Bennetts, J., Webb, G., Denton, S., Vardanega, P., & Loudon, N. (2018). Quantifying uncertainty in visual inspection data. In N. Powers, D. Frangopol, R. Al-Mahaidi, & C. Caprani (Eds.), *Maintenance, Safety, Risk, Management and Life-Cycle Performance of Bridges: Proceedings of the Ninth International Conference on Bridge Maintenance, Safety and Management (IABMAS 2018), 9-13 July 2018, Melbourne, Australia* (pp. 2252-2259). (Bridge Maintenance, Safety and Management). CRC Press.

Peer reviewed version

[Link to publication record in Explore Bristol Research](#)
PDF-document

This is the author accepted manuscript (AAM). The final published version (version of record) is available via Taylor & Francis at <https://www.crcpress.com/Maintenance-Safety-Risk-Management-and-Life-Cycle-Performance-of-Bridges/Powers-Frangopol-Al-Mahaidi-Caprani/p/book/9781138730458> . Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available: <http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/>

Quantifying Uncertainty in Visual Inspection Data

J. Bennetts

*Department of Civil Engineering
University of Bristol, Bristol, United Kingdom*

G. Webb, S. Denton

*Civil, Bridge & Ground Engineering
WSP Ltd, London & Bristol, United Kingdom*

P. J. Vardanega

*Department of Civil Engineering
University of Bristol, Bristol, United Kingdom*

N. Loudon

*Safety, Engineering and Standards
Highways England, Bedford, United Kingdom*

ABSTRACT: Visual inspection is the most common form of condition monitoring used by bridge owners. Information derived from visual inspection data is commonly used to indicate the performance of bridge stocks and inform bridge management decisions. However, several studies have highlighted that the inherently subjective nature of the methods used to record this data can result in uncertainty, due to differences between different inspectors' perceptions of the severity and extent of defects. It is important for asset managers to understand the nature of this uncertainty and the implications for decision making. This paper reports the results of a study which compared scoring of bridge defects by pairs of independent inspectors across 200 bridge structures on England's strategic road network. A sample of 200 structures was selected to be representative of Highways England's stock with regard to, *inter alia*, *age*, *condition* and *structural form*. Routine Principal Inspections for these sample structures, undertaken every six years by the relevant maintaining agents, were also attended by inspectors from WSP Ltd, with defects scored independently by each inspector. The results of these comparisons were used to derive an empirical profile of the uncertainty in different individual defect severity and extent scores. Statistical methods were then used to derive empirical probability density functions for the values of bridge and stock level condition metrics according to the widely adopted *Bridge Condition Indicator* system. The reported results highlight trends in the reliability of individual defect scores and the impact of uncertainty on commonly used performance metrics.

1 INTRODUCTION

1.1 Introduction

Visual Inspection is the primary form of condition monitoring that is currently operated by bridge owners in the UK (e.g. McRobbie et al. 2015, Bennetts et al. 2016) and can be considered as a form of 'Damage Detection' in the wider framework of structural health monitoring systems (Webb et al. 2015). However, studies of the reliability of visual inspection have shown that the recorded data can be subject to significant uncertainty due to human factors (e.g. Moore

et al. 2001, Lea & Middleton 2002). Moore et al. (2001) compared data from 49 different inspectors across 10 inspection tasks, comprising routine inspection of seven bridges and in-depth inspection of three bridges on the disused Pennsylvania Turnpike highway in the USA. Moore et al. (2001) concluded that the experience and physical capabilities of inspectors, and the environment in which inspections were undertaken had significant influence on the reliability of the inspection data recorded.

The data collected during visual inspections is increasingly used to inform decision-making, both at an individual structure level and at a strategic, whole

stock level (e.g. British Standards Institution 2014, UK Roads Liaison Group 2016). It is important to understand the reliability of the underlying data and its implications for the reliability and utility of derived *Condition Performance Indicators*, on which strategic decisions may be based.

