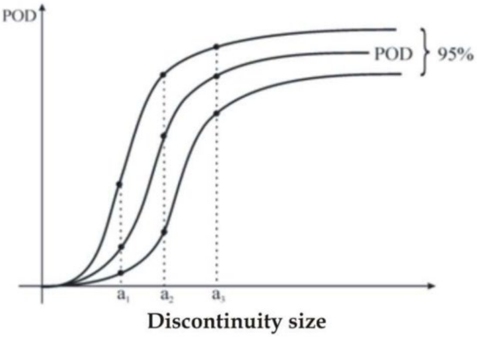
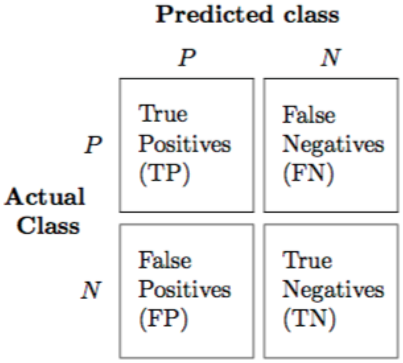
U.S. Department of
Transportation

**Federal Railroad
Administration**

Methods for Evaluation of Track Inspection Technology Effectiveness

Office of Research,
Development
and Technology
Washington, DC 20590





		Predicted class	
		P	N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

DOT/FRA/ORD-20/03

Final Report
February 2020

NOTICE

This document is disseminated under the sponsorship of the Department of Transportation in the interest of information exchange. The United States Government assumes no liability for its contents or use thereof. Any opinions, findings and conclusions, or recommendations expressed in this material do not necessarily reflect the views or policies of the United States Government, nor does mention of trade names, commercial products, or organizations imply endorsement by the United States Government. The United States Government assumes no liability for the content or use of the material contained in this document.

NOTICE

The United States Government does not endorse products or manufacturers. Trade or manufacturers’ names appear herein solely because they are considered essential to the objective of this report.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE February 2020	3. REPORT TYPE AND DATES COVERED Technical Report

4. TITLE AND SUBTITLE
Methods for Evaluation of Track Inspection Technology Effectiveness

5. FUNDING NUMBERS

6. AUTHOR(S)
Radim Bruzek, Sajjad Maymand, Gaylen Drape, Katrina Smart, John Tunna

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
ENSCO, Inc.
5400 Port Royal Road,
Springfield, VA 22151

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Department of Transportation
Federal Railroad Administration
Office of Railroad Policy and Development
Office of Research, Development and Technology
Washington, DC 20590

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
DOT/FRA/ORD-20/03

11. SUPPLEMENTARY NOTES
COR: Jay Baillargeon

12a. DISTRIBUTION/AVAILABILITY STATEMENT
This document is available to the public through the FRA website.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

Guidelines, including recommendations for sample size, are needed in the design of track inspection technology evaluation tests. This report provides and overview of recommended methods for evaluating the effectiveness of different types of track inspection systems. These methods produce performance measures that can be used to validate the technology against ground truth, which is often difficult to obtain, and to compare technologies. As sufficient test samples are needed to make valid comparisons, sample size is also addressed by, for example, grouping flaws of similar size into bins and by combining repeat inspections. The report describes each of the evaluation methods and demonstrates their application to the evaluation of a track geometry measurement system (TGMS). The relatively new Model-Assisted Probability of Detection (MAPOD) technique, which combines actual results with those from computer models of the inspection technology, could usefully be demonstrated on one or more track inspection technologies as part of future efforts.

14. SUBJECT TERMS
Correlation, receiver operating characteristics, probability of detection, repeatability, reproducibility

15. NUMBER OF PAGES
54

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT

METRIC/ENGLISH CONVERSION FACTORS

ENGLISH TO METRIC

LENGTH (APPROXIMATE)

1 inch (in)	=	2.5 centimeters (cm)
1 foot (ft)	=	30 centimeters (cm)
1 yard (yd)	=	0.9 meter (m)
1 mile (mi)	=	1.6 kilometers (km)

AREA (APPROXIMATE)

1 square inch (sq in, in^2)	=	6.5 square centimeters (cm^2)
1 square foot (sq ft, ft^2)	=	0.09 square meter (m^2)
1 square yard (sq yd, yd^2)	=	0.8 square meter (m^2)
1 square mile (sq mi, mi^2)	=	2.6 square kilometers (km^2)
1 acre = 0.4 hectare (he)	=	4,000 square meters (m^2)

MASS - WEIGHT (APPROXIMATE)

1 ounce (oz)	=	28 grams (gm)
1 pound (lb)	=	0.45 kilogram (kg)
1 short ton = 2,000 pounds (lb)	=	0.9 tonne (t)

VOLUME (APPROXIMATE)

1 teaspoon (tsp)	=	5 milliliters (ml)
1 tablespoon (tbsp)	=	15 milliliters (ml)
1 fluid ounce (fl oz)	=	30 milliliters (ml)
1 cup (c)	=	0.24 liter (l)
1 pint (pt)	=	0.47 liter (l)
1 quart (qt)	=	0.96 liter (l)
1 gallon (gal)	=	3.8 liters (l)
1 cubic foot (cu ft, ft^3)	=	0.03 cubic meter (m^3)
1 cubic yard (cu yd, yd^3)	=	0.76 cubic meter (m^3)

TEMPERATURE (EXACT)

[(x-32)(5/9)] °F = y °C

METRIC TO ENGLISH

LENGTH (APPROXIMATE)

1 millimeter (mm)	=	0.04 inch (in)
1 centimeter (cm)	=	0.4 inch (in)
1 meter (m)	=	3.3 feet (ft)
1 meter (m)	=	1.1 yards (yd)
1 kilometer (km)	=	0.6 mile (mi)

AREA (APPROXIMATE)

1 square centimeter (cm^2)	=	0.16 square inch (sq in, in^2)
1 square meter (m^2)	=	1.2 square yards (sq yd, yd^2)
1 square kilometer (km^2)	=	0.4 square mile (sq mi, mi^2)
10,000 square meters (m^2)	=	1 hectare (ha) = 2.5 acres

MASS - WEIGHT (APPROXIMATE)

1 gram (gm)	=	0.036 ounce (oz)
1 kilogram (kg)	=	2.2 pounds (lb)
1 tonne (t)	=	1,000 kilograms (kg)
	=	1.1 short tons

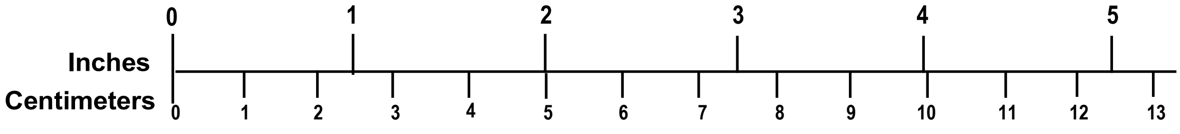
VOLUME (APPROXIMATE)

1 milliliter (ml)	=	0.03 fluid ounce (fl oz)
1 liter (l)	=	2.1 pints (pt)
1 liter (l)	=	1.06 quarts (qt)
1 liter (l)	=	0.26 gallon (gal)
1 cubic meter (m^3)	=	36 cubic feet (cu ft, ft^3)
1 cubic meter (m^3)	=	1.3 cubic yards (cu yd, yd^3)

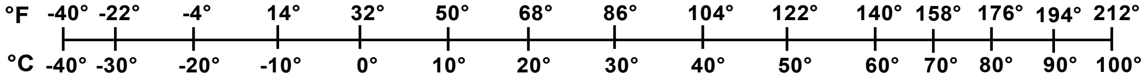
TEMPERATURE (EXACT)

[(9/5) y + 32] °C = x °F

QUICK INCH - CENTIMETER LENGTH CONVERSION



QUICK FAHRENHEIT - CELSIUS TEMPERATURE CONVERSIO



For more exact and or other conversion factors, see NIST Miscellaneous Publication 286, Units of Weights and Measures. Price $2.50 SD Catalog No. C13 10286

Updated 6/17/98

Contents

Illustrations

Tables

Executive Summary

This study, conducted by ENSCO, Inc. from July 2018 to May 2019, describes several well-documented statistical methods for evaluating the effectiveness of track inspection technologies. The methods can be used to measure the performance of new technologies and make comparisons with current track inspection methods. This process will help the Federal Railroad Administration determine if the new technologies can be used to supplement or replace established methods of track inspection.

The evaluation methods include:

1. Correlation – to measure the strength of the relationship between two sets of measurements.
2. Receiver Operating Characteristic – to compare the rate of correctly detecting defects to the rate of false alarms.
3. Probability of Detection – to measure effectiveness at different flaw sizes.
4. Repeatability – to compare measurements made under similar conditions.
5. Reproducibility – to compare measurements made under different conditions.

The choice of method depends on the objective of the evaluation, available test sample size, type of data collected by the inspection technology, and the availability of ground truth. Ground truth is the actual condition that has been ascertained by direct observation. It is often difficult and costly to establish.

The performance measures from the evaluation methods include variance, coefficient of correlation, accuracy, sensitivity, and the size of defect that can be detected with a specified probability.

This report demonstrates that sample size has a significant effect on the conclusions that can be drawn from evaluating track inspection effectiveness. If tests are performed on a few sample flaws,[1] there will be low confidence in the measured performance. As such, it will not be possible to compare different technologies. Current best practice is to have at least 30 test samples for each size of flaw for probability of detection analysis. Sample size can be increased by grouping flaws of similar size into bins as well as by combining repeat inspections.

This report recommends development of guidelines for the design of track inspection technology evaluation tests. These guidelines should include recommendations for sample size and the number of repeat inspections. Standards need to be set for acceptable track inspection performance. These could be based on the results of testing existing, so-called “gold standard” inspection methods.

A new and promising method of deriving probability of detection results is the Model-Assisted Probability of Detection (MAPOD) technique, which combines actual results with those from computer models of the inspection technology. MAPOD could be used to derive results for defects at the safety limit from tests performed with smaller flaws. This report gives examples of

[1] A flaw is defined as any type of discontinuity that must be investigated to see if it should be rejected.

MAPOD used in other industries and recommends the approach be demonstrated on one or more track inspection systems.

1. Introduction

The Federal Railroad Administration (FRA) contracted with ENSCO, Inc., to perform the work reported here.

1.1 Background

FRA safety standards require track to be inspected by various established manual and technological methods. New inspection technologies are being developed by the railroad industry that aim to improve quality and cost-effectiveness. This raises questions about the reliability and accuracy of these new inspection technologies and how well they compare to the established ones. Currently, there is no formal approach to evaluating the effectiveness of new track inspection technologies. There are no standards for acceptable reliability and accuracy of these technologies.

1.2 Objectives

The objective of this report is to provide FRA with a practical approach to quantifying the effectiveness of existing and emerging track inspection technologies. The approach is to be demonstrated on a TGMS.

1.3 Overall Approach

ENSCO, Inc., began this work by identifying the various evaluation methods and performance measures in common usage. Recommendations were then made on which method to use for each type of track inspection technology.

1.4 Scope

The work reported here only considers track inspection technologies. However, many of the methods described are applicable to other mechanical and electrical systems.

1.5 Organization of the Report

Section 2 of this report describes several commonly used evaluation methods and performance measures. Section 3 discusses the different categories of data measured by track inspection technologies. Section 4 recommends the method to be used to evaluate the effectiveness of the different categories of track inspection data. Section 5 gives an example of applying the methods to evaluate the effectiveness of a TGMS. Conclusions and recommendations for further work are given in Section 6.

The Appendix lists track inspection requirements from FRA regulations, the category of inspection, and known current technology. The Glossary gives definitions of the technical terms used in this report.

2. Evaluation Methods and Performance Measures

The following sub-sections describe the various evaluation methods and performance measures commonly used to evaluate the effectiveness of technology [1]. The last sub-section discusses the effect of sample size on the results. In this report, "flaw" is used to mean any deviation from a perfect state and "defect" is used to mean a flaw that exceeds some threshold size.

2.1 Correlation

Correlation analysis is used to measure of the strength of the relationship between two variables [2].

Figure 1 shows an example of two variables. One variable, *X*, is a track profile measured with surveying equipment (referred to as the ground truth). The other, *Y*, is a track profile measured by a geometry inspection car at 20 mph. Each variable is measured every foot, and there are 500 feet of measured data.

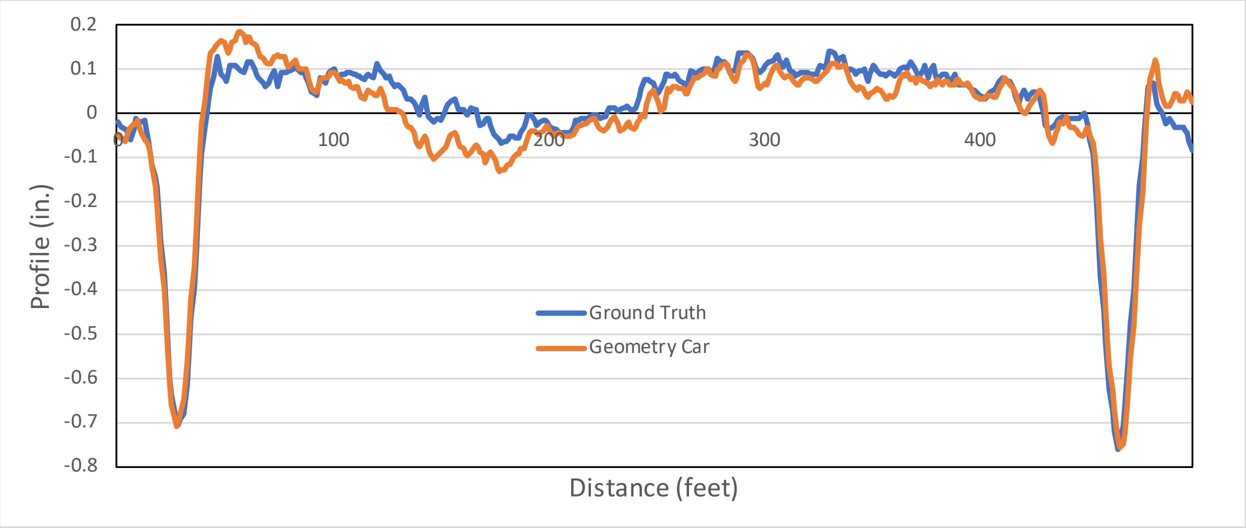


Figure 1. Track Profile Measurements

A simple way to visualize the relationship between the two variables is to plot one against the other. Figure 2 shows the result of doing so for the data in Figure 1.

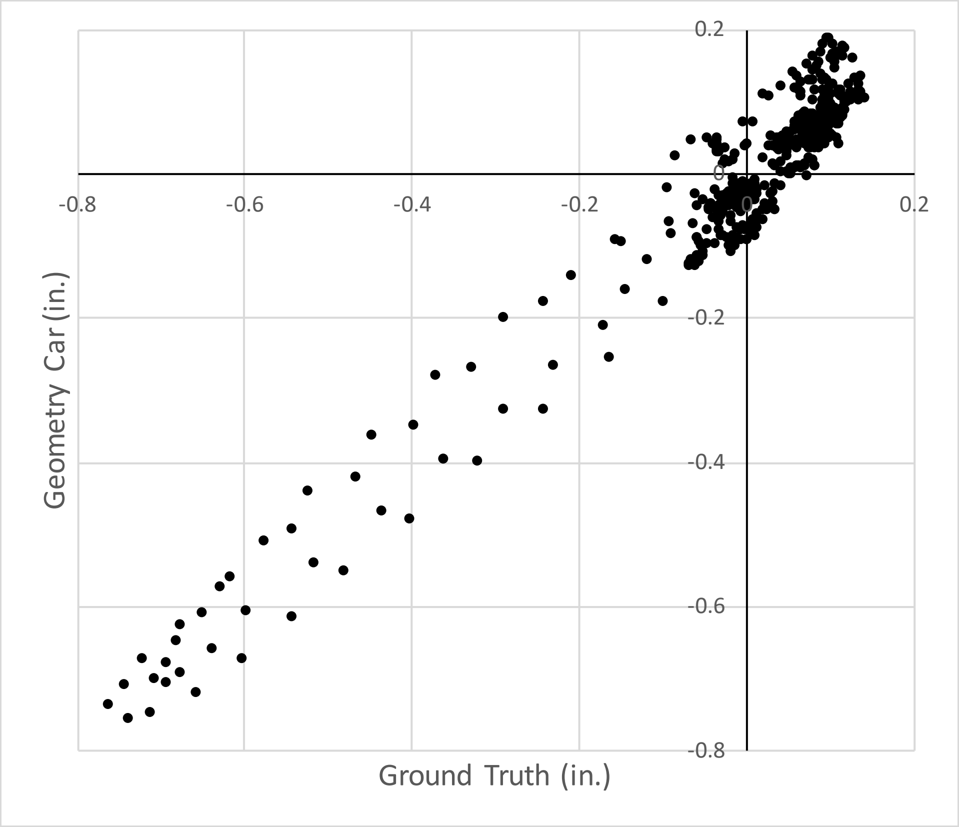


Figure 2. Plot of Geometry Car Data Against Ground Truth

Figure 2 shows an approximately linear correlation between the geometry car data and the ground truth. The best fit straight line to this data has a slope of 0.975 and a bias of -0.015 in. The formula for this line is:

$$Geometry\ Car\ Data\ (Y) = 0.975 \times Gound\ Truth\ (X) - 0.015$$

Slope and bias are two measures of the relationship between two variables. The correlation coefficient r (sometimes stated as R) is another measure of the relationship between two variables. If there are n measurements of two variables X and Y the correlation coefficient is defined as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where $\bar{x}$ and $\bar{y}$ are the means of X and Y.

r_{xy} can vary between -1 and 1. The closer the value is to -1 or 1, the better the correlation. When r_{xy} is equal to 1 or -1, X and Y are perfectly correlated.

