Accuracy of visual inspection of flood defences

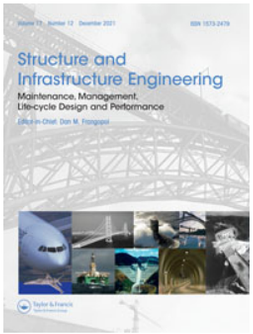
Klerk, W. J.; Kanning, W.; Kok, M.; Bronsveld, J.; Wolfert, A. R.M.

DOI
10.1080/15732479.2021.2001543

Publication date
2021

Document Version
Final published version

Published in
Structure and Infrastructure Engineering



Taylor & Francis
Taylor & Francis Group

Structure and Infrastructure Engineering
Maintenance, Management, Life-Cycle Design and Performance

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/nsie20

Accuracy of visual inspection of flood defences

W. J. Klerk, W. Kanning, M. Kok, J. Bronsveld & A. R. M. Wolfert

To cite this article: W. J. Klerk, W. Kanning, M. Kok, J. Bronsveld & A. R. M. Wolfert (2021): Accuracy of visual inspection of flood defences, Structure and Infrastructure Engineering, DOI: 10.1080/15732479.2021.2001543

To link to this article: https://doi.org/10.1080/15732479.2021.2001543

Published online: 17 Nov 2021.

STRUCTURE AND INFRASTRUCTURE ENGINEERING
https://doi.org/10.1080/15732479.2021.2001543

Accuracy of visual inspection of flood defences

W. J. Klerk[a,b], W. Kanning[a,b], M. Kok[a], J. Bronsveld[c] and A. R. M. Wolfert[a]

[a]Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, The Netherlands; [b]Deltares, Delft, The Netherlands; [c]Waterschap Rivierenland, Tiel, The Netherlands

ABSTRACT

Prioritisation of flood defence maintenance is typically based on visual inspection. However, literature shows that the Probability of Detection (PoD) of visual inspection can vary significantly. Here we investigate the PoD for visual inspections of flood defence structures, the consistency of damage classification, and the influence of different variables on the PoD, such as past experience. Four flood defence sections were inspected by 22 different inspectors for a variety of damage types, such as animal burrowing and damage to block revetments. It is found that the PoD varies significantly both per damage type and inspector. Additionally, the estimated severity of damages varies significantly in comparison to the reference situation: over half of the registered damages is assigned a different severity compared to the reference, which potentially leads to incorrect maintenance measures. A likely explanation for the variation in results is the complexity of inspection guidelines and task definitions. Therefore it is advised to simplify inspection guidelines and use more focussed inspections for the most important types of damage. This likely leads to both a reduction of the number of false negatives associated with an increase in flood risk, and better risk-based asset management and maintenance prioritisation in general.

ARTICLE HISTORY
Received 3 May 2021
Revised 16 August 2021
Accepted 26 September 2021

KEYWORDS
Visual inspection; flood risk management; flood defence; Probability of Detection; maintenance

1. Introduction

Earthen flood defences along rivers, lakes and coasts are one of the main measures for mitigating risks of flooding. Due to long term temporal developments such as socio-economic changes and climate change flood defences require reinforcement, typically every 20 to 50 years (Jonkman, Voortman, Klerk, & van Vuren, 2018). In between these reinforcements, flood defence asset managers have to maintain flood defences in the required condition. Such maintenance is for a large part aimed at the revetment, the outer protection layer that protects the flood defence from erosion through waves and currents. Examples of maintenance are repair of drought cracks, resowing grass revetments, repair works on pattern-placed revetments, and repair of damage from animal burrowing. Many of such damages are found to have a significant impact on flood defence safety (van Bergeijk, Verdonk, Warmink, & Hulscher, 2021).

The most important method for detecting such damage is visual inspection, which is typically carried out at different times throughout the year. In such an inspection flood defence inspectors walk or drive along a flood defence and register all relevant anomalies and defects resulting in a condition report. This condition report is then used as basis for maintenance planning. The International Levee Handbook (ILH) lists a variety of inspection types (CIRIA, 2013). The most important types for the countries considered in the ILH are: general inspections to determine the condition of a flood defence and/or whether maintenance works have been conducted properly, inspections before, during or after flood conditions, and special inspections aimed at detection of a specific type of damage (e.g. drought cracks). This paper focuses on general condition inspections.

