4. Comparison of Reported Delamination Against Ground Truth Cores

Consultant ID	Survey Details	Total Reported Delamination Area as % of Total Surface Area	Ground Truth Exact Match		Ground Truth Approx. Match (12-inch Buffer)		Sensitivity –Clean Match	Sensitivity –Approx. Match (12-inch Buffer)
			True Positive	False Positive	True Positive	False Positive		
			False Negative	True Negative	False Negative	True Negative		
Consultant B	Infrared Thermography	0.28%	0	1	0	3	0%	0%
			10	33	10	31		
Consultant D	Infrared Thermography	0.92%	1	0	2	1	10%	20%
			9	34	8	33		
Consultant E	Chain Drag Sounding	11.64%	4	4	7	6	40%	70%
			6	30	3	28		
Consultant F	Automated Sounding	3.23%	3	3	5	5	30%	50%
			7	31	5	29		
	Infrared Thermography	1.57%	2	5	2	5	20%	20%
			8	29	8	29		
Consultant G	Chain Drag Sounding	3.03%	1	2	1	10	10%	10%
			9	32	9	24		

Table 5 shows the Kappa coefficient ratings for the corresponding results. The higher the coefficient, the higher the agreement with the ground truth. The overall interpretation did not change for the coefficients for the methods and consultants, with higher ratings between the exact and approximate matches. Overall, manual chain drag sounding ranked at the top and bottom among the combination of methods and consultants.

Table 5. Kappa Coefficient Ratings for Nondestructive Evaluation Methods

Consultant ID	Survey Details	Cohen's Kappa Coefficient—Exact Match	Cohen's Kappa Coefficient—Approximate Match (12-inch Buffer)	Interpretation
Consultant B	Infrared Thermography	– 4%	– 12%	Less than chance agreement
Consultant D	Infrared Thermography	15%	23%	Fair agreement
Consultant E	Chain Drag Sounding	30%	47%	Moderate agreement
Consultant F	Automated Sounding	25%	35%	Fair agreement
	Infrared Thermography	6%	6%	Slight agreement
Consultant G	Chain Drag Sounding	5%	– 19%	Less than chance agreement

Time-Lapse Infrared Thermography Survey

Consultant H conducted a limited TLIRT survey covering spans 3 and 4 from the south, and the Virginia Transportation Research Council (VTRC) conducted an even more limited survey covering partial span 3 for comparison.

As Table 6 shows, this delamination prediction is much higher than other NDE methods, nearly double the second highest reported delaminations by Consultant E (Table 4). Because of the nature of the data provided to VTRC from this demonstration, it was impossible to compare

pixel by pixel, so only the ground truth values were compared. Because only 21 ground truth locations coincided with the TLIRT results, the results are not considered conclusive.

Table 6. Comparison of Time-Lapse Infrared Thermography Reported Delamination Against Ground Truth Cores

Consultant ID	Survey Details	Total Reported Delamination Area as % of Total Surface Area of Spans 3 and 4 from South	Ground Truth Exact Match		Ground Truth Approx. Match (12 inches)		Sensitivity—Clean Match	Sensitivity—Approx. Match (12-inch Buffer)
			True Positive	False Positive	True Positive	False Positive		
			False Negative	True Negative	False Negative	True Negative		
Consultant H	Time-Lapse Infrared Thermography	22.2%	3	4	3	4	60%	60%
			2	12	2	12		

On span 3 from the south, a quality check was performed using another TLIRT data collection set, and the results varied with respect to the ground truth comparison. Four ground truth locations were at the intersection of both TLIRT maps. Two out of four of Consultant H's TLIRT results matched the ground truth, and three out of four of VTRC's results matched.

Cover Depths

Appropriate concrete cover depth is an important factor that significantly contributes to the increase in the service life of bridge decks. Traditionally, manual surveying using a magnetic pachometer is conducted, and recently, GPR has been used to establish a cover depth map for a newly built or rehabilitated bridge deck in Virginia.