Applying Equation 1 to the data in Figure 1, where n = 501, gives a correlation coefficient of 0.971.

Repeating the calculation with data measured at different speeds would show how the correlation varies with speed. The method can also be used to compare the correlation between different measuring cars and the ground truth.

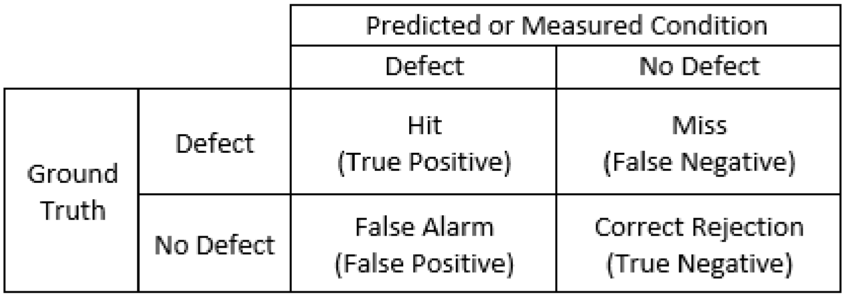
Instead of calculating the correlation coefficient using every foot-by-foot sample of data, it could be calculated for a subset of the data. For example, it could be calculated for the data in Figure 1 with an absolute value greater than 0.5. This would provide information on how well the geometry car measures extreme values in the data.

2.2 Receiver Operating Characteristic

A receiver operating characteristic (ROC) diagram can be used to evaluate track inspection technologies that look for defects. It compares the rate of correctly detecting defects to the rate of false alarms.

An example of a track defect is a broken joint bar joining two pieces of rail end-to-end. Table 1 shows the four possible outcomes when a joint bar inspection system is used.

Table 1. Defect Detection Outcomes

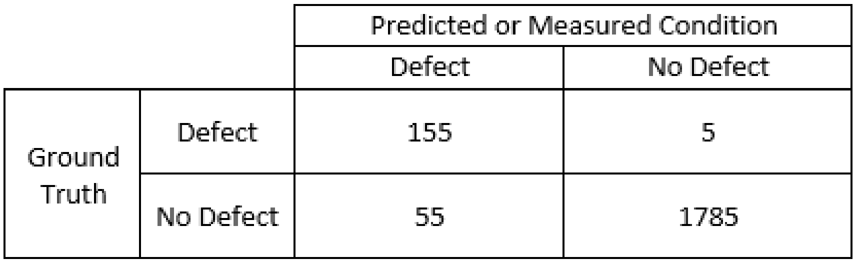


		Predicted or Measured Condition	
		Defect	No Defect
Ground Truth	Defect	Hit (True Positive)	Miss (False Negative)
	No Defect	False Alarm (False Positive)	Correct Rejection (True Negative)

A "hit" occurs when the inspection system correctly detects a broken joint bar. A "miss" occurs when the inspection system fails to detect a broken joint bar. A "false alarm" occurs when a joint bar is reported to be defective and it is not. A "correct rejection" occurs when a joint bar is correctly reported not to be defective.

Consider a section of track with 2,000 joint bars of which visual inspection (assumed to be the ground truth) has found 160 are defective. Table 2 shows a possible example of results from a joint bar inspection system operating on this section of track.

Table 2. Joint Bar Inspection Results



		Predicted or Measured Condition	
		Defect	No Defect
Ground Truth	Defect	155	5
	No Defect	55	1785

Dividing by the numbers of defective and not-defective joint bars gives the following results:

- True Positive Rate = 155/160 = 0.969
- False Positive Rate = 55/1840 = 0.030
- False Negative Rate = 5/160 = 0.031
- True Negative Rate = 1785/1840 = 0.970

Figure 3 shows the false positive rate plotted against the true positive rate on an ROC diagram.

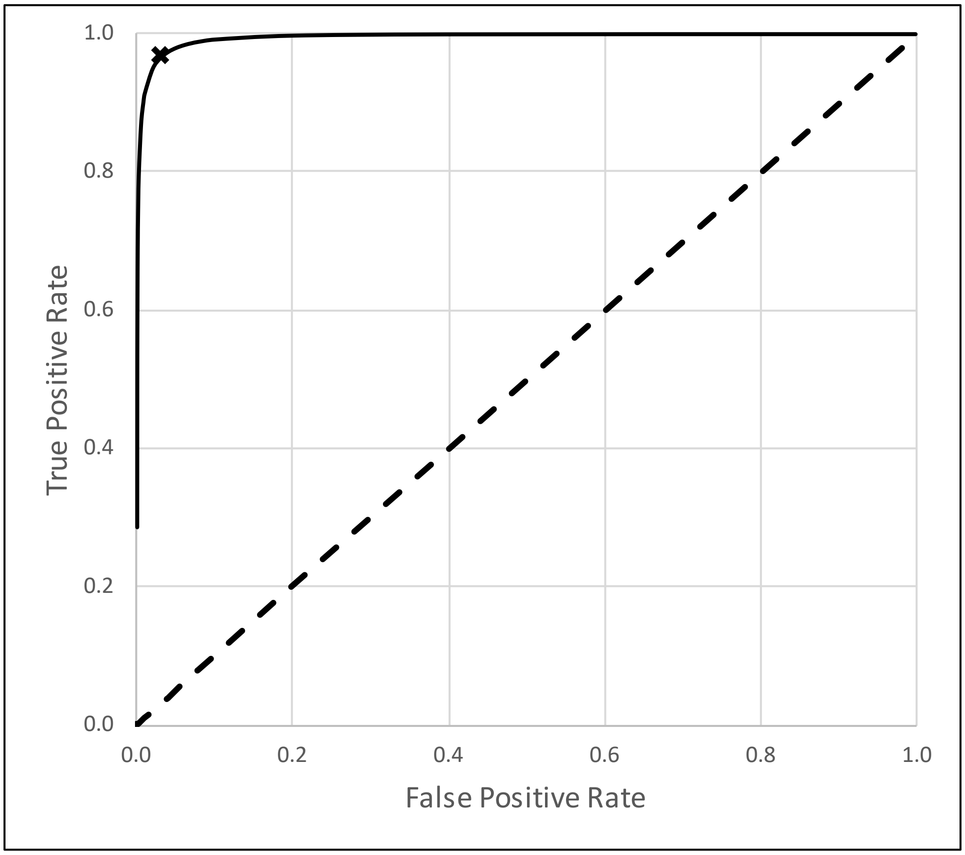


Figure 3. ROC Diagram for Joint Bar Inspection

Sensitivity, denoted by *d'* (d prime), is a measure of ROC performance. It is calculated from:

Sensitivity (d') = z(True Positive Rate) + z(False Positive Rate)

where the appropriate Z-values are obtained from a standard normal table (i.e., the Z-table).

The dashed line in Figure 3 has a sensitivity of zero. The curve that passes through the data point has a sensitivity of 3.7.

The ROC diagram can be used to compare different inspection systems. In general, the closer the result is to the top-left corner of the diagram the better the system is at classifying defects. A result lying on the dashed diagonal line in Figure 3 would show no discrimination between defective and not-defective joint bars. This result would indicate an inspection system that was no better than taking a random guess at the condition of a joint bar.

Accuracy is another measure of an inspection system's performance. It is calculated from:

$$\text{Accuracy} = \frac{(\text{True Positives}) + (\text{True Negatives})}{(\text{True Positives}) + (\text{True Negatives}) + (\text{False Negatives}) + (\text{False Positives})}$$

For the example given in Table 2, the accuracy is 0.969 or 96.9 percent.

2.3 Probability of Detection

Probability of detection (POD) analysis extends the ROC method to evaluate inspection effectiveness for different defect sizes. The method of using POD depends on whether the inspection technology simply identifies a defect, or if it quantifies the size of the defect.

2.3.1 Method A

POD Method A is used when ground truth is known and the inspection technology results in finding a defect. The same four outcomes as shown in Table 1 (i.e., hit, miss, false alarm, correct rejection) are possible.

Figure 4 shows the results of POD Method A for the inspection of defects in welds in railroad tank car shells [3]. Approximately 100 test specimens with known defects of different sizes were manually inspected.

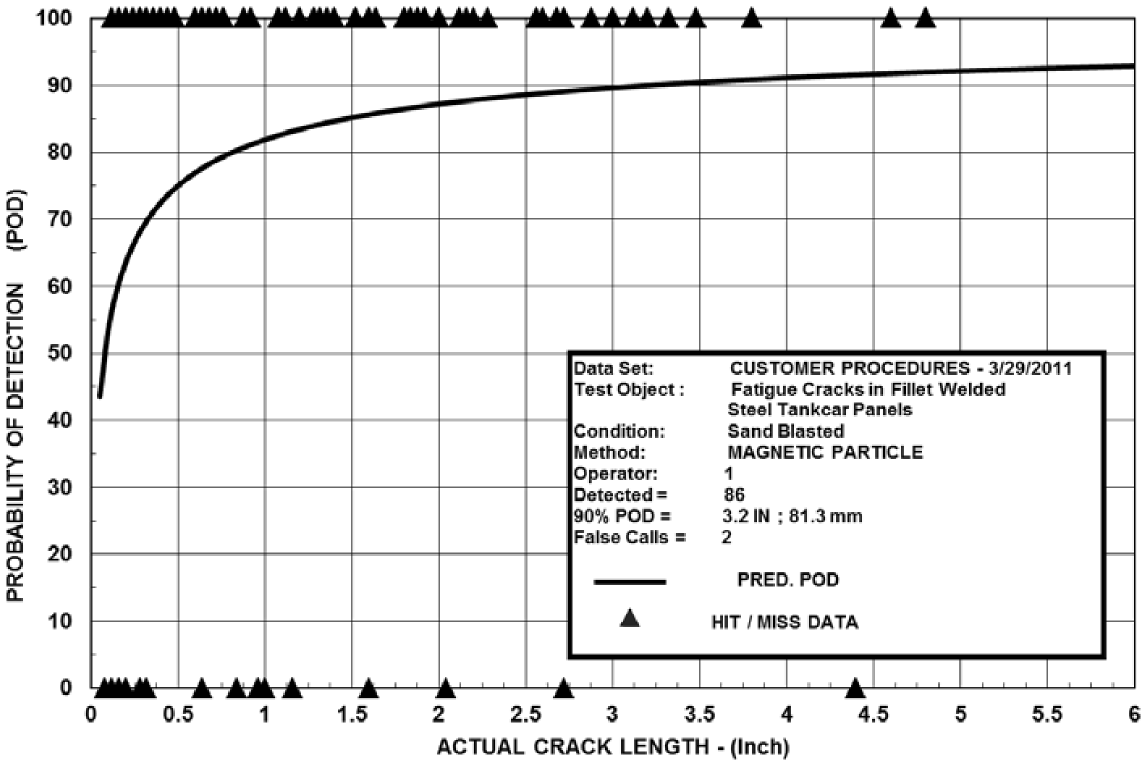


Figure 4. POD Results for Tank Car Inspection

The triangles on the upper horizontal axis of Figure 4 represent defects that the inspector detected. Those on the lower horizontal axis are defects that were not detected. The POD curve in Figure 4 is a mathematical function fitted to the results. It is used to estimate the POD for any defect size. The defect size at 90 percent POD is a commonly used performance measure for inspection technologies. In this example, a defect of 3.2 inches can be detected with an estimated probability of 90 percent.

The POD diagram does not give information on the number of false positives reported. Therefore, this is stated in the text box in Figure 4.

Figure 5 shows an example of the POD diagram being used to compare three different inspection technologies. In this example, the 90 percent POD could not be determined for any of the methods. However, the position of the magnetic particle POD curve above the other two indicates that this inspection technology is the best of the three techniques. The position of the visual inspection POD curve means that it is the worst inspection technique of the three.

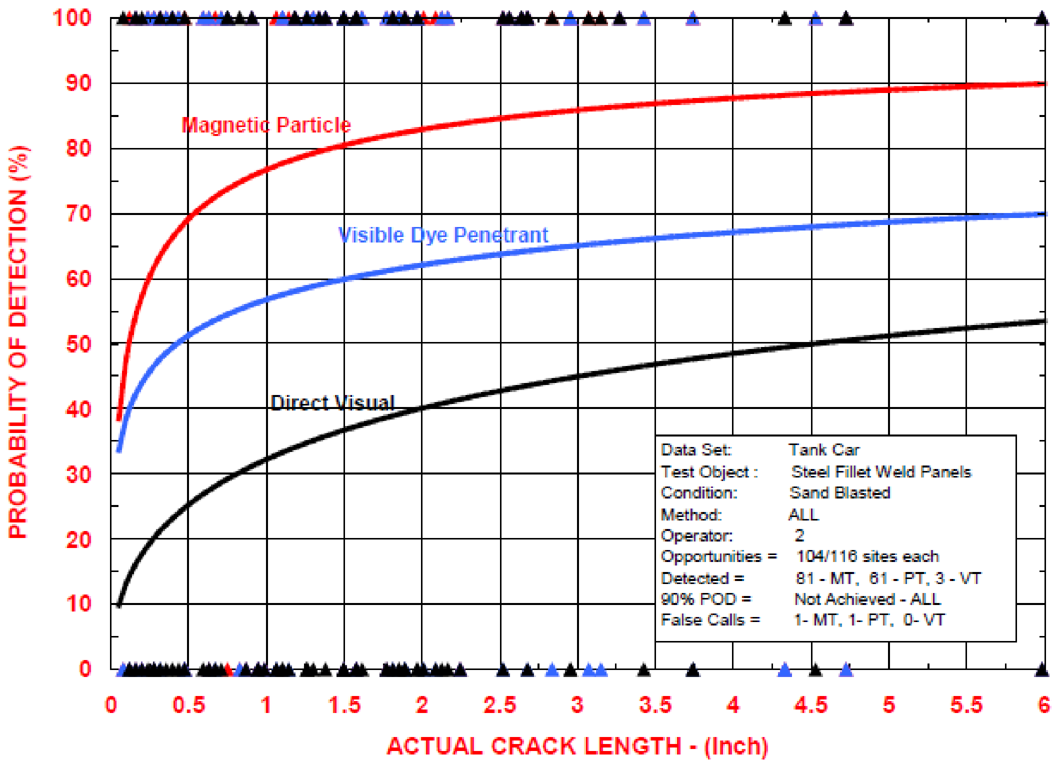


Figure 5. POD Comparison of Inspection Technologies

2.3.2 Method B

POD Method B is used when the inspection technology measures the size of a flaw and ground truth is known. For example, ultrasonic inspection can give the approximate size of individual flaws in the head of rail. In this case, there is a flaw size, a_{dec}, above which the flaw is called a defect.

For each value of the ground-truth size, a, there will be a distribution of measured values, $\hat{a}$. In most cases this distribution is assumed to be a Gaussian probability density function (PDF).

However, in some cases the values of *a* or *â* (or both) need to be normalized by logarithmic or other appropriate transformation. The portion of the PDF that is over the threshold, a_{dec}, is the POD for that value of *a*. This is illustrated in Figure 6.

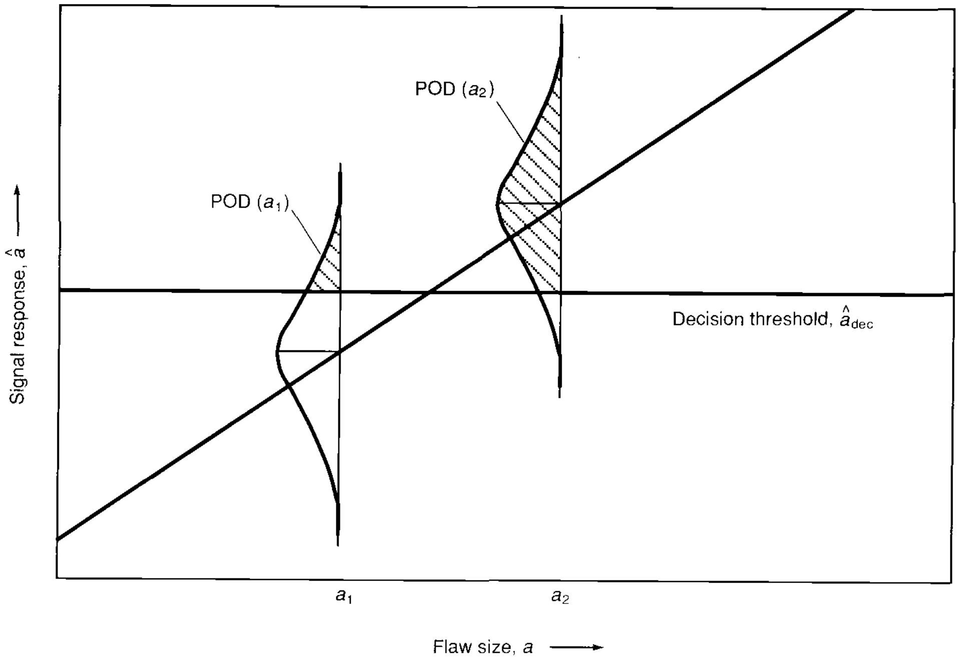

POD (a_2)
POD (a_1)
Signal response, $\hat{a}$
Decision threshold, $\hat{a}_{dec}$
a_1
a_2
Flaw size, a

Figure 6. Signal Responses at Two Flaw Sizes [4]

Figure 6 shows a Gaussian PDF at two flaw sizes, a_1 and a_2 [4]. Flaw size a_1 is below $\hat{a}_{dec}$, on the horizontal axis. However, a portion of the PDF of measured values, *â*, for a_1 fall above the threshold on the vertical axis. The portion above $\hat{a}_{dec}$ is the POD(a_1). The POD(a_2) is larger, but not 100 percent as a small portion of the PDF falls below $\hat{a}_{dec}$.