From past research in other applications, it is found that the detection rate of visual inspection displays significant variation among different applications and inspection types. In general terms, Drury and Fox (1975) report an error rate of 20–30%, but these values vary significantly among different applications and specific situations. For instance Graybeal, Phares, Rolander, Moore, and Washer (2002) carried out a field test to investigate the performance of highway bridge inspectors. This test included 49 inspectors who fulfilled 10 different inspection tasks at 7 different bridges. For general condition assessment it was found that 68% of condition ratings varies within 1 point from the assigned reference rating (10-point scale). Several potentially important variables were identified, such as visual acuity, the extent to which inspectors were rushed, and the perceptions of aspects such as the complexity of the structure and worker safety during inspection. Aside from routine inspections also in-depth visual inspections were carried out. It

CONTACT W. J. Klerk wouterjan.klerk@deltares.nl Faculty of Civil Engineering and Geosciences, Delft University of Technology, P.O. Box 5048, Delft 2600 GA, The Netherlands

Supplemental data for this article is available online at https://doi.org/10.1080/15732479.2021.2001543.

was found that these inspections were not likely to detect the types of defects such inspections are aimed at. Also a large variation in detection rates was observed, ranging from a detection rate of approximately 4% for some weld cracks to 100% for defects to the paint system (Moore, Phares, Graybeal, Rolander, & Washer, 2001). Again inspectors who spent more time, were more comfortable during inspection, and perceived the structures to be more complex, performed better overall. However, what stands out is the large variability in defect detection rates in such general in-depth inspections.

Spencer (1996) and Drury, Spencer and Schurman (1997) performed similar investigations to determine the accuracy of crack detection for airplanes. Here a detection rate of 68% was found. This is significantly higher than the weld crack detection in the research reported by Graybeal et al. (2002). Although obviously the type of cracks and type of structure differs, this might also be explained by the fact that the complexity of objects correlates with lower inspection performance (Harris, 1966). Additionally Harris (1966) found that giving more time for inspection of a complex structure does not increase the detection rate, suggesting that there is an upper limit for a given inspection type. Another explanatory factor might be that the number of fault types to be considered, and thus the complexity of the inspection task itself, strongly correlates with the error rate in inspections (Dalton & Drury, 2004; Gallwey & Drury, 1986). This is supported by findings from research on the accuracy of visual inspection of sewer systems (Dirksen et al., 2013; van der Steen, Dirksen, & Clemens, 2014). Here a clear relation between the number of False Negatives and the complexity of the used coding systems was found, with more complexity leading to lower inspection performance (van der Steen et al., 2014). These examples emphasise the variety of variables that influence inspector performance, of which See (2012) provides a structured overview. Here distinction is made between a.o. variables related to the formulation and scope of the task, the individual characteristics, and the environmental conditions and social circumstances in which inspections have to be carried out. An important example of the latter was identified by Wiener (1984), who found that both complex procedures for rejection and peer pressure to accept products, led to an increase in flinching, resulting in defect products being accepted.

Currently most risk-based assessments of flood defence safety assume that the flood defence is in good condition, and do not include the possibility of undetected defects and their potential effect on flood defence safety. While for instance the International Levee Handbook does note that inspections are not perfect, this is not translated to consequences for flood risk assessments (CIRIA, 2013). Assuming that existing (visual) inspection policies give a complete overview of defects will thus lead to an underestimation of actual flood risk. For instance, if the erosion resistance of a grass revetment is overestimated due to undetected damage, the actual flood risk might be higher than is estimated in typical flood risk assessments. As, based on the literature, it is likely that at least a part of the damaged areas remains undetected for some time, including such factors in risk assessments will improve risk estimates. Insight in the accuracy of visual flood defence inspection, and identifying factors that cause damages to remain undetected, can aid in defining targeted actions to improve inspection quality. Examples are improving task definitions, targeted training of inspectors, and improvement of inspection guidelines (See, Drury, Speed, Williams, & Khalandi, 2017). Such insights can help improve both assessment of existing and prediction of future performance (e.g. Quirk, Matos, Murphy, & Pakrashi, 2017; Ter Berg, Leontaris, van den Boomen, Spaan, & Wolfert, 2019). Obtaining estimates of flood defence inspection accuracy can therefore provide a basis for further improvement of such inspections, improve flood risk estimates, and improve flood protection performance.

This paper presents the results of a field experiment conducted in March 2020. In this experiment 4 different flood defence sections along the Dutch Rhine river were inspected 14 times by 22 different inspectors. The goal was to answer three main questions with regards to the quality of visual condition inspections of flood defences:

1. What percentage of defects is detected in a typical condition inspection?
2. What is the consistency with which defects are classified?
3. Can influential factors that impact inspector performance be identified?

Answers to these questions can be used to identify possible improvements for inspection of flood defences. Section 2 presents relevant background on the current practice of flood defence inspections, both internationally and in the Netherlands. Section 3 presents the methods and setup of the field test, as well as a description of the field test location. Results of the field test are presented in Section 4, after which Section 5 provides a discussion on findings. Conclusions are summarised in Section 6.