1.2 Inspection Data

The industry practice for visual inspection of UK highway bridges is for a cycle of ‘General Inspections’ every 2 years, with ‘Principal Inspections’ every 6 years (Highways Agency 2007). During these inspections any defects observed on the structure are noted and photographed and recorded digitally later on. These defects are assigned to individual components in the bridge inventory and, for each defect, the defect type is recorded along with a severity code (the options for which are set out in Table 1) and an extent code from *SA* to *SE* (Table 2). Severity codes are categorised into different severity types, according to the defect type (Table 3), and are suffixed with an *S* if they are considered to present an immediate safety concern. Defects observed during inspections of bridges on the Highways England network are entered into the Structures Management Information System (SMIS) database.

Table 1: Mapping of Severity Codes to Severity Scores

D	Severity Codes				Severity Score
	P	A	X		
D1	P1	A1	X1, X2		1
		A2			1.1
D2	P2	A3	X3		2
D3	P3	A4	X4		3
D3S			X4S		3.1
D4	P4		X5		4
	P4S				4.1
D4S					4.2
D5	P5				5

1.3 Condition Performance Indicators

In order to monitor the overall condition of individual structures or stocks of bridges, it is necessary to derive a measure of ‘condition’ from recorded defect information and the structure inventory. The *Bridge Condition Indicator* was developed for this purpose by WS Atkins Ltd (Sterritt 2002) for the County Surveyors’ Society (CSS, now the Association of Directors of Environment, Economy, Planning and Transport, ADEPT) in the UK, and is adopted as a metric to score the condition of a bridge asset, or a population of assets, by most highway bridge owners in the UK. An understanding of the way in which the *Bridge Condition Indicators*, BCI_{Ave} and BCI_{Crit} are calculated is critical to understand the implications of the results presented in this paper. The process that

is used by Highways England in SMIS is slightly different from the original CSS process (Sterritt 2002) and is set out as follows:

Table 2: Mapping of Extent Codes to Extent Scores

Extent Code	SA	SB	SC	SD	SE
Extent Score	0	0	0.1	0.3	0.7

Table 3: Defect Severity Types

Defect Severity Type	Description
D	Damage causing defects
P	Paint coating and protective system defects
A	Appearance related defects
X	Defects affecting adjacent areas

1. Defects in SMIS are recorded against individual components. Each defect score comprises a severity code and an extent code. These codes are mapped to numerical scores as shown in Tables 1 and 2.
2. For each component the defect with the highest severity score is selected. If there are multiple defects with the same severity then the defect with the highest extent score is selected.
3. All other defects on that component are disregarded.
4. For each component type, only defects with the same severity as the most severe are included.
5. A weighted average extent score based on the total size of all the components of that type is calculated.
6. The extent and severity scores are added to calculate an Element Condition Score (ECS) for each component type, on a scale from 1 to 5.7.
7. The Element Consequence Factor (ECF) for each component type is calculated from the ECS and the component’s importance as shown in Table 4.

Table 4: Mapping of Element Consequence Factor (ECF)

Component Importance	Element Consequence Factor (ECF)
Very High	0.0
High	$0.3 - [(ECS - 1) \times 0.3/4]$
Medium	$0.6 - [(ECS - 1) \times 0.6/4]$
Low	$1.2 - [(ECS - 1) \times 1.2/4]$

8. The Element Condition Index (ECI) is found by subtracting the ECF from the ECS. The ECI cannot take a value below 1.0.

Table 5: Mapping of Element Importance Factor (EIF)

Component Importance	Element Importance Factor (EIF)
Very High	2.0
High	1.5
Medium	1.2
Low	1.0

9. The Element Importance Factor (EIF) is determined based on the importance of the component type, as shown in Table 5.
10. The Average Bridge Condition Score BCS_{Ave} is calculated by averaging the ECI values, weighted by the EIF for each component type.
11. The Critical Bridge Condition Score BCS_{Crit} is found by taking the maximum ECS for all ‘Very High’ Importance components.
12. BCS_{Ave} and BCS_{Crit} are both scores between 1 (best condition) and 5 (worst condition). These are mapped using an assigned quadratic function to give Bridge Condition Indicators BCI_{Ave} and BCI_{Crit} between 0 (worst condition) and 100 (best condition). The mapping functions are shown in Equations 1 and 2.

$$BCI_{Ave} = 100 - 2[BCS_{Ave}^2 + 6.5BCS_{Ave} - 7.5] \quad (1)$$

$$BCI_{Crit} = 100 - 2[BCS_{Crit}^2 + 6.5BCS_{Crit} - 7.5] \quad (2)$$

To calculate stock level BCI scores these steps are repeated for all bridges in the stock and an average is taken, weighted by each structure’s deck area.