Tests at the Rail Defect Test Facility at FRA's Transportation Technology Center produced data on the variability of ultrasonic rail flaw measuring equipment [5]. The distribution of measurements of the width of transverse flaws in the head of the rail was found to be approximately Gaussian with a standard deviation of 0.4 inch. For example, making 40 measurements of a flaw with a width of 1.0 inch (a_1 in Figure 6) might result in seven measurements above a decision threshold, $\hat{a}_{dec}$, of 1.4 inches. The probability of detection, POD(a_1), is then 7/40 = 0.175. Similarly, for a flaw with width $a_2 = 2.0$ inches, might result in 37 measurements above 1.4 inches, giving POD(a_2) as 37/40 = 0.925.

This example shows that when there is variability in the measurements made by inspection systems, there is a small probability that flaws smaller than a decision threshold can be reported as being above that threshold. Conversely, not all flaws larger than the threshold will be reported as such.

Figure 7 shows the POD of several flaw sizes using this same example.

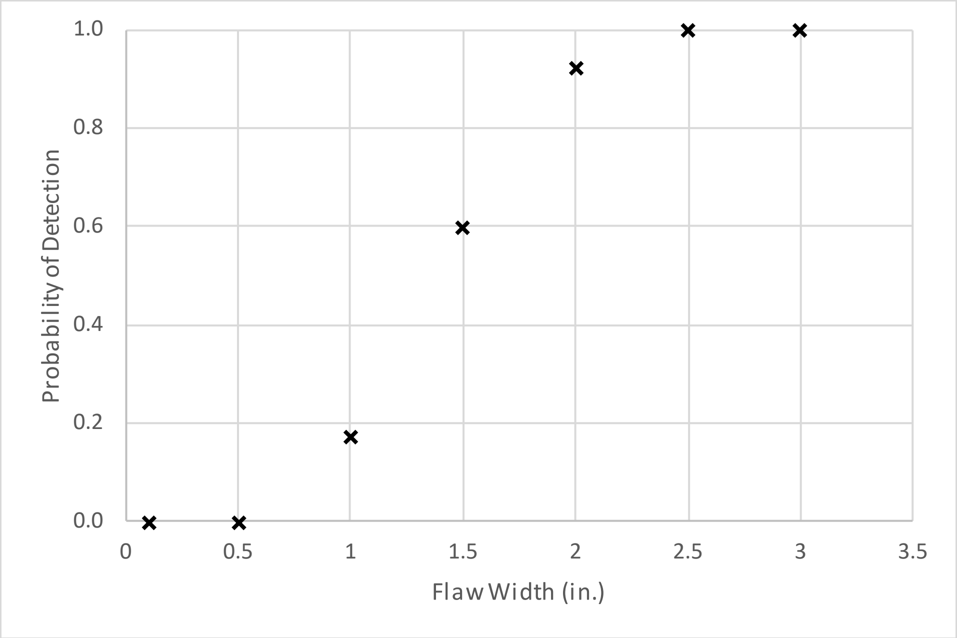


Figure 7. Probability of Detection with a Decision Threshold of 1.4 inches

2.4 Model Assisted Probability of Detection

Obtaining experimental measurements for POD analysis can be expensive due to the cost of manufacturing test specimens and operating the inspection equipment under a large variety of test conditions. The aim of the Model-Assisted Probability of Detection (MAPOD) method is to supplement experimental measurements by simulating the inspection process with a computational model. Several examples of MAPOD applications have been published [6].

The POD of an inspection technology is affected by many factors, including defect morphology, operator skill, equipment variability, and variability in procedure. Some of these factors can be easily described in a computational model. Other factors, such as human variability, must be quantified through carefully designed experiments. The earliest MAPOD studies used two non-Bayesian based different strategies:

1. Transfer Function Method (XFM), which uses a model to generate a new POD curve by transferring values of a factor (e.g., material type, curvature of a part) used in an experimentally obtained baseline curve.
2. Full Model-Assisted (FMA), which uses a model to account for the physical factors that affect the inspection results and combines this with knowledge of human factors to estimate the total variability in the system.

Most recently, a Bayesian framework has been applied to the MAPOD method. The Bayesian estimation procedure combines 'prior' information with new information obtained from experiments to provide a 'posterior' estimate of the POD curve. One type of prior information used in Bayesian MAPOD study is measurements obtained from laboratory experiments. Through application of Bayes' formula, the experimental defect measurement information is combined with similar information taken from real measurements in the field. The posterior POD curve is calculated from fewer samples of more costly field measurements by supplementing these measurements with those from the laboratory. Furthermore, the Bayesian framework allows the POD curve to be continually updated as new data is obtained [7].

Most published examples of MAPOD rely on a ground truth to validate the model. It may also be possible to use MAPOD when a ground truth is unavailable. This might be the case when the modeling software being used is well established and has previously been validated with similar, but not necessarily identical, experimental data. Finite element modeling is an example of such a well-established approach.

2.4.1 Non-Bayesian MAPOD

Table 3 lists published examples from a literature review of non-Bayesian MAPOD demonstrations of non-destructive examination (NDE) methods [8]. The references include examples where the flaw size was compared to a known size ($\hat{a}$ vs. a) and when a defect was detected or not (hit-or-miss).

Table 3. References to Non-Bayesian MAPOD Demonstrations [8]

MAPOD Approach	NDE Response	NDE Method	Applied to	Reference
XFM	$\hat{a}$ vs. a	Eddy current testing	Fatigue cracks in complex engine components	Thompson et al. [6]
FMA	$\hat{a}$ vs. a	Eddy current testing	Fatigue cracks in wing lap joints	Thompson et al.. [6]
FMA	$\hat{a}$ vs. a	Ultrasonic testing	Defects in engine disk alloys with microstructural variability	Thompson et al. [6] Smith et al. [9]
XFM	$\hat{a}$ vs. a	Ultrasonic testing	Fatigue cracks around fastener holes for aircraft	Harding et al. [10]
XFM	$\hat{a}$ vs. a	Ultrasonic testing	Fatigue cracks in aluminum components	Demeyer et al. [11]
XFM, FMA	$\hat{a}$ vs. a	Eddy current testing	Fatigue cracks in titanium plates	Rosell and Persson [12]
FMA	$\hat{a}$ vs. a	Eddy current testing	Cracks in fastener sites	Aldrin et al. [13]
XFM, FMA	$\hat{a}$ vs. a	Ultrasonic testing	Defects in railway axles	Carboni and Cantini [14]
XFM	Hit-or-miss	Ultrasonic testing	Airplane lap joint specimen sets with multiple site fatigue damage	Bode et al. [15]

Rosell and Persson give an example of the MAPOD method using transfer functions that was applied to fatigue cracks in titanium plates [12]. Eddy current scans were made over 53 plates with known crack lengths to establish the baseline POD. Two different scan grids were used with spacings of 0.5 and 1.0 mm. Figure 8 shows the POD curves derived from these experimental results.

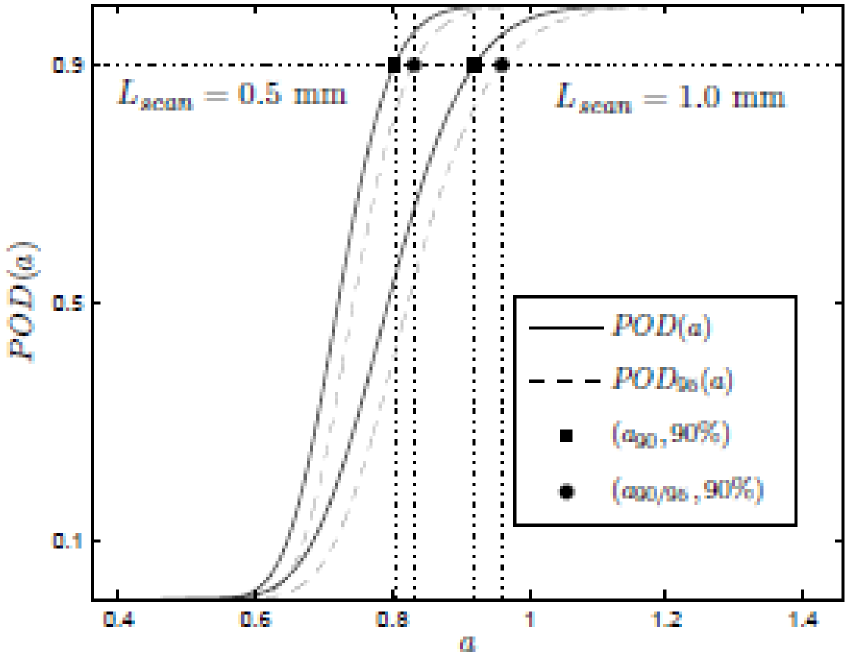


Figure 8. Experimental POD Curves [12]

The 90 percent lower confidence limits for the POD curves are plotted as dashed lines in Figure 8. An example observation from these results was that the probability of detecting a crack of length 0.84 mm using the 0.5 mm spacing was 90 percent in 90 cases when the experiment was repeated 100 times.

After the experiments were completed, a model of the eddy current inspection process was developed. The model used finite elements to simulate the cracks and predict the response of the eddy currents. Several variables were modeled, including grid spacing, crack orientation and shape, and probe handling. These were assumed to be uniformly or normally distributed with appropriate statistical properties. The model was then used to predict POD curves for the two different grid spacings.

Figure 9 compares the POD curves from the model (dotted lines) with those from the experimental results (solid lines) for the two different grid spacings. The upper 95 percent confidence limits for the model (dot-dash lines) and the experimental results (dashed lines) are also shown.

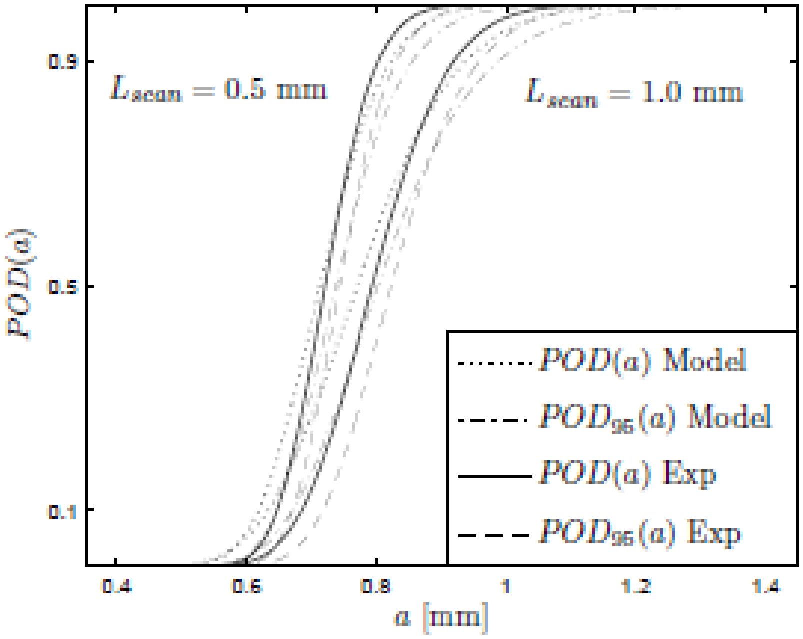


Figure 9. Modeled and Experimental POD Curves [12]

The next step would be to make a judgement on the agreement between the model and the experimental results. If better agreement was needed, the model could be developed further. If the agreement was acceptable, the model could be used to predict POD curves for other conditions such as larger grid spacing.

2.4.2 Bayesian MAPOD

Table 4 lists published examples from a literature review of Bayesian MAPOD demonstrations of NDE methods [8].

Table 4. References to Bayesian MAPOD Demonstrations [7]

NDE Response	NDE Method	Applied to	Reference
Hit-or-miss	Visual testing	Demonstrate Bayesian approach to POD determination based on limited field data.	Leemans and Forsyth [16]
Hit-or-miss	Eddy current testing	Demonstrate Bayesian approach to POD determination using computer models to generate additional information to supplement limited data from experiment.	Jenson et al. [17]
$\hat{a}$ vs. a	Radiographic testing	Demonstrate Bayesian approach to POD demonstration using information from artificial flaws to supplement that from a limited set of real flaws.	Kanzler et al. [7]

NDE Response	NDE Method	Applied to	Reference
â vs. *a*	Eddy current testing	Demonstrate Bayesian approach to POD demonstration using computer model generated information.	Aldrin et al. [18]
â vs. *a*	Not specified	Demonstrate method for determination of both POD and crack size distribution from in-service inspection data.	Hovey [19]
Hit-or-miss	Magnetic leakage	Apply a hierarchical Bayes approach to incorporate influence of spatial distributed uncertainties on in-line inspections of pipelines.	Dann and Maes [20]

Kanzler et al. give an example of using the Bayesian approach to estimate POD for radiographic testing (RT) of flaws in welded copper canisters [7]. The results of testing on artificial flaws of known sizes were used to develop a model to predict the relationship between RT response and flaw size. The model could also predict the variance in RT response at any artificial flaw size, and hence it could predict the POD using Method B.

A similar model was developed from a small number of tests with real flaws. The real flaws were broken open after testing to determine their sizes.

Bayes' theorem was then used to combine the model from the small number of tests on real flaws with the model from the larger number of tests on artificial flaws. This improved the confidence in the results. For example, the size of the defect that could be detected with 90 percent probability and 95 percent confidence reduced from 1.2 mm with the model from the small number of real flaws to 1.0 mm with the combined model from real and artificial flaws.

2.5 Repeatability and Reproducibility

For any inspection technology there will be (1) variability when measuring the same test sample and (2) factors that the operator can change that may result in a difference in the measurements. The first situation describes repeatability, which is the result of measuring the same test sample multiple times keeping the conditions of the measurement as consistent as possible. Reproducibility, on the other hand, is the result of measuring the same test sample but adjusting conditions that the system operator can control (e.g., changing the measurement speed, direction, and orientation). Ground truth is not needed for repeatability or reproducibility analysis.

For repeatability, it is important that external sources of variation are avoided so the differences between measurements are only due to the inspection technology. Figure 10 shows an example of two measurements of track profile over the same 500 feet of track.

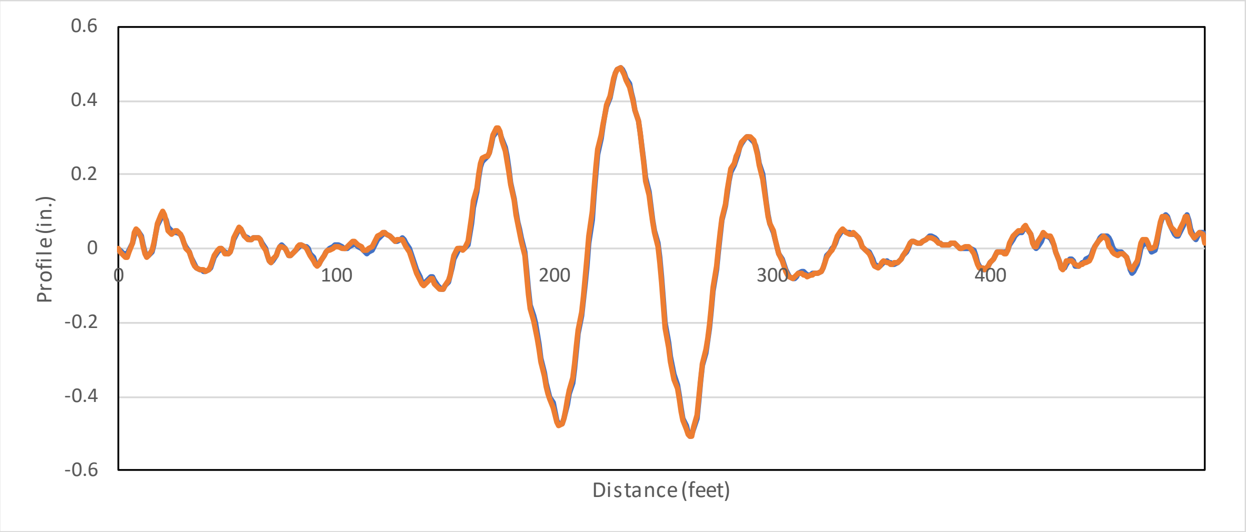


Figure 10. Repeated Track Profile Measurements

The measurements in Figure 10 were made at the same speed and in the same direction to avoid the effect of these variables on the analysis of repeatability. Although other variables such as sunlight and humidity were not controlled, one measurement was made immediately after the other. This minimized the effect of the uncontrollable variables.