Table 7 shows the summary of the cover depth survey results from three consultants. Two ground truth cores were used to compare the results. Because Consultant A obtained cover depths at a grid of 5 by 10 feet, the ground truth locations did not exactly match but were within 2 feet of the cover depths. Despite that mismatch, the magnetic pachometer result matched reasonably well.

Table 7. Cover Depth Results and Ground Truth from Southbound Bridge

Consultant ID	Survey Details	Ground Truth Comparison		Summary Statistics for Entire Deck Provided by Consultant			
		Actual (Core)	Consultant Report				
Consultant A	Magnetic Cover Meter	2.25 in.	2.55 in. (closest location)	< 1 in.	1 to 1.99 in.	2 to 2.99 in.	3+ in.
		2.625 in.	2.75 in. (closest location)	0%	0.6%	21%	78.4%
Consultant D	Ground Penetrating Radar	2.25 in.	1.2 to 1.3 in.	Avg.	St. Dev.	Min.	Max.
		2.625 in.	1.8 to 1.9 in.	2 in.	0.2 in.	1.2 in.	2.7 in.
Consultant F	Ground Penetrating Radar	2.25 in.	2.25 to 2.5 in.	Avg: 3.5 inches			
		2.625 in.	2.75 to 3.0 in.				

From the two GPR scans, Consultant D's results were far from reality. However, Consultant F's GPR scans produced particularly good results that were close to the ground truth. The summary statistics for the whole deck show that Consultant D's scans indicated more cover depths than the actual values. It is very likely that the calibration of the dielectric constant for the concrete cover was not performed properly.

Corrosion Deterioration

Corrosion deterioration is a difficult factor to validate because it is a probabilistic process, and given the extent of corrosion, visible rust may not always appear in all locations. Traditionally, half-cell potentials are used to estimate the probability of corrosion. Increasingly, GPR is considered a possible indicator of future corrosion because GPR is sensitive to the presence of moisture and likely chloride ions.

Half-cell potentials are measured at discrete points, and GPR scans are essentially line scans interpolated into an area map. Figure 7 shows the interpolated half-cell potential map from a 5-x-10-foot grid. Figure 8 shows the GPR results interpolated from line scans taken 3 feet apart.

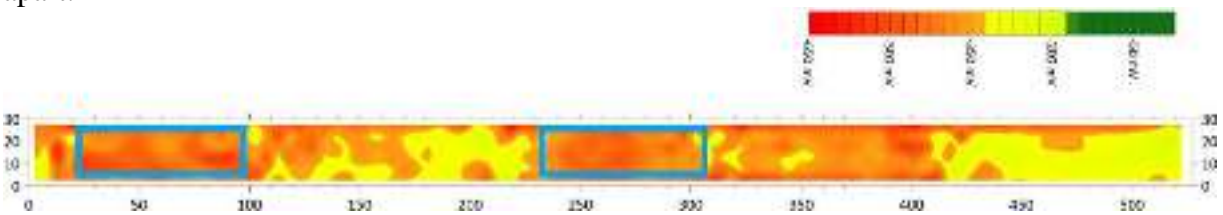
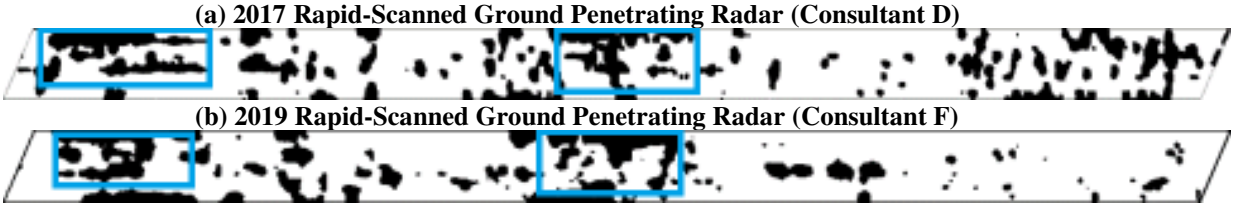