2 METHODOLOGY

2.1 Random Sample

A stratified sample of 200 bridges was selected by the authors from Highways England’s stock for inclusion in this research. This sample was selected to be representative of the whole bridge stock, taking into account the following attributes: *Condition*; *Deck Type*; *Construction Type*; *Road Class*; *Age Group*; *Structure Type*; *Maintenance Region*; and *Type of Tensioning*. For example, the proportion of bridges in the sample with post-tensioning was selected to be the same as the proportion of post-tensioned structures in Highways England’s whole stock of bridges.

The sample was selected computationally by randomly choosing several thousand possible samples of 200 from the list of structures due to be inspected over the duration of the project, and then selecting the random sample which was the best fit for the whole stock. The quality of fit, Q_{sample} , was evaluated as the sum of the mean of the squares of the differences in the proportions of each attribute value in each attribute, with the lower numbers indicating a better fit as shown in Equation 3.

$$Q_{sample} = \sum_{attributes} \frac{\sum_{attribute\ values} [P_{stock} - P_{sample}]^2}{N_{attribute\ values}} \quad (3)$$

2.2 Independent Scoring of Defects

For each of the 200 bridges in the sample, a WSP bridge inspector attended the site to independently inspect and record a sample of the defects present on the structure. For the large majority of the bridges in the sample, this coincided with the Service Providers’ Principal Inspection. However, for a small number of the sample bridges, the independent inspectors visited the structure during a separate shift. For each visit, the inspectors recorded independent defect type, severity and extent information for a subset of the defects on the structure. These were limited to those visible at close quarters on the structure using the available access. 1373 individual defects were independently recorded by WSP’s inspectors, of which 988 could be directly compared to those entered by the Service Providers’ inspectors.

3 DATA ANALYSIS

3.1 Reliability of defect recording

For 385 (28%) of the defects recorded by WSP’s inspectors, it was not possible to identify an equivalent defect in the Service Providers’ Principal Inspection reports. This indicates variability in the reliability of defect identification by inspectors during visual inspections. Across the 988 independently recorded defects that could be directly compared to those recorded by the Service Providers as part of their Principal Inspection, 21% agreed on the assigned defect type, defect severity and extent. Overall, for 58% of the compared defects there was agreement on the type (e.g. *Map cracking*) of defect assigned, and agreement on the overall class of defect (e.g. *Cracks in concrete or masonry*) for 78%.

The level of agreement of defect types varied between different classes of defects, as shown in Figure 1. Particularly low reliability of defect class assignment was found within the *Cracks in concrete or masonry* and *Loss of concrete or masonry* defect classes. Investigation into the causes of these inconsistencies suggests that the process allows too many options of defect type within these classes. Consequently accuracy has been compromised in exchange for the availability of greater precision in defect type assignment. Furthermore, it was noted that while some defect types can only be confirmed through laboratory testing, the codes for these defects could, and had been, applied based only on visual inspection.

Similar results were found for the probabilities of the two inspectors recording the same defect severity or extent, with agreement of defect severity assignment for 53% of the compared defects, and agreement on defect extent for 59%. The same defect severity and extent were applied by the two inspectors for 34% of defects compared.

Figure 1: Comparison of defect type allocated by Service Provider inspector and WSP inspector for the defects included in this study, grouped by defect class

3.2 Element Condition Scores

For each recorded defect the extent and severity scores have been added to calculate an element condition score (ECS), giving a single measure to allow a comparison to be made between the two different inspectors (i.e. one each from WSP Ltd and Highways England’s Service Provider for the maintenance area). These comparisons are presented in Figure 2, with each of the 988 comparisons represented as a blue marker. Due to the discrete nature of defect scores there are many overlapping markers. Red shading has therefore been added to indicate the density of markers at any point on the plot. Also shown, at the top and right of the plot, are histograms and gaussian kernel density estimations to provide an indication of the distribution of the scores from each inspector. While the two inspectors assigned the same, or a similar, defect severity and extent for many of the observed defects, there were substantial numbers of defects with a difference in ECS of more than 1. There were considerable discrepancies between the two inspectors’ assessments of the severity of some defects, suggesting that the condition data for some structures could be markedly different, depending which individual inspector conducted the inspection.