A measure of repeatability is the variance (or standard deviation) calculated for a group of measurements collected under similar conditions. The variance can be calculated using all data points to give an overall variance. Alternatively, when repeatability at extreme values is of interest, the variance of maxima and minima within groups of measurements can be calculated.

The repeatability of an established inspection technology can be measured and used to set a gold standard. Then, the repeatability of new inspection technologies can be compared to this standard.

Reproducibility is the closeness of agreement between average measurements taken under varied conditions. Figure 11 shows an example of track profile measurements made in the forward and reverse directions over the same 500 feet of track. The speed was the same for both measurement runs.

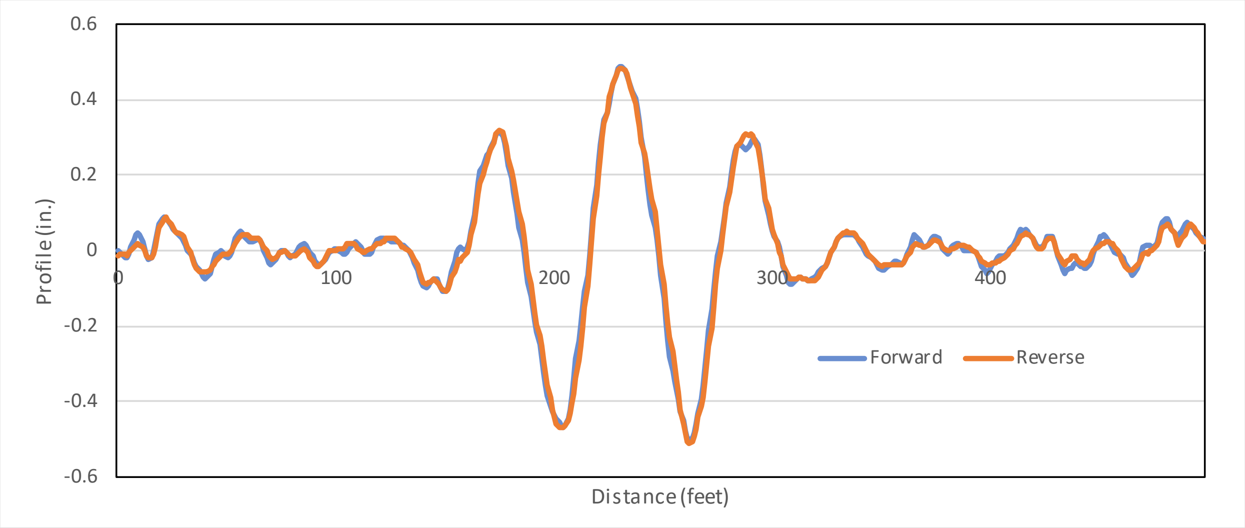


Figure 11. Track Profile Measurements in Different Directions

Figure 12 shows an example of track profile measurements made at 20 and 100 mph over the same 500 feet of track. The direction was the same for both measurement runs.

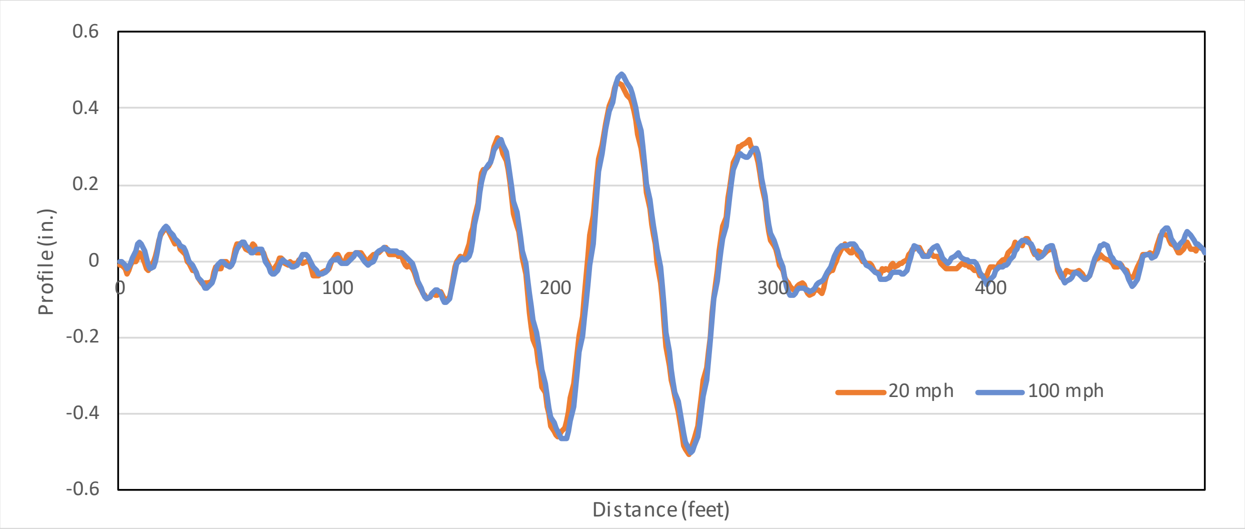


Figure 12. Track Profile Measurements at Different Speeds

Reproducibility can be measured by comparing the variance (or standard deviation) between groups of measurements made under different conditions with the inherent variance of the inspection technology (calculated from its repeatability). Analysis of variance (ANOVA) methods can be used to determine if the reproducibility is worse than that expected.

As with repeatability, the variance in reproducibility can be calculated using the averages of all data groups. Alternatively, the variance of maxima and minima within the measurement groups can be calculated. This alternative is useful when reproducibility at extreme values is of interest.

The reproducibility of an established inspection technology can be measured and used to set a gold standard. In turn, the reproducibility of new inspection technologies can be compared to this standard. This comparison can be made for different operating conditions, such as speed and direction of travel.

2.6 Sample Size

The number of test samples available to evaluate an inspection technology has a significant effect on the results. Current literature recommends 30 samples per flaw size for POD. In the POD Method A example in Section 2.3.1, there were approximately 90 test samples with a range of crack lengths up to 6 inches. No single crack length had more than six samples. To satisfy current best practice, there should have been at least 30 test samples for each crack length. This large number of test samples is usually impractical.

Error bounds are the range of numbers within which a certain percentage of a population are expected to fall, the percentage being the confidence interval. When the sample size is less than 30, the error bounds are given by:

$$B = \pm \frac{s * t_{confidence}}{\sqrt{n}} + \bar{X}$$

where n is the sample size, s is the standard deviation of the sample measurements, $\bar{X}$ is the mean of the sample measurements, and $t_{confidence}$ is a function of the confidence level. When the sample size is 30 or more, the parameter $t_{confidence}$ is replaced with Z. Table 5 gives Z-values for common confidence intervals.

Table 5. Z-values for Common Confidence Intervals

Confidence Interval (%)	Z-value
80	1.282
90	1.645
95	1.960
99.9	3.291

The user determines the confidence level and, if greater confidence is required, the error bounds widen. Given a small sample size, an acceptable confidence level must be balanced with acceptable error bounds.

In the example in Section 2.3.1 there were six samples with cracks 2.75 inches long. The estimated POD at this crack length was 0.88. T-tables give a value of $t_{confidence}$ = 0.727 for this sample size and a 75 percent confidence level [22]. Assuming s = 1, the formula above gives the lower error bound as 0.58 and the upper error bound as greater than 1.0. Thus, there is a 75 percent confidence that the true POD for this crack length lies between 0.58 and 1.0 (since the maximum possible value is 1.0).

This example shows how small sample sizes can weaken the conclusions that can be drawn from experiments.

Figure 13 shows an example of how error bounds vary with sample size for different confidence levels. In this example the standard deviation, s, is 1, and the sample mean, $\bar{X}$, is again 0.88. As before, the lower error bound with a sample size of six and a confidence level of 75 percent is seen to be 0.58. Reducing the sample size to four, for example, reduces the lower error bound at the same confidence level to 0.55. Alternatively, increasing the confidence level to 90 percent and keeping the sample size at six reduces the lower error bound at the same confidence level to 0.28.

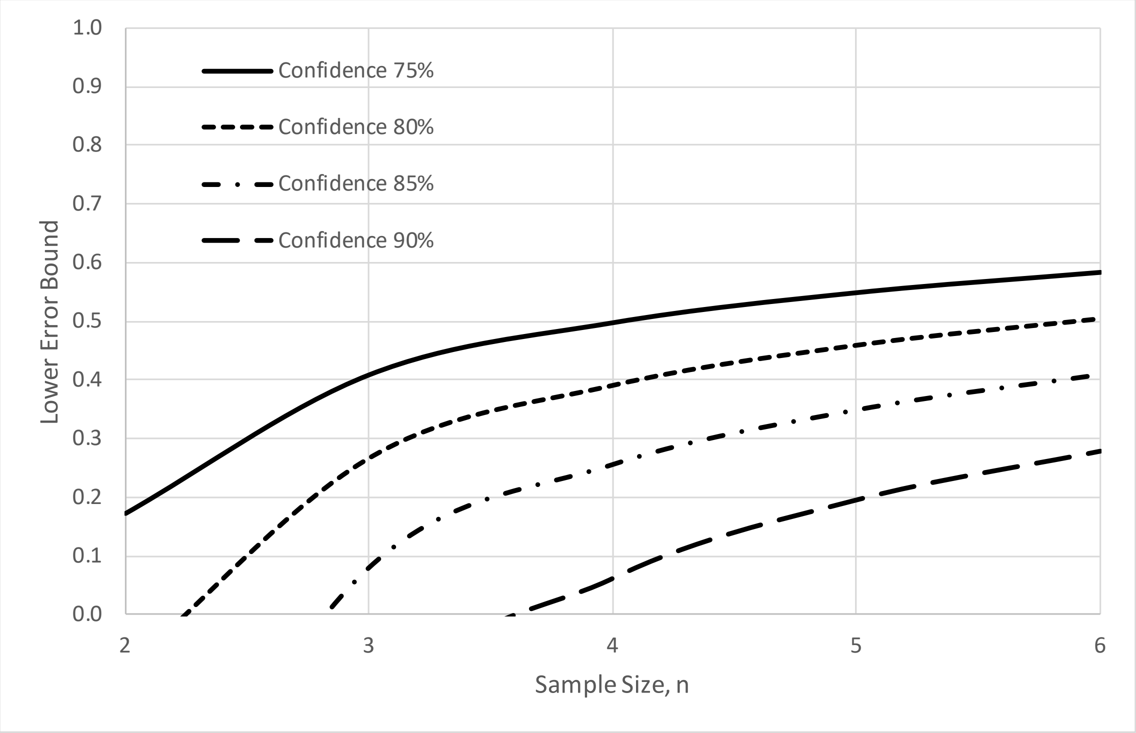


Figure 13. Effect of Sample Size on Lower Error Bounds

When there are few samples for analysis there are two methods recommended in the literature to achieve a higher confidence with smaller error bounds: binning and supplementing [4].

The binning of flaw sizes allows for several samples to be grouped, and each sample within the group can be added to the value n. The POD is of most importance on either side of the critical flaw size, a_{dec}, which is the size at which a flaw is considered a defect. This should be considered when determining bin ranges. Bins on either side of a_{dec} should be roughly equivalent and ideally the most populated. Bins at the lower and upper ends of flaw sizes should populated by enough samples to achieve a reasonable confidence level and tolerable error bounds.

Supplementing the sample size with data from different sources of measurements increases confidence in the results. It is an acceptable approach if the different measurements are shown to be repeatable and reproducible.

3. Categories of Inspection Technology

Section 2 highlighted the importance of ground truth measurements in evaluating the effectiveness of inspection technologies. Strong conclusions can only be made when ground truth is available. Thus, there are two basic types of inspection technologies—where ground truth is available and where it is not.

In practice, absolute ground truth is rarely available. In the case of the track profile measurements described earlier, ground truth was taken from accurate surveys, but even these were not free from errors.

When ground truth is available it may be time-consuming and costly to establish. For example, in the case of the joint bars described earlier, the ground truth could be established by visually inspecting all 2,000 samples.

In some cases, it may be acceptable to replace ground truth with an established method of inspection, referred to as a gold standard. For example, a gold standard for track geometry measurement might be an inspection system that has been used by the industry for many years and is acknowledged as giving valid results. It could also be an inspection technology that has previously been validated using ground truth.

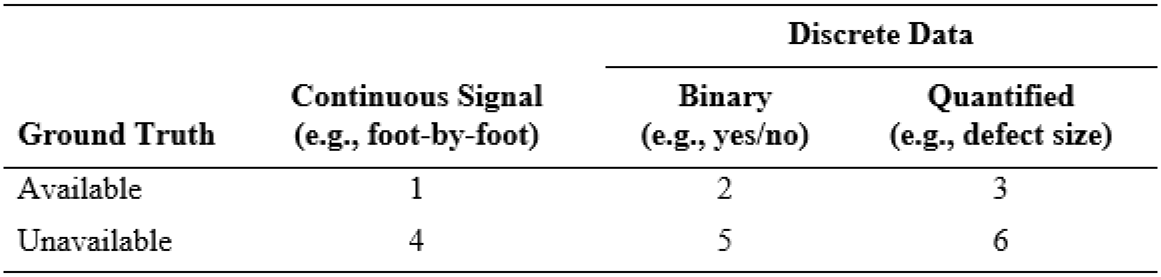
The data gathered by track inspection technologies falls into two categories:

1. Continuous signals, such as the foot-by-foot data recorded by TGMS
2. Discrete data (i.e., data at a particular track location), such as the condition of a joint bar and the flangeway clearance at a crossing diamond

Discrete data can be further categorized as binary (e.g., a fatigue crack is present or not) or quantified (e.g., the length of a fatigue crack).

Table 6 allocates numbers from 1 through 6 to the different categories of data produced by track inspection technologies.

Table 6. Categories of Track Inspection Data



		Discrete Data	
Ground Truth	Continuous Signal (e.g., foot-by-foot)	Binary (e.g., yes/no)	Quantified (e.g., defect size)
Available	1	2	3
Unavailable	4	5	6

The Appendix lists currently available track inspection technologies and shows which categories of data they produce. Note that some inspection technologies can produce more than one category of data.

4. Recommended Methods of Evaluation

Table 7 shows the recommended methods of evaluation for the data categories defined in Table 6.

Table 7. Recommended Evaluation Methods

Data Category	Description	Method
1	Continuous signal with ground truth	Correlation
2	Binary discrete data with ground truth	ROC Curve, POD Method A, MAPOD
3	Quantified discrete data with ground truth	Correlation, POD Method B, MAPOD
4	Continuous data – no ground truth	Repeatability, Reproducibility
5	Binary discrete data – no ground truth	MAPOD, Repeatability, Reproducibility
6	Quantified discrete data – no ground truth	MAPOD, Repeatability, Reproducibility

Where more than one method is listed in Table 7, the preferred method depends on the availability of data, costs of experiment and analysis, objective of the evaluation, and the criteria being used to judge the effectiveness of the inspection technology. Using more than one method can give additional insight into the capabilities of the system being evaluated.

5. Evaluation of Track Geometry Measuring System Effectiveness

This section gives detailed examples of applying the methods described in Section 2. Data from a TGMS are used. The purpose of the analysis is to answer the question: How effective is the system at measuring track geometry?

Table A.1 in Appendix A shows that track geometry data covers data categories 1, 2, and 3. Since the data is measured continuously it is in category 1. The data can also be analyzed and compared with limits set out in track safety standards and regulations. This pass-or-fail criteria is an example of binary data in category 2. The analysis can also quantify the magnitude of track geometry flaws, which puts the data in category 3.

Table 6 recommends that data in categories 1, 2, and 3 be analyzed by Correlation, ROC Curve, POD Methods A and B as well as MAPOD. Examples of applying each of these methods to track geometry data are provided in the following sub-sections. Examples of repeatability and reproducibility analysis are also provided.

The data used in these examples were measured on the High-Speed Adjustable Perturbation Slab Track section of the Railroad Test Track (RTT) at FRA's Transportation Technology Center. The 31-foot mid-chord offset for the vertical track profile of both rails was the measurement selected for analysis.

Figure 14 shows the ground truth geometry at the section of track used in the analysis. Measurements with high-accuracy surveying equipment were considered ground truth for this analysis. The measurements were made every foot for 500 feet.