2. Practice of flood defence inspections

2.1. Approaches for flood defence inspections in different countries

Crespo Márquez (2007) defines inspection as a 'check for conformity by measuring, observing, testing and gauging the relevant characteristics of an item'. For the process of translating inspection findings to maintenance actions, Bakkenist, van Dam, van der Nat, Thijs, and Vries (2012) distinguish 4 steps:

- Observation: observing a defect, anomaly or condition.
- Diagnosis: assessing the nature and type of a defect or the condition, as well as the severity, based on relevant (predefined) characteristics (Crespo Márquez, 2007).
- Prognosis: assessing whether the severity of the defect or the general condition will change in the future.
- Operationalization: defining appropriate actions to deal with the observed defect, such as repair, overhaul or doing nothing (Crespo Márquez, 2007).

In many countries the basic principles for flood defence inspection are in line with those outlined in the International Levee Handbook (CIRIA, 2013). Generally inspections focus on the first three steps, and always at least combine observations and diagnosis: defects or conditions are classified using different parameters and severity classes. In some cases inspections are aimed at observing and diagnosing defects, in some cases at observing the general condition of a structure. Inspections aimed at observing defects (e.g. animal burrows, rutting and corrosion) are our main focus here. Inspections at assigning condition ratings to flood defence sections are not further considered here, although it should be noted that such ratings are sometimes achieved by translating observed defects into condition ratings (CIRIA, 2013). E.g. the US Army Corps of Engineers translates ratings for 125 specific items considered in the inspection to 'acceptable', 'minimally acceptable' and 'unacceptable' ratings.

In the UK a similar approach is used, but additionally the condition grade (scale 1–5) established from visual and other types of inspection is used to predict remaining life using deterioration curves (Flikweert & Simm, 2008). As such, the condition grades are directly translated into a prognosis of the future condition. However, defects that determine the condition grade might often be caused by shock-based processes rather than continuous degradation processes (Sanchez-Silva, Klutke, & Rosowsky, 2011). It was demonstrated in Klerk and Adhi (2021) that the condition of a certain flood defence section can vary significantly over time, indicating that the use of standardised deterioration curves might not correctly reflect the actual degradation behaviour.

Several factors for high quality flood defence inspections are mentioned in literature. Specific focus is often on training, and in the UK new inspectors first have to gain in-field experience under supervision of a more experienced inspector. Compared to the factors mentioned by See (2012), being able to evaluate flood risks based on an understanding of the failure mechanisms, experience with inspections and computer literacy are some of the factors mentioned that ensure consistent, efficient and thorough inspections (CIRIA, 2013). Long, Mawdesley, and Simm (2006) describe a blueprint of an 'ideal condition indexing' process. This mostly concerns more extensive use of information from other sources, adding other types of measurements, and increasing the range of condition values to enable greater gradation of asset condition. While such efforts can indeed lead to a better overall estimate of flood defence condition, it is doubtful whether the condition estimates from visual inspection itself would improve, as the task complexity will increase, which typically results in lower defect detection and less consistency in classification of defects as was shown by a.o. van der Steen et al. (2014), Gallwey and Drury (1986) and Dalton and Drury (2004).

2.2. Routine condition inspections of flood defences in The Netherlands

As the field test reported here was carried out using the general approach used in the Netherlands, this is described in more detail in this section. The main focus of the field test was to mimic a spring inspection, usually carried out in March, after the winter season, i.e. the period between October and March during which most storms and flood waves occur. The goal of the spring inspection is to identify defects and anomalies at all dike sections such that, if required, repair or overhaul works can be carried out before the next winter season. While other inspections are also of importance, the spring inspection is the backbone of maintaining the overall condition, as most of the repair and maintenance works are based in the spring inspection results.

Spring inspections are typically carried out using the Digigids guideline (Het Waterschapshuis, 2016). The Digigids is a comprehensive guideline with many different types of damage for different types of flood defence elements. Inspectors have to classify defects/damages in three variables: the flood defence element (e.g. grass revetment), the damage parameter (e.g. animal burrowing or bare spots) and the severity, which is a classification on a 4 point scale: good, reasonable, mediocre and bad. The Digigids provides descriptions for each category, as well as reference photos of damages. Figure 1 illustrates this for bare spots. The definitions for severity are not explicitly related to failure behaviour or failure mechanisms, although these can be related to the sod quality of the grass revetment, which is an important input parameter in reliability assessments (Klerk & Adhi, 2021). In principle however, the Digigids is aimed at inspecting the condition of the revetment and not at assessing the risk of failure.

Although the Digigids does facilitate registering using severity estimation 'good', in practice this is not done and only points with severity 'reasonable' or worse are registered. For spring inspections some prioritization is made in terms of the parameters to be inspected. For instance, most water authorities do not register flotsam on slopes, and burrowing by moles and mice is typically also not registered as it is dealt with in routine maintenance.

Despite some differences in rating systems, interpretation of results, and specific prioritizations, the approaches towards inspections are fairly similar in other countries. In most countries (e.g. France, UK, USA, and Ireland) also a system of 3 to 5 condition grades is used for diagnosis of the severity (CIRIA, 2013). Additionally similar types of defects are considered, for instance: unwanted (woody) vegetation, bare spots in the grass cover, deformations, erosion, cracks and animal burrows. Inspections are typically carried out by 2 inspectors, both to ensure worker safety and to ensure completeness of the data. In most countries registrations are made including photographs and GPS coordinates. Data is reported to the flood defence asset manager and stored for future analysis.