Figure 7. Half-Cell Potentials Results from Consultant A (2015)

(a) 2017 Rapid-Scanned Ground Penetrating Radar (Consultant D)



(b) 2019 Rapid-Scanned Ground Penetrating Radar (Consultant F)

Figure 8. Ground Penetrating Radar Deterioration Map Results

Consultant D's GPR deterioration map indicated a 25.2% area of potential problems. Consultant F's GPR deterioration showed 20.15% after 2 years, and the overlap was 8.05%. Figure 7 and Figure 8 show that some agreement is present on the far left span (north) and the middle span. However, none of the results between the two spans on the right side (south) of the deck showed any agreement. Because of the inconsistency, it is unclear if actionable information can be gathered from this corrosion deterioration mapping for maintenance planning.

CONCLUSIONS

- *NDE methods, both manual and modern, employed by the consultant teams reveal significant inconsistencies in their results, with little agreement.* Among the four consultants employing

modern technologies and between the two consultants employing manual methods, no agreement was found.

- *Delamination surveys produced results of the lowest agreement among all surveys.* The delamination locations identified had the worst agreement among six teams of operators, using three methods (manual sounding, IRT, and automated sounding) from five consultant firms.

- *Cover depth results varied significantly among the consultants.* One of the two consultants using GPR produced highly erroneous cover depth values across the board. In contrast, the other GPR scan was much closer to the magnetic pachometer readings and ground truth cores. These findings indicate the need for proper calibration of the GPR raw data.

- *Currently, manual chain drag sounding has the highest potential to provide high-quality results.* Two consultant teams produced results that ranked the highest and lowest compared with the ground truth. However, the consultant team that produced the worst results had obvious, avoidable problems that could be filtered out through simple quality checks.

- *Rapid-scanning IRT demonstrates limited effectiveness in detecting delaminations, as evidenced by consistently poor results across evaluations conducted by three consultant teams.*

- *TLIRT results look encouraging compared with rapid-scanning IRT.* However, advanced inspection methods are expected to be more accurate.

- *Automated sounding has the potential to catch up with manual sounding and could improve the detection of delaminations.*

- *Corrosion deterioration mapping with half-cell potentials and GPR exhibits marginal agreement.*

- *Despite employing a more robust ground truth evaluation with a reasonably larger sample size than existing literature, limitations in terms of statistical significance continue to hinder the identification of a definitive control method and the establishment of statistically robust conclusions regarding the performance of the evaluated technologies.*

RECOMMENDATIONS

1. *VDOT Structure and Bridge Division and VTRC should utilize sections from demolished decks or design and develop a mockup with field-realistic flaws for qualification of technologies and operators, with the consensus of VDOT bridge end users, before awarding contracts.* Considering the rapid evolution of technologies, it is important to filter suitable technologies based on the need and efficacy. Achieving statistical significance on bridge decks in service is nearly impossible because of the high demand for sample sizes. Therefore, qualification through mockup testing is a reliable and practical alternative that has worked well for VDOT on acoustic monitoring and complex concrete casting projects.

2. *VDOT Structure and Bridge Division and VTRC should develop specifications and engineering guidance for NDE inspection methods.*

IMPLEMENTATION AND BENEFITS

Researchers and the technical review panel (listed in the Acknowledgments) for the project collaborate to craft a plan to implement the study recommendations and to determine the benefits of doing so. This process is to ensure that the implementation plan is developed and approved with the participation and support of those involved with VDOT operations. The implementation plan and the accompanying benefits are provided here.

Implementation

With regard to Recommendation 1, the VDOT Structure and Bridge Division will look for upcoming deck replacement projects to identify appropriate demolished deck sections to be transported to a VDOT area headquarters yard for storage within 2 years from this report's publication date. If appropriate deck sections are unavailable, VTRC will investigate developing a mockup slab with field-realistic flaws for qualifying consultants and technologies. This investigation will involve VTRC conducting a research study after discussion and ranking in the Bridge Research Advisory Committee.