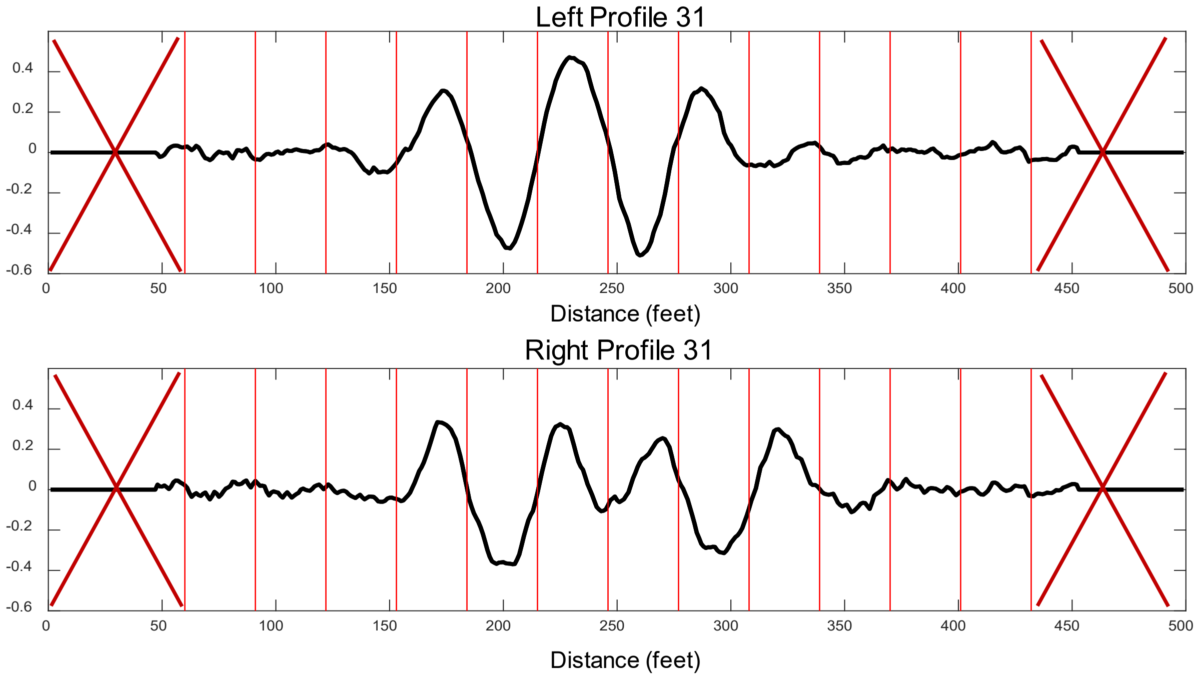


Figure 14. Ground Truth Geometry at Test Section

To allow run-in and run-out of the 31-foot mid-chord measurement, 54 feet of data was eliminated from the start and end of the measurement. This left 393 feet of individual data points that could be divided into 12 segments, 31 feet long on both the left and right rails (24 segments in total), as shown in Figure 14.

The TGMS made measurements over the test section at six different speeds ranging from 20 to 100 mph. Measurements were made in clockwise and counter-clockwise directions on the RTT. They were also made with the TGMS running in the forward and reverse orientations. A total of 72 measurement runs were available for analysis—three measurement runs for each combination of speed, direction, and orientation. The measurements were made in July 2017.

5.1 Repeatability and Reproducibility

The repeated measurements at the same speed, direction, and orientation allow the repeatability of the TGMS to be analyzed. Once that has been established, the TGMS reproducibility can be analyzed by comparing results at different speeds, directions and orientation.

In this analysis, the means and variances of the 31-foot mid-chord offset were calculated in each of five bins, each 0.1 in wide. Table 8 lists the number of ground truth samples in each bin.

Table 8. Foot-by-Foot Samples for Each Combination of Speed, Direction, and Orientation

31-Foot Mid-Chord Offset Bin	Samples
0.0 to 0.1 in.	541
0.1 to 0.2 in.	70
0.2 to 0.3 in.	87
0.3 to 0.4 in.	57
Greater than 0.4 in.	31

Figure 15 shows a whisker plot of the results for the measurements at 20 mph. Each data point indicates the mean and variance (shown as ±3 standard deviations from the mean) for the three measurement runs at a combination of direction and orientation.

The top-leftmost data point in Figure 15 shows the repeatability of the TGMS in the clockwise direction, operating in the forward orientation, for the data in the bin from 0.0 to 0.1 inch. The whisker length is smallest in this bin and increases as the magnitude of the bin increases. This could be partially due to the fewer data points in the higher bins.

The four top-leftmost data points in Figure 15 show the reproducibility of the TGMS when the direction and orientation are varied for the data in the bin from 0.0 to 0.1 inch. The overlap of the four results (whiskers) indicate no statistically significant difference among the means of the four measurements. The TGMS has a good reproducibility in this range. The results also overlap in most of the higher bins. One exception occurs in the bin from 0.2 to 0.3 inch where there is a significant difference in the counterclockwise measurement run in the forward orientation and the clockwise measurement run in the reverse orientation, denoted by the red data points in the figure. A similar mismatch occurs in the bin from 0.3 to 0.4 inch.

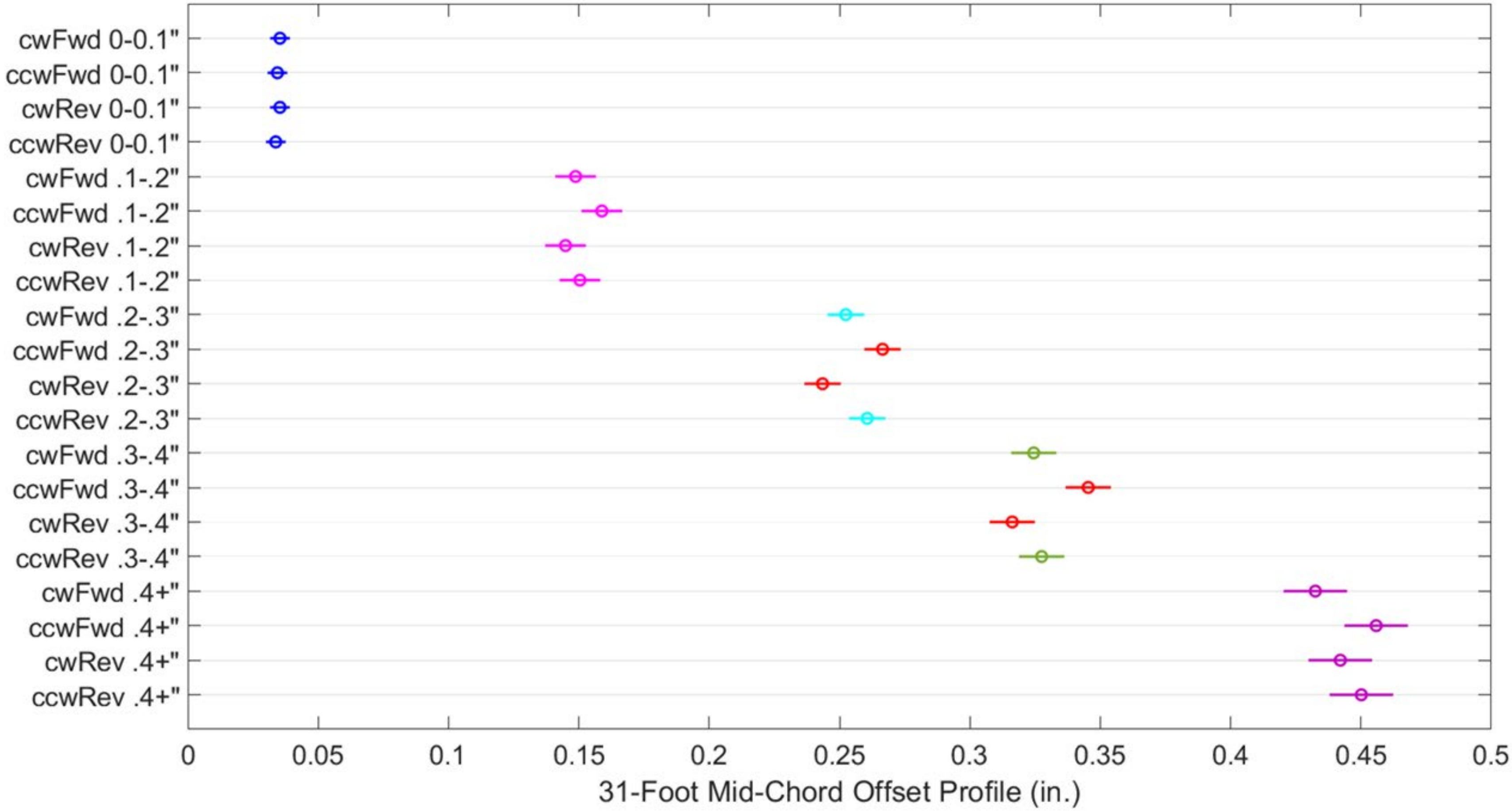


Figure 15. Repeatability and Reproducibility Results for Foot-by-Foot Data at 20 mph

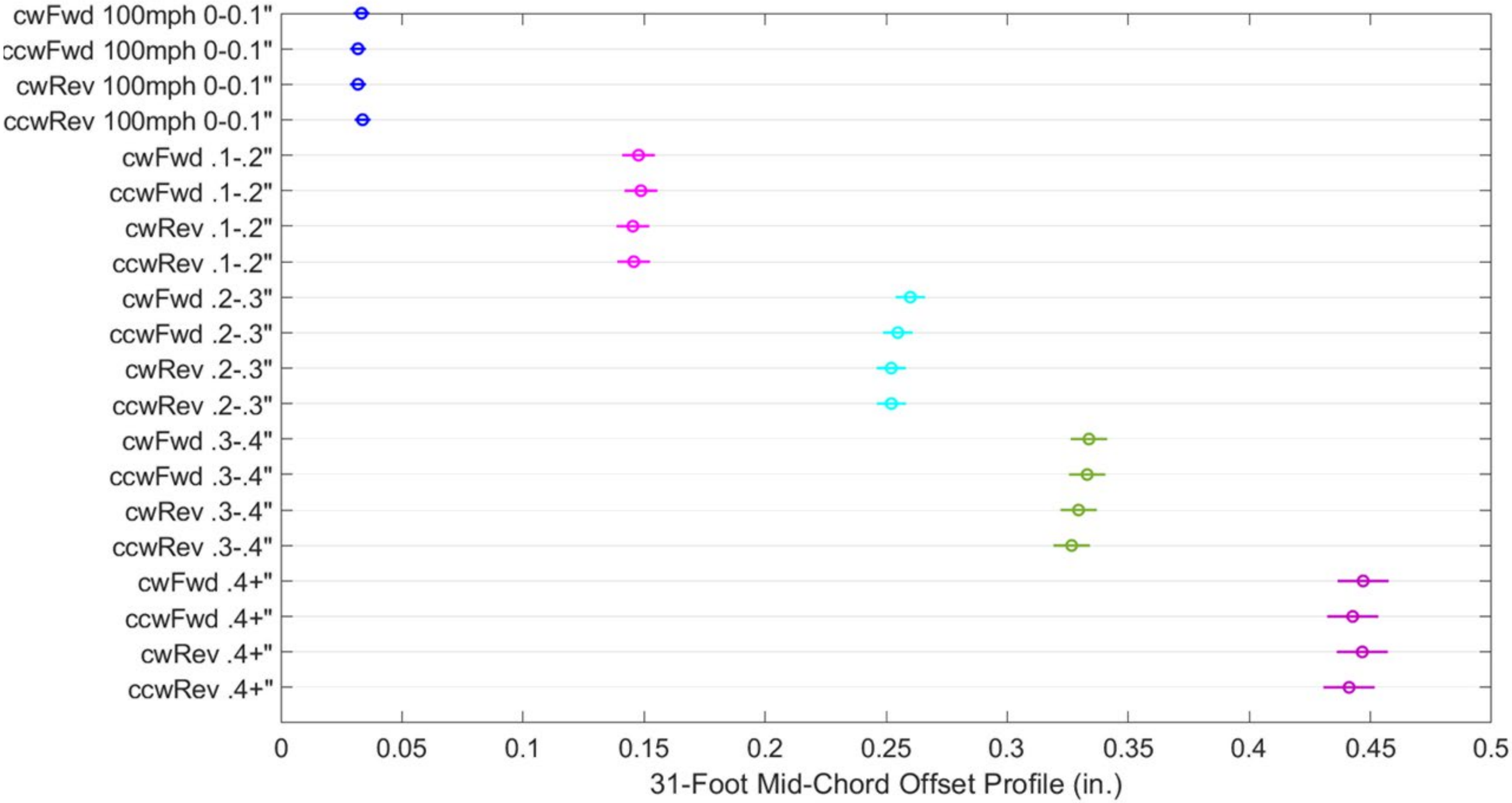


Figure 16. Repeatability and Reproducibility Results for Foot-by-Foot Data at 100 mph

Figure 16 shows the repeatability and reproducibility results for the TGMS measurements at 100 mph. The results show good repeatability and reproducibility in all bins.

Comparing Figure 15 with Figure 16 shows there is less variance in the 100 mph measurements than at 20 mph. The TGMS has better repeatability and reproducibility at 100 mph compared to at 20 mph.

5.2 Correlation

Correlation analysis shows the strength of the relationship between the TGMS measurements and the ground truth. Figure 17 shows the correlation for a single measurement run at 20 mph (sample size of 393). The absolute value of the data at each foot is plotted. The straight line is the least squares best-fit to the data. Figure 18 shows the correlation for a single measurement run at 100 mph.

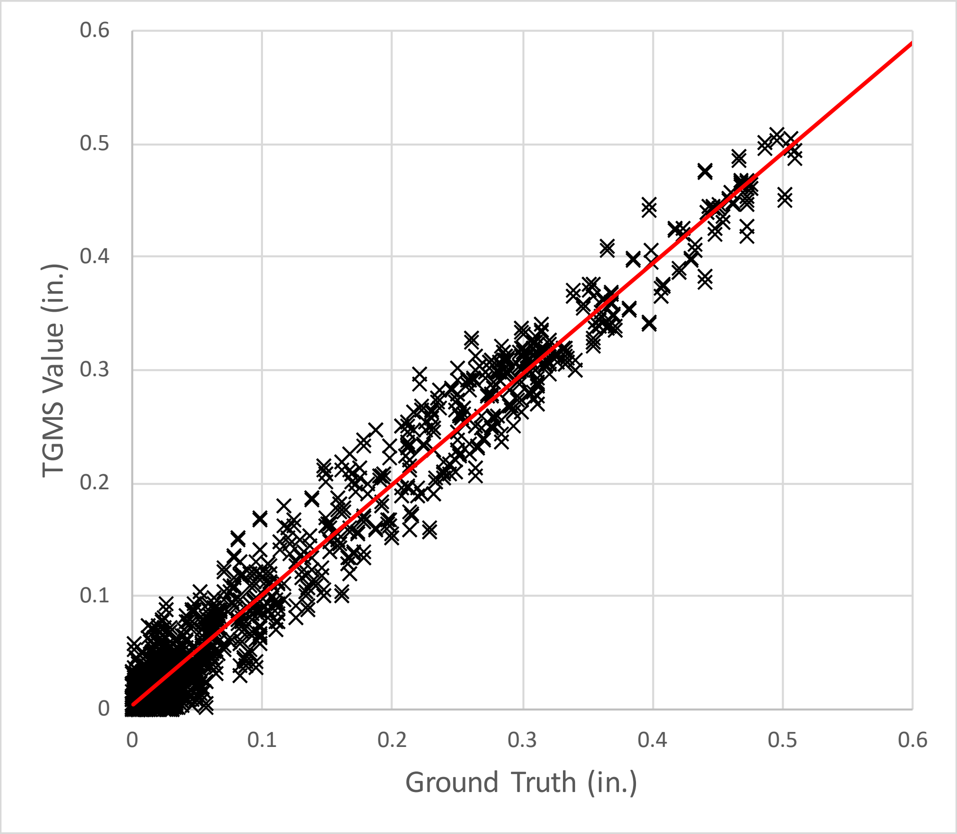


Figure 17. Foot-by-Foot Correlation at 20 mph

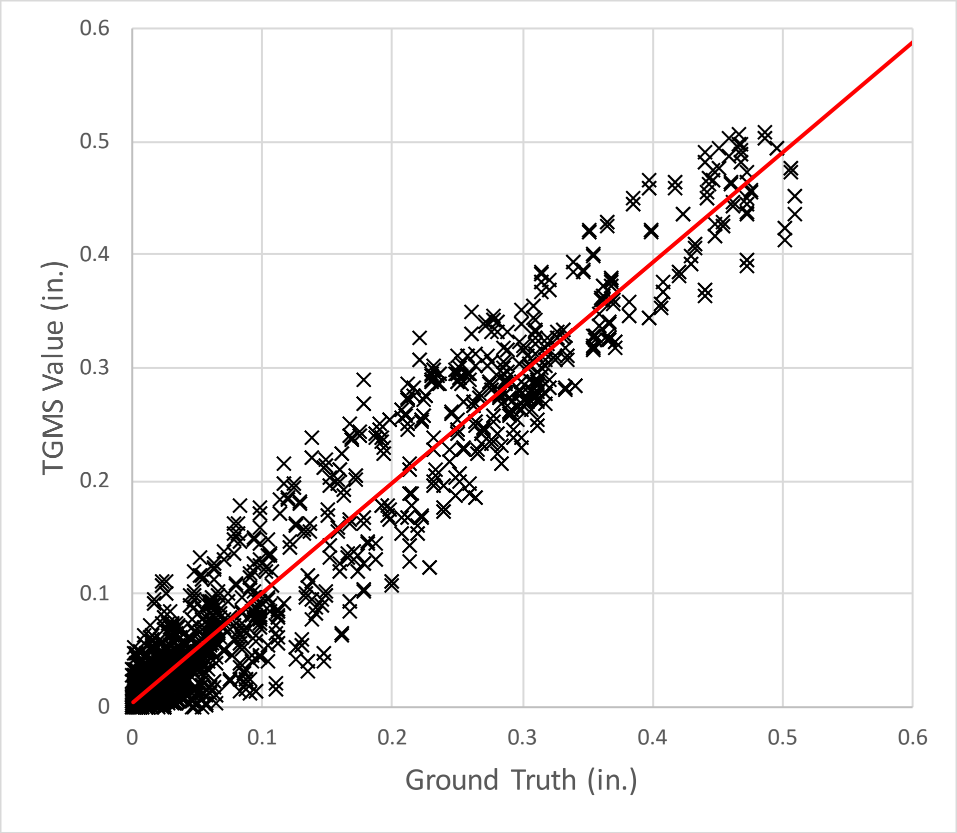


Figure 18. Foot-by-Foot Correlation at 100 mph

Figure 17 and Figure 18 show a reasonably linear relationship between the TGMS data and the ground truth. There appears to be a slightly lower correlation at 100 mph compared to that at 20 mph. Table 9 shows the correlation statistics for the foot-by-foot data at 20 and 100 mph.