3. Methods

3.1. Quantifying the accuracy of inspection

Results from an inspection can be classified in 4 different categories (Keprate & Chandima Ratnayake, 2015):

- True Positive (TP): a defect exists and is detected.

(a)
Good: no bare spots.

(b) Reasonable: at most 5 spots with a diameter <0.2 m where vegetation is gone.

(c) Mediocre: at most 5 spots with 0.2 m < diameter < 0.3 m where vegetation is gone.

(d) Bad: > 6 spots with diameter > 0.2 m, or 1 spot with diameter > 0.3 m where vegetation is gone.

Figure 1. Example of Digigids classification for bare spots. panes a-d show increasing severity (good, reasonable, mediocre, bad). Captions for subfigures give description of category. All descriptions apply to an area of 25 square metres. All figures originate from Het Waterschapshuis (2016). (a) Good: no bare spots. (b) Reasonable: at most 5 spots with a diameter <0.2 m where vegetation is gone. (c) Mediocre: at most 5 spots with 0.2 m < diameter < 0.3 m where vegetation is gone. (d) Bad: > 6 spots with diameter > 0.2 m, or 1 spot with diameter > 0.3 m where vegetation is gone.

- False Positive (FP): a defect does not exist, but is detected.
- True Negative (TN): a defect does not exist, and is not detected.
- False Negative (FN): a defect exists, but is not detected.

The effectiveness of non-destructive evaluation techniques such as visual inspection is typically quantified using the Probability of Detection (PoD). The PoD can be computed using:

$$\mathrm{PoD} = \frac{TP}{TP + FN}. \tag{1}$$

It should be noted that in some fields (e.g. pattern recognition) this parameter is named 'recall'. The other way around, the probability that a registered defect does not exist can be quantified using the Probability of False Alarm (PFA):

$$PFA = \frac{FP}{TN + FP}. \tag{2}$$

As hardly any false positives are registered in the field test, the PFA will not be considered any further in the analysis of the results.

Another important point is the distinction between classification and detection errors. Practically, any defect that has not been registered can be considered a False Negative. However, in some cases an inspector might detect a damage (with for instance severity 'mediocre'), but classify its severity as good. As points with severity 'good' are not registered in spring inspections, such a case might be incorrectly judged as a detection error, while it is in fact a classification

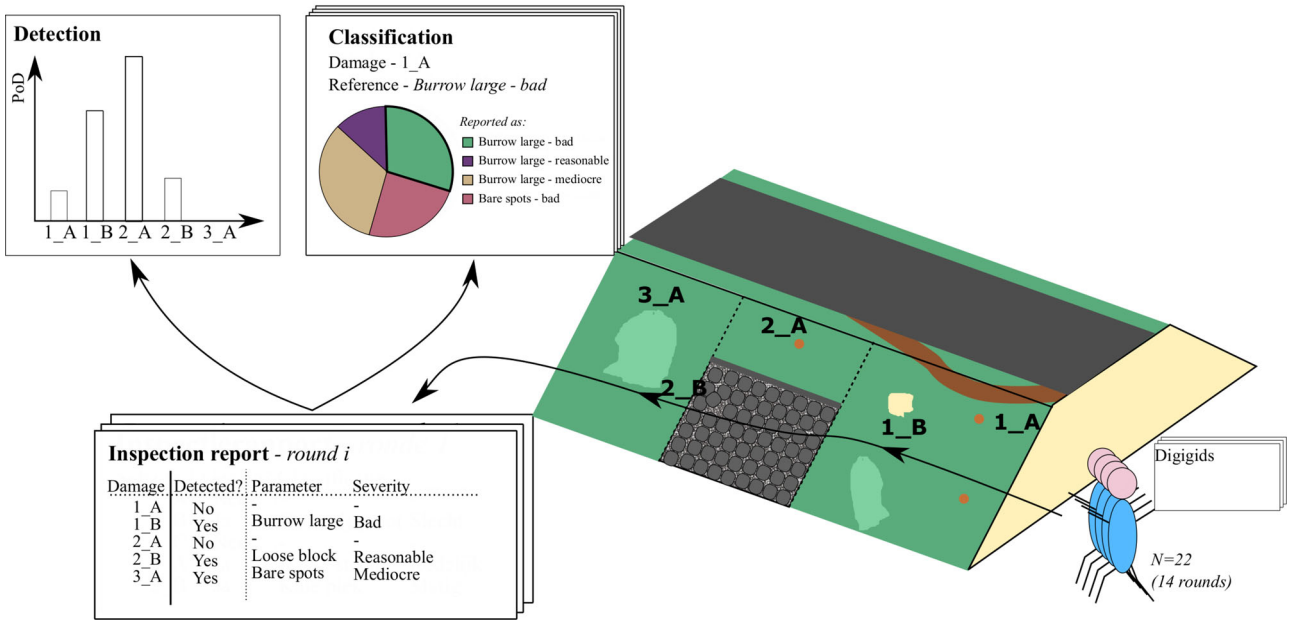


Figure 2. General approach to the field test. Inspectors inspect several dike sections as by their normal practice. This results in an inspection report for each inspection round where for each damage present it is indicated whether it was detected and how it was classified. This enables analysis of the detection accuracy and classification consistency based on the predefined reference situation.