3.3 Trends in the reliability of defect scoring

3.3.1 Decision Trees

Optimal decision trees allow presentation of the categorical factors that most influence a target variable (Quinlan 1986). To explore the factors affecting the reliability of defect scoring, a decision tree was generated to find the most informative multi-factor trends in the variation of defect score assignment (ECS) between two inspectors. For each compared defect the absolute difference between the ECS recorded by the WSP inspector and the Service Provider was calculated and banded into five groups as evenly as possible from those which were the same, to those which were the most different. Each pie chart in the decision tree has been drawn to represent the distribution of the ECS difference bandings within the sub-population below that point in the tree structure (Figure 3). The top ‘All comparisons’ pie-chart represents the distribution of ECS difference in the full set of defect comparisons.

At each splitting point, the optimum attribute has been selected to partition the data in the way which provides the most information. The reduction in Shannon entropy (Shannon 1948) has been used to measure the information gained by a partition. For this measure, a perfect partition would be an attribute that perfectly split the structures into each of the ECS difference bands, whereas the worst possible partition would result in sub-populations that each had an even distribution of ECS difference within them. The size of each node’s pie-chart has been drawn inversely proportional to its entropy and larger pie-charts have been plotted further from their parent node. This allows the most informative partitions to be readily identified. The following attributes were given to the algorithm from which it selected the optimum tree topology: *Component*, *WSP Class*, *WSP Extent*, *WSP Severity*, *Structure type*, *Construction type*, *Deck type*, *Structure use*, *Structure condition band*, *Distance from coast*, *Region*, *Age group*. Where:

WSP Class, WSP Extent & WSP Severity are the defect class, extent and severity, as recorded by the WSP inspector on site.

Age group splits the bridges into five bands from Youngest to Oldest with the same number of bridges in each band.

Structure condition band splits the bridges into five bands by their BCI_{Ave} condition score, from Best to Worse.

Nodes, and corresponding branches, have only been drawn if they represent at least 10 comparisons. The depth of the tree was limited to 3 partitions.

3.3.2 Trends

Figure 3 shows the most informative variables which affected the variation in ECS between two inspec-

Figure 2: Comparison of element condition score allocated by Service Provider inspector and WSP inspector for the 988 defects included in the study. Red shading has been added to indicate the density of markers at any point on the plot.

tors. The *defect extent* had a significant influence. There was a much greater variation in ECS for defects with greater extent, than for those with lesser extents. This suggests that it is much harder to judge the severity of extensive defects. *Component type* was an important factor. Components such as *Road Vehicle Restraint Systems* and *Supports* were seen to exhibit much greater variation in scoring than components such as *Crossheads*, *Expansion joints* and *Wingwalls*. There was also a large degree of variation between different maintenance regions, suggesting that inspection guidance and practice are not consistent across the country. Structure details such as *age*, *structural form*, *construction material*, *road class*, and overall *structure condition* have a lesser influence on the reliability of defect scoring.

4 IMPLICATIONS OF DEFECT RATING VARIABILITY ON CONDITION METRICS

Given the variability of the visual inspection data collected by the owners of highway bridges, it is important to understand how this uncertainty affects derived metrics, which are used to inform strategy and decision making. The *Bridge Condition Indicator* calculation process is convoluted, and includes some non-linear steps, which means that it is difficult to derive directly the uncertainty in the resulting metrics from the uncertainty in the underlying data. Instead, Monte

Carlo simulations have been used to derive this uncertainty empirically.

4.1 Monte Carlo Simulations

The 988 individual defect comparisons have been used to generate a table of probabilities mapping, for each extent code allocated by one inspector, the probabilities that a second inspector would allocate each of the possible extent codes. To avoid an implicit assumption that one inspector was ‘correct’, each observation was counted twice: once with the WSP inspector as Inspector 1 and once with the Service Provider’s inspector as Inspector 1. This exercise was repeated for the severity scores. The resulting probabilities are presented in Table 6 and Table 7.