Table 9. Foot-by-Foot Correlation Statistics

Speed (mph)	20	100
Slope	0.976	0.971
Bias	0.004	0.005
R^2	0.969	0.941
Correlation Coefficient	0.984	0.970

The slope of the relationship between TGMS data and ground truth is similar at 20 and 100 mph. The TGMS slightly underestimates the ground truth. The bias is similar at 20 and 100 mph and is very small.

The R^2 and correlation coefficient values in Table 9 confirm the observation that there is better correlation at 20 mph compared to 100 mph.

An example of an alternative correlation analysis is to consider the peak measurement (maximum absolute value) in each 31-foot track segment. Figure 19 and Figure 20 show this correlation at 20 and 100 mph, respectively. The data is combined from three measurement runs at each speed in the same direction and orientation (sample size of 3 × 24 = 72).

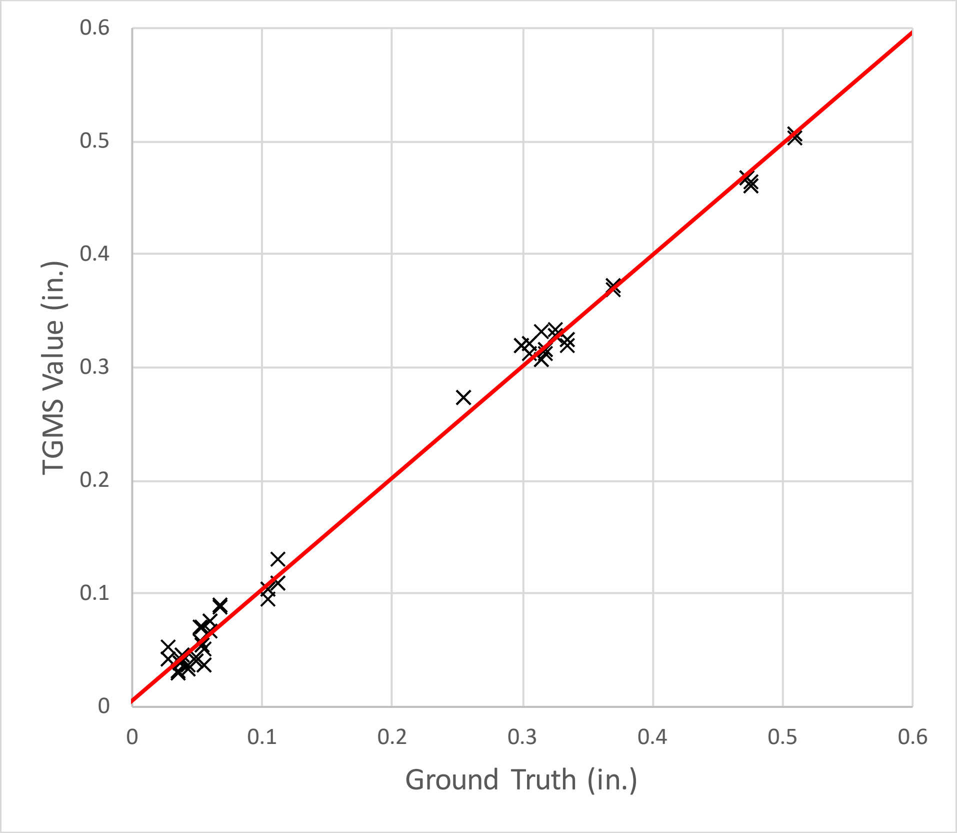


Figure 19. 31-Foot Correlation for Three Measurement Runs at 20 mph

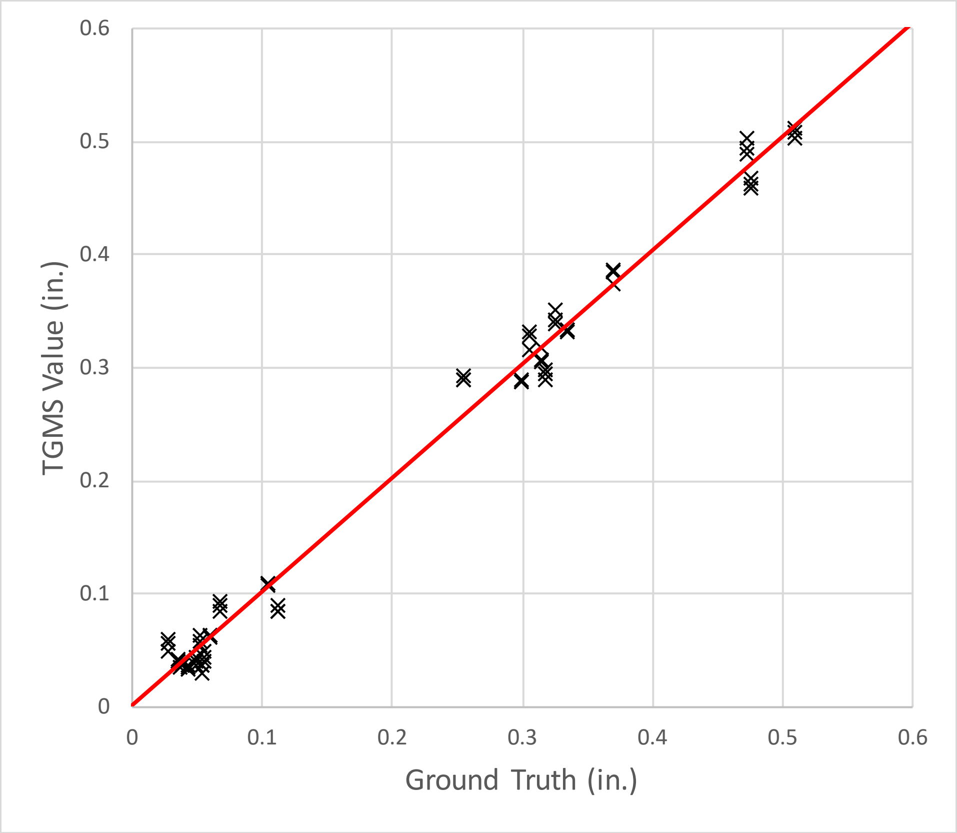


Figure 20. 31-Foot Correlation for Three Measurement Runs at 100 mph

Figure 19 and Figure 20 show a reasonably linear relationship between the TGMS data and the ground truth. Table 10 shows the correlation statistics for the 31-foot segment data at 20 and 100 mph.

Table 10. 31-Foot Correlation Statistics

Speed (mph)	20	100
Slope	0.986	1.008
Bias	0.006	0.001
R^2	0.995	0.990
Correlation Coefficient	0.998	0.995

The statistics in Table 10 show a very good correlation between TGMS data and ground truth. The bias is very small, and the other statistics are close to the ideal value of unity. There is little difference between the correlation at 20 mph and 100 mph.

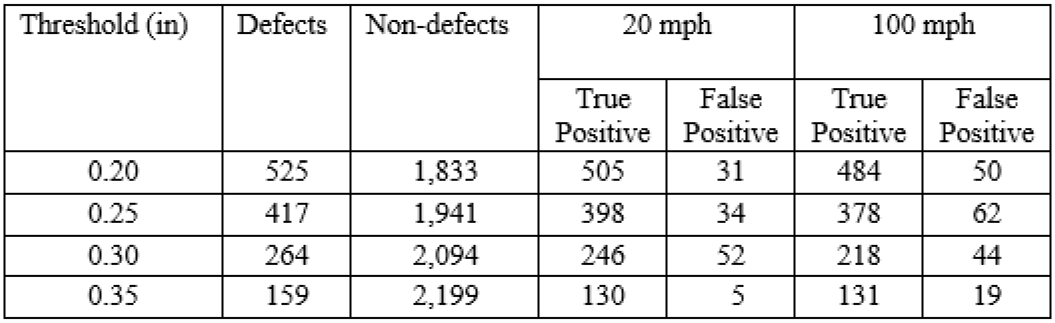
Comparing Table 9 with Table 10 shows there is improved correlation with the 31-foot data than with the foot-by-foot data. This is likely due to the difficulty in exactly aligning foot-by-foot data—a problem that does not significantly affect the 31-foot results.

5.3 ROC

ROC analysis compares the rate of correctly detecting defects to the rate of false alarms. It requires thresholds to be set that define the magnitude of defects. For example, a foot-by-foot measurement of the 31-foot mid-chord offset in the ground truth that exceeds a threshold may be considered a defect. The ROC curve plots the success rate of the TGMS in finding those defects (true positives) against the rate at which defects are reported that are below the threshold (false positives).

Table 11 shows the results of this analysis for four different thresholds. The data comes from three measurement runs at each speed in the same direction and orientation.

Table 11. Foot-by-Foot ROC Results at 20 and 100 mph

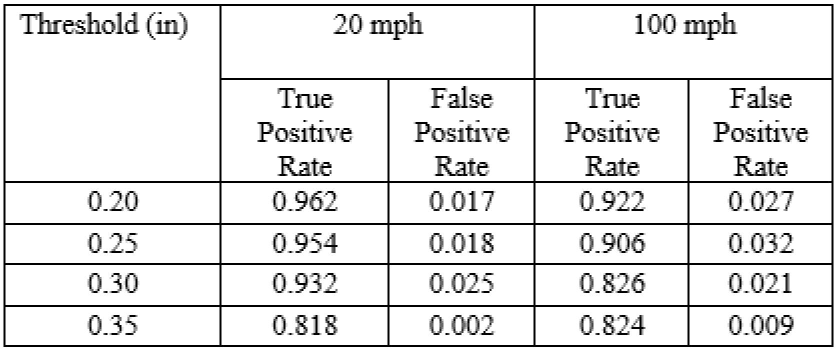


Threshold (in)	Defects	Non-defects	20 mph		100 mph	
			True Positive	False Positive	True Positive	False Positive
0.20	525	1,833	505	31	484	50
0.25	417	1,941	398	34	378	62
0.30	264	2,094	246	52	218	44
0.35	159	2,199	130	5	131	19

From Table 10, the ground truth has 525 foot-by-foot values that exceed a threshold of 0.20 inch and are considered defects. It also has 1,833 foot-by-foot values that are below the threshold. At 20 mph, the TGMS correctly identified 505 foot-by-foot values above the threshold of 0.2 inch. It also reported 31 foot-by-foot values above that threshold which, according to the ground truth, were not defects.

Table 12 shows the true and false positive rates calculated from the results in Table 10.

Table 12. Foot-by-Foot True and False Positive Rates at 20 and 100 mph



Threshold (in)	20 mph		100 mph	
	True Positive Rate	False Positive Rate	True Positive Rate	False Positive Rate
0.20	0.962	0.017	0.922	0.027
0.25	0.954	0.018	0.906	0.032
0.30	0.932	0.025	0.826	0.021
0.35	0.818	0.002	0.824	0.009

Figure 21 shows the rates in Table 11 plotted on a ROC diagram.

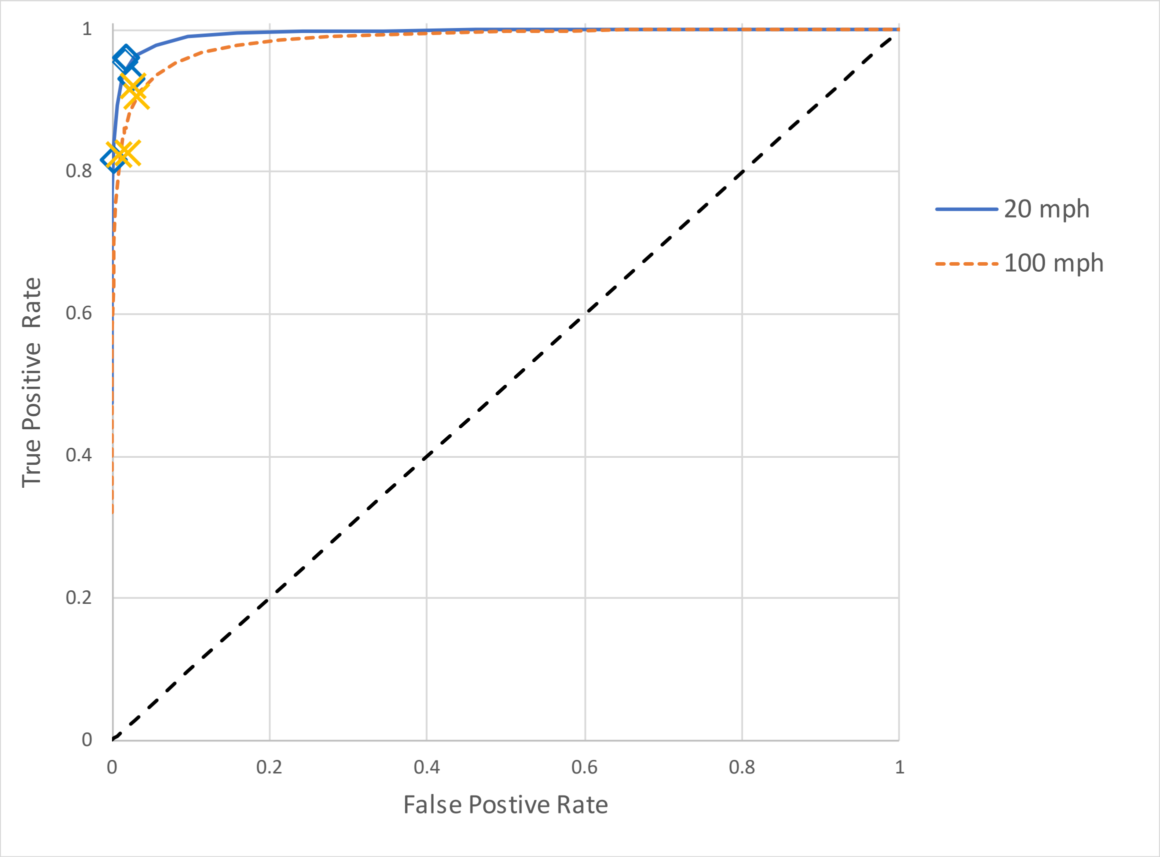


Figure 21. ROC Curves at 20 and 100 mph

Figure 22 shows the detail in the top-left corner of the ROC diagram in Figure 21.

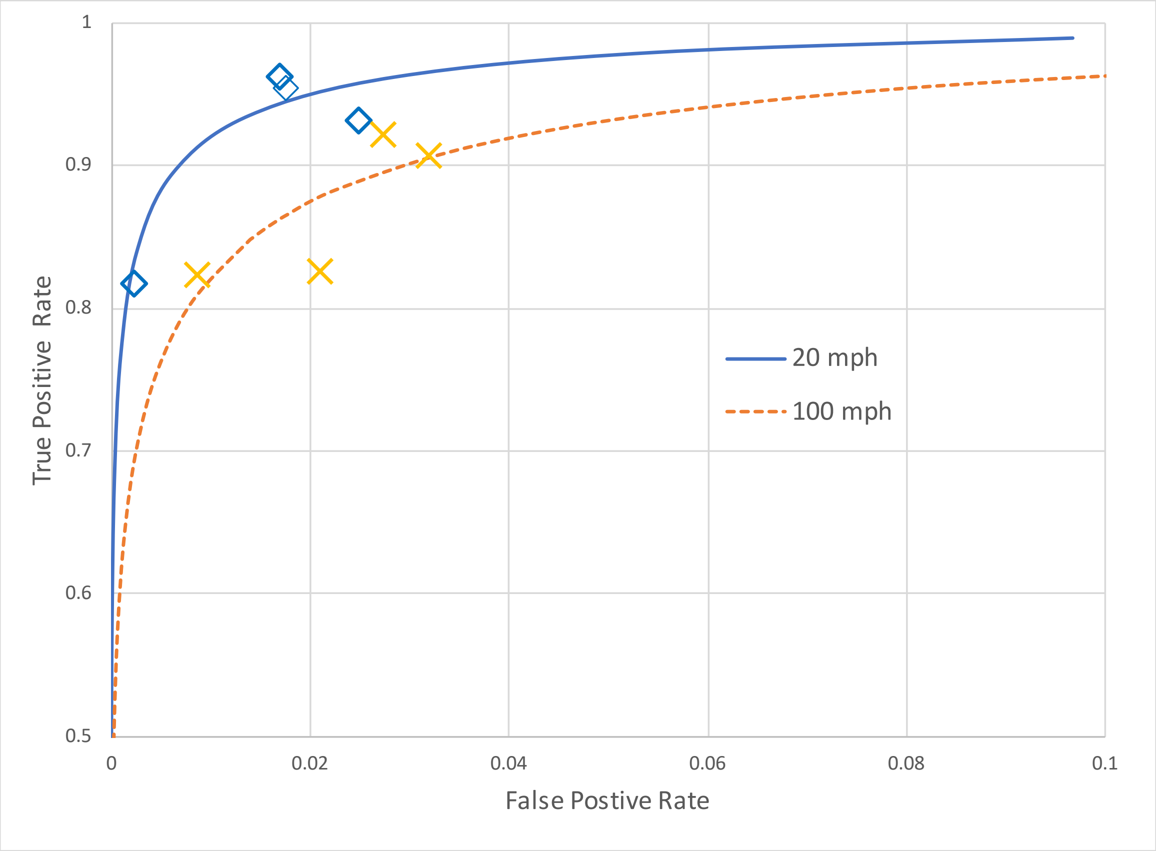


Figure 22. Detail of ROC Curves at 20 and 100 mph

Figure 21 and Figure 22 show ROC curves that fit the results at 20 and 100 mph. The TGMS performs slightly better at 20 mph compared to 100 mph. The sensitivity at 20 mph is approximately 3.7, and at 100 mph is approximately 3.2.