With regard to Recommendation 2, the VDOT Structure and Bridge Division will work with VTRC to develop specifications and engineering guidance documents as a part of the Chapter 32—Bridge Maintenance Manual within 2 years from the date of publication.

Benefits

As per Recommendations 1 and 2, the reliability of the inspection methods, the inspecting team, and the equipment will be judged before a contract can be awarded, reducing the chances of poor results.

Early and accurate detection of bridge problems is crucial for preventing significant issues that require expensive repairs or even bridge closures. If a technology misses critical damage to a bridge, the oversight could lead to more extensive and costly repairs. Although inspection costs for several modern methods are decreasing because of the ease of deployment and partial automation of the post-processing methods, given the inconsistency in detecting flaws in bridge decks, this study encourages further improvement in detection technologies.

Identifying the advantages and disadvantages of the current technologies for bridge deck inspection can lead to the allocation of bridge maintenance budgets directed toward more reliable methods. Other approaches to bridge deck inspection include identifying the accuracy and reliability of each method, further research and development to improve the technology, or lastly, finding alternative solutions. This process can ultimately lead to more effective and cost-efficient bridge inspection methods in the future.

ACKNOWLEDGMENTS

The researchers are grateful to the champions, Rex Pearce, District Bridge Engineer, Staunton District, and Adam Matteo, Assistant State Structure and Bridge Engineer for Maintenance, Central Office Structure and Bridge Division, for their support. Specifically, this research would not have been successful without the Staunton District bridge deck, which has a long history of multiple inspections, contributed by Rex Pearce for research purposes. The researchers are also grateful to the previous champion, now retired, Jeff Milton, former Bridge Preservation Program Manager, for his tremendous help and expertise in Bridge Engineering. The researchers are grateful for the guidance regarding the statistical methods from Justice Appiah, Associate Principal Research Scientist, and Jim Gillespie, Senior Research Scientist.

REFERENCES

Blume, J.D. Bounding Sample Size Projections for the Area Under a ROC Curve. *Journal of Statistical Planning and Inference*, Vol. 139, No. 1, 2009, pp. 711–721. https://doi.org/10.1016/j.jspi.2007.09.015.

Buderer, N.M. Statistical Methodology: I. Incorporating the Prevalence of Disease into the Sample Size Calculation for Sensitivity and Specificity. *Academic Emergency Medicine: Official Journal of the Society for Academic Emergency Medicine*, Vol. 3, No. 9, 1996, pp. 895–900.

Chase, S.B., and Anderson, C.M. *Time-Lapse Infrared Thermography Applied to Concrete Bridge Deck Inspection Surveys*. VTRC 20-R22. Virginia Transportation Research Council, Charlottesville, VA, 2020.

Gucunski, N., Imani, A., Romero, F., Nazarian, S., Yuan, D., Wiggenhauser, H., Shokouhi, P., Taffe, A., and Kutrubes, D. *Nondestructive Testing to Identify Concrete Bridge Deck Deterioration*. Transportation Research Board, Washington, DC, 2012.

Jia, Y., Williams, C.S., Baah, P., and Bowman, M.D. *Long-Term Project and Network-Level NDT Implementation Plan for Indiana*. Joint Transportation Research Program Publication FHWA/IN/JTRP-2022/31. Purdue University, West Lafayette, IN, 2022. https://doi.org/10.5703/1288284317582.

Sultan, A.A., and Washer, G.A. Comparison of Two Nondestructive Evaluation Technologies for the Condition Assessment of Bridge Decks. *Transportation Research Record*, Vol. 2672, No. 41, 2018, pp. 113–122.

White, A., and Ernst, E. The Case for Uncontrolled Clinical Trials: A Starting Point for the Evidence Base for CAM. *Complementary Therapies in Medicine*, Vol. 9, No. 2, 2001, pp. 111–116. https://doi.org/10.1054/ctim.2001.0441.