The probabilities have been derived from data from the sample of 200 structures only, however, since the structures were selected as a representative sample of the entire stock, it has been assumed that they are applicable to data from all structures in the stock. It is noted that no defects with a severity score of 4.1 or 5 were observed, by either inspector, on any of the structures in the sample. In the absence of additional information it has been assumed, for this analysis, that there was no variability in the scoring of defects with these severities. This will not significantly affect the results since defects with a severity score of 4.1 or 5 represented fewer than 1% of the approximately

Figure 3: Decision Tree, showing the factors that most affect defect scoring agreement between inspectors. (The blue number next to each pie chart denotes the number of comparisons that are included in that sub-population. The red text on each branch of the tree represents the attribute that the dataset has been partitioned by at that level, and the black text represents the value of the attribute.)

Figure 4: Standard deviation of condition scores obtained for each individual structure during Monte Carlo analysis

500,000 currently valid defects in the SMIS database.

The probability tables were used to vary the severity and extent scores for each of the currently valid defects in the SMIS database. These altered scores were then used to calculate new BCI_{Ave} and BCI_{Crit} scores for each bridge. This analysis was repeated for 1000 simulations.

Table 6: Probability table for defect extent codes

		Extent Code Allocated By Inspector 1			
		SB	SC	SD	SE
Probability of	SB	0.735	0.436	0.266	0.312
Inspector 2	SC	0.196	0.419	0.293	0.219
Allocating	SD	0.043	0.105	0.351	0.177
Extent Code	SE	0.026	0.040	0.090	0.292

4.2 Uncertainty in BCI scores for individual structures

This analysis gives a set of plausible alternative values for the BCI scores for each individual structure. The level of uncertainty in BCI scores calculated by this method is dependent on a structure's inventory and defect assignments. It is therefore not possible to derive a definitive probability density function around a mean that would be valid for all structures. Instead, the standard deviations of each of these sets of alternative values have been plotted in Figure 4 for each of the approximately 7100 bridges on Highways England's network for which data was available. This shows that there is considerable uncertainty in the condition indicator scores for individual bridges; the BCI_{Ave} scores for most bridges have standard deviations of between 2 and 4 points, and the BCI_{Crit} scores for most bridges have standard deviations of between 10 and 15 points. The larger degree of uncertainty for BCI_{Crit} scores is due to them being governed by a very small number of defects on key structural components. In contrast, the BCI_{Ave} scores in-

clude defects from multiple component types and are therefore much less sensitive to variations in the scoring of individual defects.

In considering the implication of these results, it is useful to consider approximately what magnitude of change in a structure's condition score between consecutive inspections would be considered statistically significant. For a 95% confidence level, it is typically assumed that a change of two standard deviations would be required for the result to be considered significant. It follows that an apparent change in BCI_{Ave} score between inspections of less than about 6 points cannot be considered statistically significant. The larger uncertainty in BCI_{Crit} scores means that a change of around 25 points would be required to be significant.

A further consequence of this variation in individual structure BCI scores is that ranking structures by score (especially BCI_{Crit}) in order to prioritise maintenance would produce unreliable results. Thus, it would be undesirable to implement contractual conditions or targets that encourage prioritising interventions purely by structure condition scores.

Figure 5: Empirical probability density function for the value of the BCI_{Ave} (top) and BCI_{Crit} (bottom) score for all structures in the stock

4.3 Uncertainty in BCI scores for stocks of bridges

Given that the computed variation in defect scores does not exhibit significant skew, convergence to the mean dictates that the uncertainty is reduced for metrics that include more data, as is the case for stock-