5.4 POD Method A

POD Method A analyzes the success of TGMS in finding defects. As with the ROC analysis, it requires thresholds to be set that define the magnitude of defects.

Figure 23 shows POD Method A results for foot-by-foot data from one TGMS measurement run at 20 mph with a threshold set to 0.2 inch.

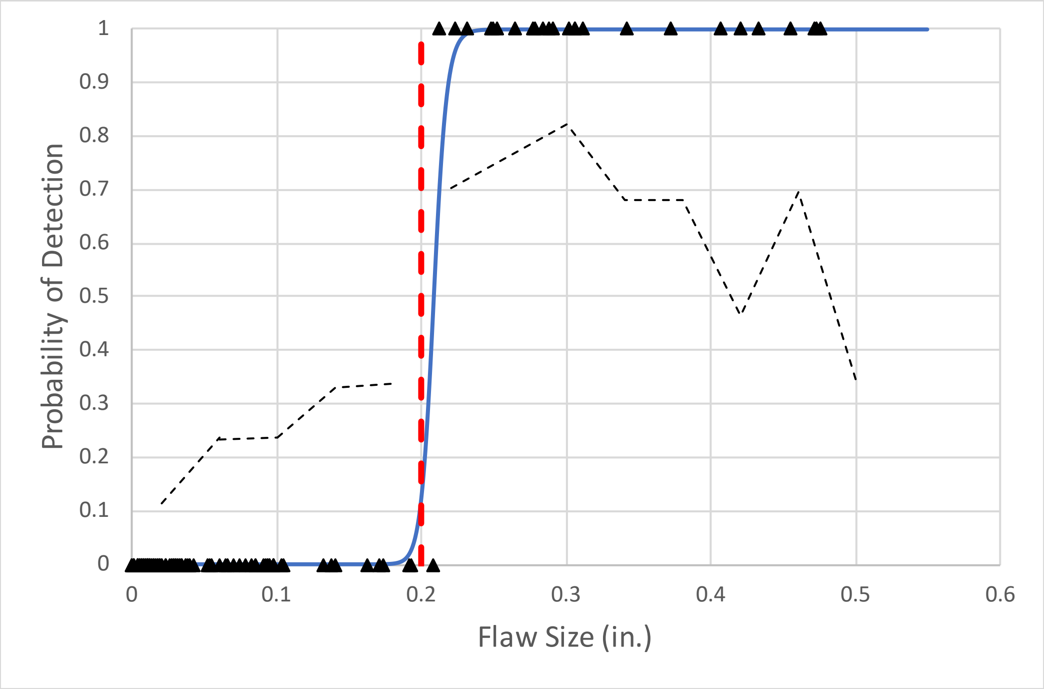


Figure 23. POD Method A Results for Foot-by-Foot Data with a Threshold of 0.2 inch – One Measurement Run at 20 mph

The triangles in the top-right of Figure 23 indicate foot-by-foot measurements that the TGMS correctly identifies as defects (true positives). The triangles in the bottom-left indicate correct rejections of defects by the TGMS (true negatives). The triangle on the x-axis to the right of the vertical threshold line is a defect that the TGMS missed (false negative). In this example there are no results in the top left of the diagram (false positives).

The POD curve fitted to the results in Figure 23 has a value of 0.9 at 0.22 inch. This means the TGMS has an estimated 90 percent probability of detecting defects (defined as being larger than 0.2 inch) when the defect size is 0.22 inch.

The 0.9 POD at 0.22 inch is an estimated value because it is based on a limited set of results. The dashed lines in Figure 23 are the 90 percent confidence limits calculated from the number of samples in each flaw size bin. For the data from one measurement run at 20 mph the lower 90 percent confidence limit at 0.22 inch is 0.70. This means the probability of detecting defects (defined as being larger than 0.2 in) when the defect size is 0.22 inch lies between 0.70 and 1.0 ninety percent of the time.

The confidence interval is wide for the largest flaw sizes due to the small number of samples. Although the POD for large flaw sizes might be expected to show better confidence in detection, the data used for Figure 23 does not prove this.

Using the data from only one measurement run at 20 mph clearly results in significant uncertainty over the results. It would not be possible to compare two different TGMS with such limited data.

The uncertainty can be reduced by supplementing the number of samples. The repeatability and reproducibility results from Section 5.1 justify combining all 12 measurement runs at 20 mph. Figure 24 shows the results of this approach.

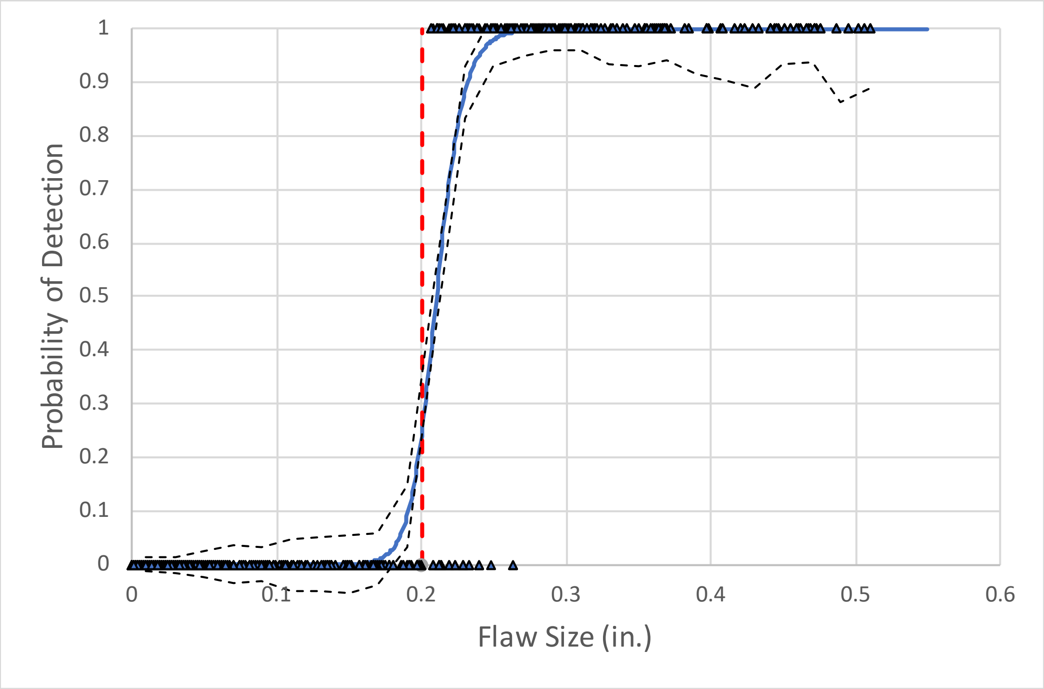


Figure 24. POD Method A Results for Foot-by-Foot Data with a Threshold of 0.2 inch – 12 Measurement Runs at 20 mph

Combining all 12 measurement runs at 20 mph gives many samples around the threshold of 0.2 inch. It results in a 0.9 POD at 0.23 inch with upper and lower 90 percent confidence limits at 0.93 and 0.83, respectively. With this greater confidence it should be possible to find any significant differences between the probabilities of detection of different TGMS.

Figure 25 shows POD Method A results for 31-foot segment data from all 12 TGMS measurement runs at 20 mph with a threshold set to 0.2 inch.

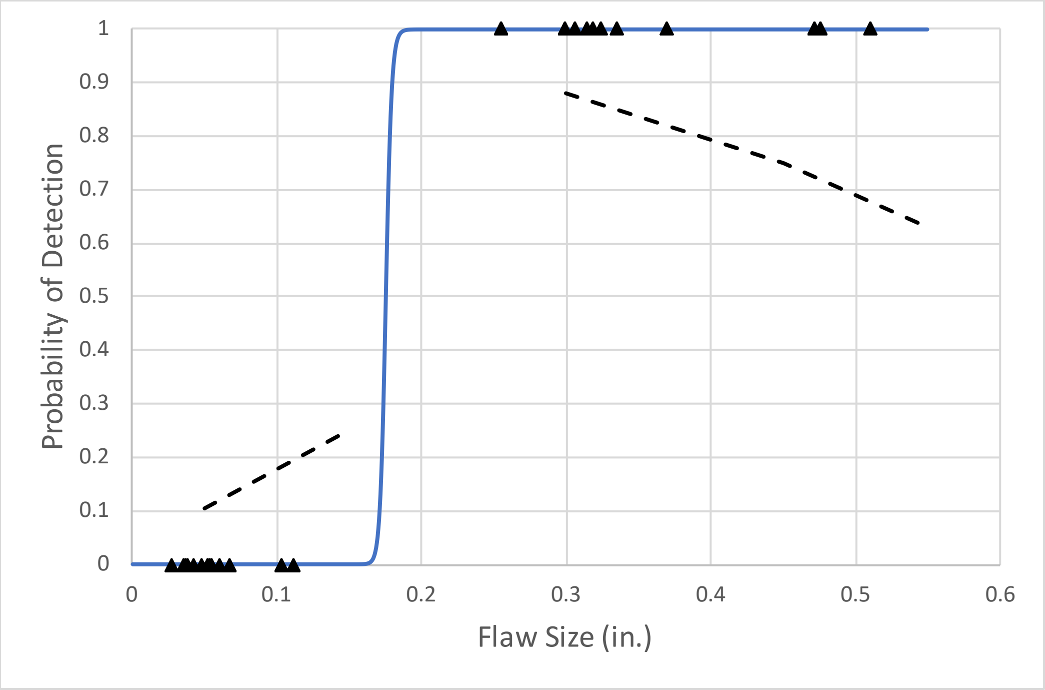


Figure 25. POD Method A Results for 31-Foot Segment Data with a Threshold of 0.2 inch – 12 Measurement Runs at 20 mph

The dashed lines in Figure 24 are 90 percent confidence limits calculated for bins 0.1 inch wide. Although all available data at 20 mph has been combined, the small sample sizes result in wide confidence bands. In the region around the threshold used to define a defect (0.2 inch), there are no samples and the confidence interval is too wide to calculate. For the same reason there is significant uncertainty about the shape of the POD curve around the threshold flaw size.

One way to achieve narrow confidence bands is to design the evaluation test with many defects around the threshold value. A minimum of 30 samples will generally give useful results.

5.5 POD Method B

POD Method B analyzes the success of TGMS in finding defects by looking at the distribution of measured results at different flaw sizes. It uses the ability of the TGMS to measure the size of track geometry flaws.

Figure 26 shows POD Method B results and 90 percent confidence limits for foot-by-foot data from all 12 TGMS measurement runs at 20 mph with a threshold set to 0.2 inch. The POD curve in Figure 26 is generated from the distribution of flaw measurements at various flaw sizes, whereas the POD curve in Figure 24 was generated from hit-and-miss results.

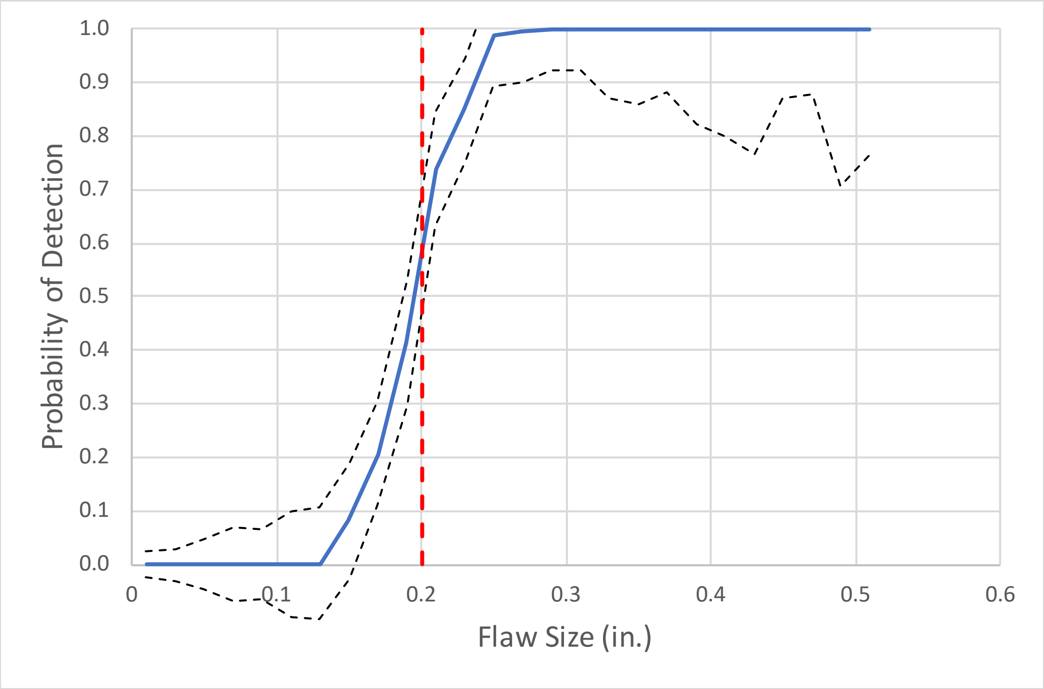


Figure 26. POD Method B Results for Foot-by-Foot Data with a Threshold of 0.2 inch – 12 Measurement Runs at 20 mph

From Figure 26, the 0.9 POD is 0.24 inch with upper and lower 90 percent confidence limits of 1.0 and 0.82, respectively. This means the probability of detecting defects (defined as being larger than 0.2 inch) when the defect size is 0.24 inch lies between 0.82 and 1.0 ninety percent of the time. This result is similar to that from the POD Method A analysis in the previous sub-section.

Another way to interpret the confidence intervals in Figure 26 is to consider the confidence in the flaw size that has a 90 percent POD. The lower 90 percent confidence limit crosses the 0.9 POD at a flaw size of 0.24 inch. This means a flaw of 0.24 inch has a 90 percent POD 9 out of 10 times the evaluation is repeated.

Figure 27 shows POD Method B results and 90 percent confidence limits for foot-by-foot data from all 12 TGMS measurement runs at 100 mph with a threshold set to 0.2 inch.

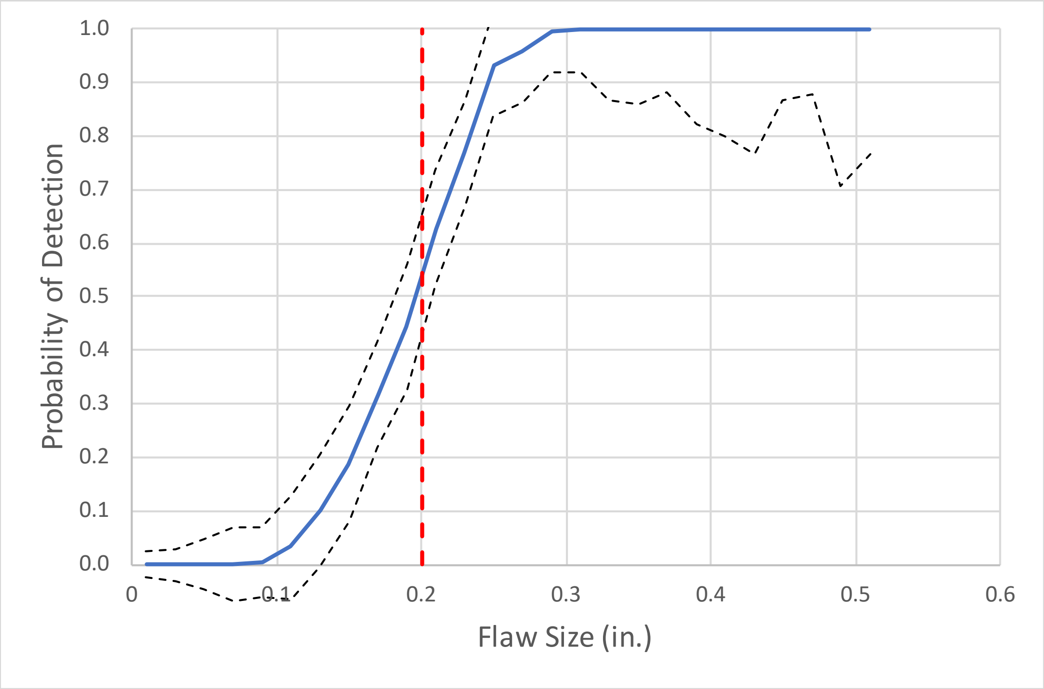


Figure 27. POD Method B Results for Foot-by-Foot Data with a Threshold of 0.2 inch – 12 Measurement Runs at 100 mph

At 100 mph the 0.9 POD is 0.25 inch, which is similar to that at 20 mph. The upper and lower 90 percent confidence limits are 1.0 and 0.84, respectively, which are also similar to those at 20 mph. Comparing Figure 26 with Figure 27 shows the POD at 20 mph lies within the 90 percent confidence limits for 100 mph and vice-versa. This means there is good confidence that the POD is the same at 20 and 100 mph.