error. Hence, the data would suggest it is a False Negative, whereas it is in fact a True Positive, with an error in the classification of severity. Therefore all PoD-values computed for the field test are lower limits. It has to be noted that the practical effect of such classification errors is that the damage remains unknown to the asset manager.

3.2. General set up of field test

The goal of the field test was to answer the three main research questions outlined in Section 1. Therefore it was set up to mimic the actual spring inspection for flood defences using the typical guidelines in the Netherlands as closely as possible. Figure 2 shows the general setup of the test. Different dike sections (3 are displayed here) were inspected by different (teams of) inspectors. During the field test inspectors registered defects on their smartphone or tablet. This was done using a cloned version of the ESRI Survey123 application they normally use during inspections, in order to avoid issues in registering defects. Inspectors always had to register the coordinates, parameter, severity, dimensions, other remarks, and add a detail and overview picture of the defect. The database from these registrations enables analysis of the detection accuracy, as well as the consistency of classification using parameter and severity in accordance with the Digigids. To enable the analysis of detection accuracy and classification consistency all considered dike sections were pre-inspected in order to map all defects and determine the reference classifications.

To facilitate the analysis of influential variables, several questionnaires were presented to the inspectors at different times. Before the test, an extensive questionnaire on a.o. several personal characteristics, training, experience and their common inspection approach was filled in by the participants. Additionally throughout the experiment inspectors were inquired about a.o. their feeling during the day, their experiences during the test and whether they thought the field test was representative for their normal inspections. All questions to individual inspectors have been listed in Table A1 in Appendix A. Answers to questions have several categories: numerical values, yes/no answers, multiple choice questions and questions on a 1 to 5 rating (very bad-very good or disagree-agree). In the analysis of influential variables only those of which an effect on inspection performance could be expected have been included. All considered variables including the questions they originate from are listed in Table A2 in Appendix A and will be further discussed in Section 4.3. Before and after inspection rounds, the inspectors involved were inquired (as a pair) about the difficulty of that round, and whether they experienced any time pressure. This was only the case for a few inspection rounds, and these questions have not been analysed further.

During the test a supervisor was present at each dike section. These supervisors posed questions before and after inspection rounds, gave participants general instructions and observed the general behaviour of participants during the test. Supervisors were given smartphone applications to log important events during the inspection, and general remarks on participant behaviour. Examples are people passing by, the walking routes of inspectors, and remarks about the collaboration between participants.

Supervisors ensured that inspection rounds lasted no longer than 25 minutes. This time frame was determined based on the typical time used for spring inspections and was chosen such that no additional time pressure was imposed, as was confirmed from the questions posed to the

inspectors. Also, nearly all inspections were finished within the given time frame.

An important working agreement for spring inspections is that defects with severity 'good' are not registered. This means that based on the registered points, it is impossible to determine whether a damage was not detected, or whether it was detected but classified with severity 'good'. The logging of the supervisors was used to gather evidence on points that were detected but not registered in the Survey123 database. Such points where added to the database with classification 'good'. Thus the database consists of two parts: the registrations in Survey123, and additional damage registrations based on the logging. These datasets contain all required information for determining the Probability of Detection for damage points and section damages, and analysing the classification consistency. It should be noted that, as the supervisor logging is not entirely complete, the computed PoD-values are still lower bounds.

3.3. Damage classes considered

In the analysis two damage classes are considered. The first class are the damage points, which are specific damage spots such as a specific animal burrow. As in some cases the overall damage to a flood defence section is a better measure for the state of the flood defence, section damages are also considered. If for instance a certain dike section contains 3 animal burrows, this yields 3 individual damage points and a section damage 'burrowing' that encompasses all three burrows for that specific section. Whether section damage 'burrowing' has been detected is computed as:

$$I(\text{burrowing}) = \max(I_{\text{burrow1}}, I_{\text{burrow2}}, I_{\text{burrow3}}), \qquad (3)$$

where $I(..)$ is an indicator function that indicates whether a damage is detected ($I(..) = 1$) or not. Logically the PoD for section damages will always be equal or larger than for the damage points it encompasses. Determining the PoD for section damages is relevant, as the maintenance of some damage types is carried out at a section level, meaning that all individual damages at a certain section will be repaired. Based on the severity of a damage, and its potential consequences for failure distinction can be made between essential and non-essential section damages and damage points. Essential damages are those that should not be missed in an inspection as these likely have a direct impact on flood defence safety. To summarise: a single damage point (e.g. a specific animal burrow) is part of the subset that forms a section damage 'burrowing' and both the damage point and main damage are categorised as (non-)essential. Note that non-essential damages could develop into essential damage over time. Whether damages are essential has been based upon an expert assessment of their potential consequences for the strength of the flood defence. A description of the damages at the different sections is given in Section 3.5.