Table 7: Probability table for defect severity codes

		Severity score allocated by Inspector 1								
		1	1.1	2	3	3.1	4	4.1	4.2	5
Probability of Inspector 2 allocating Severity score	1	0.46	0.23	0.10	0.05	0.04	0.00	0.00	0.00	0.00
	1.1	0.03	0.31	0.01	0.01	0.00	0.00	0.00	0.00	0.00
	2	0.39	0.31	0.66	0.36	0.33	0.16	0.00	0.33	0.00
	3	0.11	0.15	0.20	0.50	0.25	0.53	0.00	0.33	0.00
	3.1	0.00	0.00	0.01	0.01	0.25	0.01	0.00	0.33	0.00
	4	0.00	0.00	0.01	0.06	0.04	0.29	0.00	0.00	0.00
	4.1	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
	4.2	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.00
	5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

level BCI scores. The sets of plausible scores for each individual structure have been combined into a set of plausible stock level benchmarks by taking the mean of the structure values in each set, weighted by each structure's deck area. The results have been plotted in Figure 5. Despite the variation in individual defect scores, the average BCI_{Ave} and BCI_{Crit} scores for the entire stock of structures show very little deviation, with computed standard deviations of 0.06 and 0.20 respectively.

5 CONCLUSIONS

The following concluding remarks are made:

- Significant uncertainty was found in the classification and grading of individual defects during the visual inspection process.
- Defect classifications were particularly inaccurate for common concrete defects, where a large range of possible defect types is available.
- Defect grading by severity and extent was seen to be most unreliable for defects with high extents, and for component types such as *Vehicle Restraint Systems*.
- The degree of variation in defect grading between inspectors differed between different regional inspection teams, suggesting variation in practice.
- The industry standard *Bridge Condition Indicator* metric was shown to be sensitive to uncertainty in the underlying defect data, with the BCI_{Crit} metric particularly unreliable for comparing the performance of individual structures.
- Use of the *Bridge Condition Indicator* system at stock-level was shown to be considerably more reliable, with low levels of uncertainty.

DATA AVAILABILITY STATEMENT

The authors thank Highways England and Lockheed Martin for use of portions of the Structures Manage-

ment Information System database as part of this research project. New experimental data collected during this study is owned by Highways England.

ACKNOWLEDGEMENTS

The authors would like to thank Highways England, Professor Colin Taylor and the Engineering and Physical Sciences Research Council (**Grant No. EP/G037353/1** Industrial Doctorate Centre: Systems).

REFERENCES

- Bennetts, J., Vardanega, P. J., Taylor, C. A., & Denton, S. R. 2016. Bridge data - What do we collect and how do we use it? In Mair, R. J., et al. (Eds.), *Transforming the Future of Infrastructure through Smarter Information: Proceedings of the International Conference on Smart Infrastructure and Construction, 27-29 June 2016*, Cambridge, pp. 531–536. London, UK: ICE Publishing.
- British Standards Institution 2014. *BS ISO 55000:2014 Asset management: Overview, principles and terminology*. London, UK: British Standards Institution.
- Highways Agency 2007. BD 63/07 Inspection of Highway Structures. In *Design Manual for Roads and Bridges*, Volume 3. London, UK: HMSO.
- Lea, F. & Middleton, C. 2002. Reliability of Visual Inspection of Highway Bridges, CUED/D-STRUCT/TR.201. Cambridge, UK: University of Cambridge - Department of Engineering.
- McRobbie, S. G., Wright, M. A. & Chan, A. 2015. Can technology improve routine visual bridge inspections? *Proceedings of the Institution of Civil Engineers - Bridge Engineering* 168(3), 197–207.
- Moore, M., Phares, B., Graybeal, B., Rolander, D., & Washer, G., 2001. Reliability of Visual Inspection for Highway Bridges, Volume I: Final Report, FHWA-RD-01-020. Washington DC, USA: US DOT, Federal Highway Administration.
- Quinlan, J. 1986. Induction of decision trees. *Machine learning* 1(1), 81–106.
- Shannon, C. E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27(3), 379–423.
- Sterritt, G. 2002. Bridge Condition Indicators, Volume 1. Technical report, WS Atkins for CSS Bridges Group, London, UK.
- UK Roads Liaison Group 2016. *Well-Managed Highway Infrastructure: A Code Of Practice*. London, UK: Department for Transport.
- Webb, G. T., Vardanega, P. J. & Middleton, C. R. 2015. Categories of SHM Deployments: Technologies and Capabilities. *Journal of Bridge Engineering (ASCE)* 20(11), 04014118.