5.6 MAPOD

MAPOD could be used to establish the TGMS probability of detection for defects beyond those available as ground truths. A computer model of the TGMS would be created, then calibrated with a limited set of data.

This method avoids the need to make measurements with the TGMS on track with defects close to or at a safety limit. The model would be validated with measurements over smaller flaws, then used to predict the POD for larger defects.

A computer model of the TGMS might have three parts:

1. Physical model – The TGMS vehicle would be modelled with a commercially available software package. The mass and inertia of the vehicle's body and main components would be represented, as would the stiffness and damping of the suspension elements. The TGMS physical instrumentation, including the geometry beam and its connection to the vehicle, would be represented the same way.
2. Electronics model – Transfer functions would describe how physical displacements, velocities, and accelerations are transformed into electrical signals by the TGMS instrumentation. This model would include the optical devices used by the TGMS to

measure displacements. Signal-to-noise ratios, variations in calibration, the effects of filtering, synchronization of inertial and optical signals, time domain to distance domain conversion, and analog to digital conversion would be included.

3. Software model – The TGMS software would then be used to process the digital data into track geometry output parameters, such as the 31-foot mid-chord offset.

The inputs to the three-part model would be track geometry flaws representing the ground truth. The outputs would be measurements of the same flaws, from which a POD could be calculated by Methods A and B.

Confidence intervals could be derived by varying the random parameters of the model and repeating the POD calculations. The random parameters could include noise in the track geometry inputs, variability in wheel-rail contact conditions, calibration variability, effects of sunlight, and other instrumentation noise.

6. Discussion and Conclusions

There are many established assessment methods available for evaluating the effectiveness of track inspection systems. Some, such as correlation, receiver operating characteristic, and POD compare the measured results to known defects that are referred to as ground truth.

Ground truth is often hard to establish. For example, it is difficult to know the size of a defect in the head of a rail so that it can be used to evaluate rail defect detection systems. The rail could be broken open and the defect size accurately measured after the test had been completed. However, although this would establish the ground truth, it would destroy the rail sample for future tests.

The track geometry ground truth at the High-Speed Adjustable Perturbation Slab Track section of the Railroad Test Track at FRA's Transportation Technology Center is established by making an accurate survey of the track's position. However, even this method is not precise and there is a possibility of slight track movement after the survey has been made. Surveying the track's position frequently will confirm the ground truth, but it will increase the cost of evaluating the effectiveness of TGMS.

When the ground truth is too difficult or costly to establish there are several alternative approaches. One approach is to compare a new track inspection system with a system already established in the industry. The established system—referred to as a "gold standard"—would have previously been verified and would have known effectiveness.

A relatively new approach when the ground truth is sparse or non-existent is called Model Assisted Probability of Detection (MAPOD). A computer model of the inspection system is used to derive the POD. Once the model has been calibrated using any available ground truths it can predict PODs for other conditions. For example, a computer model of a TGMS could simulate the dynamic response of the system as it ran over a perturbed track section. The response of the instrumentation and data analysis could be modelled, and the outputs compared with the known inputs. Once calibrated, the model could be used to calculate PODs for other types of perturbations and operating conditions.

Three approaches to evaluating track inspection system effectiveness do not require ground truth: correlation, repeatability, and reproducibility. A correlation coefficient provides the degree of linear agreement between results of an inspection system versus ground truth, or agreement between two inspection systems. Repeatability evaluates repeat performance under identical conditions. Reproducibility evaluates performance when conditions, such as operating speed, are varied. Repeatability and reproducibility of a new track inspection system can be compared with that for an established gold standard system.

The confidence in the results from evaluating a track inspection system depends on the sample size. Best practice suggests that at least 30 samples of each flaw are needed to give acceptable levels of confidence. For evaluating track inspection systems this is typically impractical or unaffordable. However, if a system can be shown to have good repeatability and reproducibility, then repeat measurements over fewer samples can improve the confidence in results.

Another way to increase confidence is to use flaw samples that are close to the threshold at which the system is being evaluated. If only a limited number of samples can be tested, then the size of flaws should be close to the threshold of interest.

This report describes several methods for evaluating the effectiveness of track inspection technologies. It illustrates these methods with the example of a TGMS. Further work is recommended to:

- **Determine acceptance criteria.** This report describes several evaluation parameters such as correlation coefficient, sensitivity, and 90 percent POD without specifying acceptable values. Further work is needed to say which parameters should be used and what the acceptable values are. The choice of parameters will likely depend on the type of inspection technology and the goal of the evaluation. Acceptable values could be derived from current, gold standard systems.
- **Develop a process for evaluating repeatability and reproducibility.** These are two common measures of effectiveness since they do not require ground truth. Further work could establish a process using analysis of variance methods to quantify repeatability and reproducibility. The process could be applied to existing inspection technologies to establish benchmarks.
- **Develop guidance for the design of evaluation tests,** showing how to use the findings from this report in practical applications. Guidance would need to be given on sample size and the number of measurements under different conditions. It would also need to show how testing objectives and costs affect the choice of evaluation method.
- **Apply MAPOD to a railroad inspection technology,** building on the approach outlined in this report. One or more MAPOD examples could be developed to demonstrate the approach. If successful, the models could be used to determine the PODs for the example systems beyond currently known conditions.

7. References

1. U.S. Department of Defense. (2004)., Nondestructive Evaluation System Reliability Assessment [MIL-HDBK-1823A]. Retrieved from http://www.statisticalengineering.com/mh1823/MIL-HDBK-1823A(2009).pdf
2. Draper, N. & Smith, H. (1981). *Applied Regression Analysis* (2nd ed.). New York: John Wiley and Sons, Inc.
3. Federal Railroad Administration. 2016). Probability of Detection Evaluation Results for Railroad Tank Cars [DOT/FRA/ORD-16/13]. Washington, DC: U.S. Department of Transportation.
4. Berens, A.P. (1989). NDE Reliability Data Analysis. ASM Handbook, Volume 17: Nondestructive Evaluation and Quality Control, p. 689-701, 1989. ASM Handbook Committee.
5. Federal Railroad Administration. (2014). Quantification of the Effectiveness of Handheld Equipment for Ground Verification of Detected Rail Internal Defects [DOT/FRA/ORD-14/07]. Washington, DC: U.S. Department of Transportation.
6. Thompson, R.B. et al. (June 2009). Recent Advances in Model-Assisted Probability of Detection. Presented at 4th European-American Workshop of Reliability of NDE.
7. Kansler, D. et al. (May 2012). An Approach to the POD Based on Real Defects Using Destructive Testing and Bayesian Statistics. In *9th International Conference on NDE in Relation to Structural Integrity for Nuclear and Pressurized Components*, pp. 191-199.
8. Meyer, R.M. et al. (September 2014). Review of Literature for Model Assisted Probability of Detection. Report prepared for U.S. Nuclear Regulatory Commission, U.S. Department of Energy, Pacific Northwest National Laboratory.
9. Smith, K. et al. (August 2006). Model-Assisted Probability of Detection Validation for Immersion Ultrasonic Application." *Review of Progress in Quantitative Nondestructive Evaluation*, *26A/26B*, 1816-1822.
10. Harding, C.A., Hugo, G.R., & Bowles, S.J. (July 2008). Application of Model-Assisted POD Using a Transfer Function Approach. In *35th Annual Review of Quantitative Nondestructive Evaluation, 28*, 1792-1799.
11. Demeyer, S. et al. (July 2011). Transfer Function Approach Based on Simulation Results for the Determination of POD Curves. In *Proceedings of 38th Annual Review of Progress in Quantitative Nondestructive Evaluation*, pp. 1757-1764.
12. Rosell, A. & Persson, G. (2013). Model Based Capability Assessment of an Automated Eddy Current Inspection Procedure on Flat Surfaces. *Research in Nondestructive Evaluation, 24*(3),154-176.
13. Aldrin, J.C. et al. (July 2008). Model-Assisted Probability of Detection Evaluation for Eddy Current Inspection of Fastener Sites. In *35th Annual Review of Quantitative Nondestructive Evaluation, 28*, 1784-1791.

14. Carboni, M., & Cantini, S. (April 2012). A “Model Assisted Probability of Detection” Approach for Ultrasonic Inspection of Railway Axles. In *Proceedings 18th World Conference on Non-Destructive Testing*, pp. 2457-2466.

15. Bode, M.D., Newcomer, J., & Fitchett, S. (July 2011). Transfer Function Model-Assisted Probability of Detection for Lap Joint Multi Site Damage Detection. In *Proceedings of 38th Annual Review of Progress in Quantitative Nondestructive Evaluation*, pp. 1749-1756.

16. Leemans, D.V., & Forsyth, D. (2004). Bayesian Approaches to Using Field Test Data in Determining the Probability of Detection. *Materials Evaluation, 62,* 855-859.

17. Jenson, F. et al. (2012). A Bayesian Approach for the Determination of POD Curves from Empirical Data Merged with Simulation Results. In *Review of Progress in Quantitative Nondestructive Evaluation, 32,* pp. 1741-1748.

18. Aldrin, J.C. et al. (2012). Bayesian Methods in Probability of Detection Estimation and Model-Assisted Probability of Detection Evaluation. In *Review of Progress in Quantitative Nondestructive Evaluation*, pp. 1733-1740. Mellville, NY: American Institute of Physics.

19. Hovey, P.W., "Using the Information in Field Service Inspections to Assess the Damage of the Aircraft and the Detection Capability of the Inspection System." In *35th Annual Review of Quantitative Nondestructive Evaluation, Volume 28*, pp. 1855-1861. July 22-25, 2008, Chicago, Illinois. American Institute of Physics, Melville, New York. AIP Conf. Proc. 1096. 2009

20. Dann, M.R., & Maes, M.A. (2012). Spatial Hierarchical POD Model for In-line Inspection Data. In M Faber, J Koehler and K Nishijima (Eds.), *Applications of Statistics and Probability in Civil Engineering,* pp. 2274-2282. London: Taylor & Francis Group.

21. Gerstman, B.B. StatPrimer Version 7.0. Retrieved from http://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf

Appendix A.
Track Inspection Technologies and Data Categories

The following tables list the categories of data generated by current inspection technologies. Table A.1 covers the inspection requirements of 49 CFR §213. Table A.2 covers technologies that address track inspection not currently required by FRA regulations.

Table A.1(a). Technologies Used for 49 CFR §213 (Part I)

Measured Characteristic	Parameter	Data Category	Technology
Drainage	Water flow obstructions	2	Imaging systems
			Theromographic systems
			LiDAR
Vegetation	Visibility obstruction	2	Imaging systems
			Theromographic systems
			LiDAR
Track Geometry	Gage, Alignment, Crosslevel, Runoff, Surface	1,2,3	Track Geometry Measurement System
			LiDAR based
Ballast	Ballast profile	1,2	Aurora system
	Fouling, moisture content, layering	4,5,6 (1,2,3 - lab testing on samples)	GPR
			NMR-TMMS, Imaging Systems
Crosstie	Defective Ties	2,3	Imaging Systems
			Laser Profiling System
			Aurora system
	Defective Ties Internally	2,3	Aurora X-Ray system
Gage restraint	ULG, LG, PLG24, GWP	4,5,6 (1,2,3 for ULG)	Gage Restraint Measurement System (GRMS)
			PTLF

Table A.1(b). Technologies Used for 49 CFR §213 (Part II)

Measured Characteristic	Parameter	Data Category	Technology
Defective rails	Internal Rail Flaw	2,3	Ultrasonic Rail Flaw
			Induction Systems
			Radiography
			Magnetic Anomaly Detection
Defective rail (contd.)	Rail Surface Defects	2,3	Imaging Systems
	Broken Rail	2	Imaging System
Rail joints	Broken or cracked joint bars, missing bolts	2,3	Joint Bar Inspection System (JBIS)
			Other Imaging systems
Torch cut rail	NA	NA	NA
Tie Plate	Broken or missing tie plates	2	Imaging Systems
	Metal objects present	2	Imaging Systems
Missing or Defective Fasteners		2	Imaging Systems, GRMS
Turnouts and track crossings generally	Fastenings, obstructions	2	Imaging Systems, laser profile
	Flangeway width	1,2,3	Imaging Systems, laser profile
Switches	Stock rail seating, braces	2	Imaging Systems, laser profile
	Switch points	2	Imaging Systems, Laser Profiling Systems
	Wheel thread stock rail contact	2	Imaging Systems, Laser Profiling Systems
	Heel bolts	2	Imaging Systems
	Stand, connecting rod, throw lever	NA	NA
	Switch position indicator visibility	2	Imaging Systems

Table A.1(c). Technologies Used for 49 CFR §213 (Part III)

Measured Characteristic	Parameter	Data Category	Technology
Frogs	Flangeway depth	1,2,3	Imaging Systems, laser profile
	Frog point is chipped, broken, or worn	2	Imaging Systems, laser profile
	Tread portion of a frog casting wear	1,2,3	Imaging Systems, laser profile
Spring and self-guarded frogs	As the applicable portions of "Frogs"	2	Imaging Systems, laser profile
Frog guard rails and guard faces and gage.	Guard check gage, guard face gage	1,2,3	Imaging Systems, laser profile

Table A.2. Technologies Not Used for 49 CFR §213

Vertical Deflection		1,2,3 (4,5,6)	Mrail (VTDMS)
Rolling Contact Fatigue		2,3	Eddy Current, Imaging systems
Accelerations and Rail Impact Forces		4,5,6	VTI Monitor
			Ride quality
Wheel Rail Forces		2,3	Instrumented Wheelsets (IWS)
Wheel Loads		2,3	WILD (rail circuits)
Corrugation		1,2,3	Rail Corrugation Measurement System
Grade Crossing (Profile and Layout)		2,3	LiDAR
			Imaging System
Track Circuits		2	Signaling Inspection Systems
Clearance		1,2,3	Theromographic systems
			LiDAR
Catenary Wire Geometry		1,2,3	Catenary Inspection Systems
Pantograph (Acceleration/Voltage/Force)		1,2,3	Catenary Inspection Systems
Rail Profile		1,2,3	Rail Profile System
Rail Cant		1,2,3	Rail Profile System
PTC Asset			
Third Rail (present or not)		2	Rail Profile System

Abbreviations and Acronyms

Abbreviation or Acronym	Name
FRA	Federal Railroad Administration
FMA	Full Model-Assisted
MAPOD	Model-Assisted Probability of Detection
PDF	Probability Density Function
POD	Probability of Detection
RTT	Railroad Test Track
ROC	Receiver Operating Characteristic
TGMS	Track Geometry Measuring System
XFM	Transfer Function Method

Glossary

a – Physical dimension of flaw or target.

â, (a-hat) – Measured response of an inspection system to a target of size a.

a_{dec} – Dimension of target at the decision threshold.

Accuracy – The probability of either true positive or true negative.

Bayesian method – An analysis method that updates a hypothesis when new data becomes available.

Calibration – Process of determining the performance parameters of a system by comparing them with measurement standards.

Confidence – The long run frequency of being correct, e.g., a 95 percent confidence value for a_{90} will be greater than the true a_{90} in 95 percent of similar experiments.

Decision threshold – Value of a-hat which the signal is interpreted as a hit and below which the signal is interpreted as a miss.

Defect – A flaw that is to be rejected, i.e., one that does not meet acceptance criteria,

Discrete variable – Measurement variable having discrete levels or categories

False negative – Inspection system response interpreted as having failed to detect a flaw when it is present at the inspection location,

False positive – Inspection system response interpreted as having detected a target when none is present at the inspection location,

Flaw – A type of discontinuity that must be investigated to see if it should be rejected,

Ground Truth – Actual data ascertained by direct observation.

Hit – Affirmative inspection system response (detection) when flaw is present,

Inspection threshold – Smallest value of a-hat the inspection system records; the value of a-hat below which the signal is indistinguishable from noise,

Inspection system – Ensemble that can include hardware, software, materials, and procedures intended for application of a specific inspection technology,

Likelihood ratio method – Method for producing confidence bounds on hit-or-miss POD(a) curves.

Miss – Inspection system response interpreted as not having detected a target when one is present at the inspection location,

Repeatability – Variation among repeated runs over the same object or area being tested with the same conditions; e.g., comparison of multiple TGMS runs with the same system over the same track configuration at the same speed, direction, and car orientation,

Reproducibility – Variation among repeated runs using the same equipment and different conditions; e.g., comparison of multiple TGMS runs with the same system over the same track configuration at the different speed, direction, and car orientations,

Saturation – Value of a-hat as large or larger than the maximum output of the system or the largest value of a-hat that the system can record.

Sensitivity – Probability of a true positive.

Specificity – Probability of a true negative.

Target – Object of an inspection; e.g. a crack, flaw, defect, anomaly, or measurement discontinuity.

$t_{confidence}$ – A parameter from student's t-distribution to give the confidence interval based on the size and standard deviation of a population.

True negative – Inspection system response interpreted as having failed to detect a target when none is present at the inspection location.