3.4. Approach for analysis of influential variables

The influential variables based on the questionnaires are of two main types: categorical variables typically consisting of 2 or 3 categories, and variables based on questions that were answered using a 5-point scale or using numerical values (such as years of experience).

For the analysis Bayesian Parameter Estimation as outlined by Kruschke (2013) is used. Using data from the experiment, prior assumptions on uncertain parameters are updated to obtain posterior distributions of for instance different categories. Based on these distributions the difference between groups or the influence of a numerical variable can be assessed. This estimation of the posterior distribution is done by generating multiple samples using Markov chain Monte Carlo.

The advantage of Bayesian Parameter Estimation over for instance null hypothesis significance testing is that it provides richer information, for instance with regard to the influence of uncertainty on the differences between groups. As sample sizes are relatively small, this provides more insight in the influence of different variables, rather than a simple acceptation/rejection of the null hypothesis of, for instance, two groups originating from the same distribution. Nevertheless, given the relatively small sample sizes, all results should be interpreted as an exploration of what parameters might influence flood defence inspection performance, rather than statistical evidence of the importance of a certain parameter.

For categorical variables estimates of distribution parameters of a three-parameter Student-t distribution of the number of detected damage points d for both groups are obtained (e.g. whether inspectors used a tablet or smartphone for registration). For numerical variables the parameters of a linear regression between the number of damage points detected d and a numerical variable y (e.g. inspector age) are estimated, such that $d = a + b \cdot y + \epsilon$. More details on the precise formulations and prior assumptions are given in Appendix B.

In line with Kruschke (2013) the 95% Highest Density Interval (HDI) for all parameters is computed. This is the interval that contains 95% of the posterior density of the distribution with estimated parameters. For categorical variables the main indicator is the effect size, which is defined as: $(\mu_1 - \mu_2)/\sqrt{(\sigma_1^2 + \sigma_2^2)/2}$ for groups 1 and 2. If the HDI is (almost) entirely negative or positive, this means that it is highly likely that there is a difference between the categories. For numerical variables the main indicator is the HDI for slope b. If the HDI for b is (almost) entirely positive or negative, this indicates a relation between the detected damage points d and the considered variable y. Note that only influential variables for damage points are analysed, not for section damages.

3.5. Description of field test location

The field test was conducted at 4 dike sections near the city of Tiel, along the Waal river in the Netherlands. Pictures of the 4 sections are shown in Figure 3. The sections were

Figure 3. Impression of the 4 sections. Photos for sections 1, 2 and 3 were taken by inspectors during the test. Photo 4 was taken by one of the supervisors.

chosen as these are representative for the flood defences in the considered area. For each dike section a set of reference damage points and section damages was derived based on the pre-assessment and field test. An overview of all reference damages is provided in Table 1.

Section 1 is approximately 185 metres long. Here the inner slope, which consists of a grass revetment, was inspected. Several small damages are present to the grass cover, both burrows and other types of damage. Additionally the slope is deformed locally. Both the presence of burrowing at a section level, and the slope deformation are essential damages in an inspection of this section.

Section 2 is 200 metres long, and here the outer slope was inspected. The general shape of the grass revetment is good except for a few spots with weeds. However, there is a significant number of animal burrows, mostly caused by dogs who enlarged smaller pre-existing mice or rabbit burrows. It has to be noted that damage point 2_6 developed during the field test and could therefore only be observed in 5 of 14 rounds. This is the only damage that developed in the period the field test was executed. Although there was high water during the test, the water level did not influence the inspectability at this dike section.

Section 3 is 200 metres long, and the outer slope was inspected. The lower two-third of the slope is covered with a block revetment (consisting of pattern-placed basalt), and the upper third is covered with grass. The block revetment is in a relatively bad shape: there are several loose and missing blocks, the joint fill material has washed out and there are tree trunks that penetrated the revetment and displaced the blocks. Many of the damage points at this section are therefore classified as essential. The grass at the higher part of the slope is also not in good shape. It has to be noted that at section 3 the influence of the high water level conditions influenced the outcome of the test, as it was impossible to walk on the maintenance path at the lower part of the revetment. During some test sessions the inspectability was lowered as the revetment was wet due to rainfall and therefore difficult to access.

Section 4 is approximately 80 metres long, but here both inner and outer slope as well as the crest and the inlet structure had to be inspected. In general there is a significant amount of rough vegetation at this section, and the transition between revetment and structure is an important point of attention. Some parts of the outer slope contain a concrete lawn grid which is deformed in 1 location. The high water conditions had no influence on inspectability. Figure 4 provides an example of the results from the Survey123 database for dike section 4. Here both registrations from the pre-assessment and the field test are shown for each damage point. All registrations have been manually coupled to the registrations, based on the attached photographs, description and location.

Table 1. Overview of reference damages for all sections.

	Damage point		Section damage		Reference classification		
	ID	*Description*	*ID*	*Description*	*Parameter*	*Severity*	Remarks
Section 1	1_1	*Large burrow 1*	1_B	Burrowing	Burrowing large	Bad	
	1_2	*Small burrow 1*	1_B	Burrowing	Burrowing small	Bad	
	1_3	*Weeds*	1_A	Grass cover	Weeds	Reasonable	
	1_4	*Cover & bare spots*	1_A	Grass cover	Bare spots/Coverage	Mediocre	
	1_5	Slope deformation	1_C	Deformation	Slope deformation	Bad	
	1_6	*Small burrow 2*	1_B	Burrowing	–	–	Not in pre-assessment
	1_7	*Small burrow 3*	1_B	Burrowing	–	–	Not in pre-assessment
Section 2	2_1	*Burrow 1*	2_A	Burrowing	Burrowing large	Bad	
	2_2	*Burrow 2*	2_A	Burrowing	Burrowing large	Bad	
	2_3	*Crack*	2_B	*Grass cover*	Cracks	Mediocre	
	2_4	*Burrow 4*	2_A	Burrowing	Burrowing large	Bad	
	2_5	*Weeds*	2_B	*Grass cover*	Weeds	Reasonable	
	2_6	*Burrow 5*	2_A	Burrowing	Burrowing large	–	Developed during field test
	2_7	*Burrow 6*	2_A	Burrowing	Burrowing large	Bad	
	2_8	*Other burrows*	2_A	Burrowing	Burrowing large	Bad	Not in pre-assessment
Section 3	3_1	*Bare spots*	*3_A*	*Grass cover*	Bare spots	Bad	
	3_2	Washed out joint fill	3_B	Loss of clamping force	Joint fill washout	–	
	3_3	Tree trunks	3_C	Woody vegetation	Woody vegetation	Bad	
	3_4	*Displaced block*	3_B	Loss of clamping force	–	–	Not in pre-assessment
	3_5	Loose block 1	3_D	Loose or missing blocks	Loose blocks	Bad	Not in pre-assessment
	3_6	Loose block 2	3_D	Loose or missing blocks	Loose blocks	Bad	Not in pre-assessment
	3_7	Missing block	3_D	Loose or missing blocks	Holes	Bad	
	3_8	Loose block 3	3_D	Loose or missing blocks	Loose blocks	Bad	Not in pre-assessment
	3_9	Loose block 4	3_D	Loose or missing blocks	Loose blocks	Bad	
	3_10	*Small burrow*	3_A	*Grass cover*	–	–	Not in pre-assessment
Section 4	4_1	*Woody vegetation 1*	4_A	Rough and woody vegetation	Woody vegetation	Bad	
	4_2	*Woody vegetation 2*	4_A	Rough and woody vegetation	Rough vegetation	Bad	
	4_3	Transition 1	4_B	Transition with structure	–	–	
	4_4	*Transition 2*	4_B	Transition with structure	–	–	
	4_5	*Lawn grid deformed*	4_C	*Grass cover*	–	–	
	4_6	*Tree growth*	4_A	Rough and woody vegetation	Woody vegetation	Reasonable	
	4_7	Rough slope	4_A	Rough and woody vegetation	Woody vegetation	Bad	
	4_8	*Small burrow*	4_C	*Grass cover*	Burrowing small	Bad	
	4_9	*Weeds*	4_C	*Grass cover*	–	–	Not in pre-assessment

Notes. Descriptions of damages marked as non-essential are displayed in italics. Both damage points and corresponding section damages are given, as well as reference classification. Not in all cases a reference classification is given as this could not be determined because damage points were not in the pre-assessment or the classification was ambiguous.

3.6. Conditions during the field test

During the field test the river water levels were relatively high. Several inspectors indicated that this impacted their performance and approach, especially at dike section 3. At the 3rd of March, water levels were lower than at the 6th of March. Weather conditions differed slightly between the two dates: the 3rd of March was dry, and mostly sunny, at the 6th of March rainfall in the night before caused wetter slopes. Especially at section 3 this had influence on the accessibility of the block revetment.

All inspectors received a time schedule for their inspections. Time schedules were generated based on a randomised algorithm to ensure each section was inspected once, and inspectors always inspected with a different partner. In a few occasions inspectors arrived slightly late or were rescheduled to a different time slot. Nevertheless, except for 1 inspector who did not inspect section 3, all inspectors inspected all sections, each time with a different partner.

Due to absence of 6 of the total 28 scheduled inspectors not all inspections were done by a pair of inspectors. 25 out of 56 inspection rounds were carried out by 1 inspector (7 at section 3, 6 at the other sections) and each participant inspected 0 to 2 rounds on their own. From the results it is demonstrated that this had no influence on scores at sections 1, 2 and 4, but at section 3 single inspections are found to result in lower detection rates.

An important remark with regards to the supervisors at different sections is that each used a slightly different approach to log events during the test. In some cases supervisors recorded many voice messages, others took plenty of photos. This might have influenced the number of damage registrations that were added based on the logging. Additionally, at the 6th of March inspectors were given specific instructions to pay attention to damage points that were detected but not registered in Survey123. In total 29 damage registrations were added based on the supervisor logging, of which 14 concerned the 5 inspection rounds at the 6th of March, and 15 registrations the 9 rounds on the 3rd of March.

4. Results

4.1. Accuracy of flood defence inspections

First the overall Probability of Detection (PoD) for damage points and section damages is analysed. Figure 5 shows the PoD for all damage points per section. Grey bars indicate registrations based on the supervisor logging, and hatched bars denote damage points that were marked as essential in

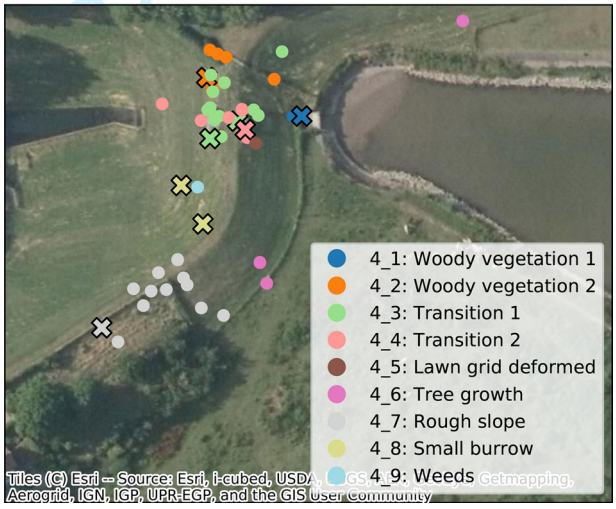


Figure 4. Results for dike section 4 obtained from the Survey123 database. Dots are registrations during the test, crosses indicate registrations during the pre-assessment.

the reference. Overall, the PoD varies significantly, ranging from 0 to almost 0.9 per damage point. Most of the registrations added based on the logging concern smaller issues, such as small burrowing and the quality of the grass cover. Also note that there is no clear difference in the PoD for essential and non-essential damages, except for section 4 where the two essential damage points were registered by the majority of the inspectors.

The PoD is lowest for issues with the block revetment at section 3. Here the loose and missing blocks have been sparsely or not detected during the test. There are several explanations for this: first of all, the high water conditions hampered inspectors (which was also indicated in the questionnaires). Secondly, it should be noted that only damage point 3_9 (Loose block 4) was detected during the pre-assessment, when conditions were much more favourable. A second explanation is therefore that inspecting block revetments is generally more difficult as it is harder to see and process all the details. This especially holds for blocks that are loose, but still in their place. This could also explain why the missing block was detected more often, even though its proximity to some of the loose blocks.

At sections 1 and 2, there were multiple burrows for which the PoD varies significantly. The question is whether the main reason is failure to detect, as inspectors indicated that in many cases they do not register all the burrows at a section. This reduces the work load during inspection, and common maintenance works to deal with it will be done at a section level anyway, so all burrows will be repaired together.

From that perspective, it is more relevant to look at burrowing as a section damage. Figure 6 shows the PoD for all section damages. By definition these are higher than for individual points (see Equation (3)). It can be seen that at section 2 most inspectors (PoD = 0.86) registered at least 1 burrow.

Nearly half of the inspection rounds was done by a single inspector instead of a pair. For most damage points and section damages this caused no major differences in the estimated PoD, except for section damages 3_B (loss of clamping force) and 3_D (Loose or missing block), and the corresponding damage points. For these damages individual inspectors scored much lower: for 3_B the PoD for a pair and individual were found to be 0.86 versus 0.14, for 3_D the PoD was 0.71 versus 0.29. Figures for all damage points and section damages for individual/pairs of inspectors can be found in Appendix C.

An important aspect in the context of risk-based inspection is not only the average PoD, but also the variation among different inspectors. Figure 7 shows the variation among inspectors for both damage points and section damages, and subsets of the essential damages. For damage points, the PoD ranges between approximately 0.25 and 0.55, with a bit more variation for the essential damages. Note that this is strongly influenced by the low detection percentages of the various loose blocks at section 3. The PoD for section damages ranges between 0.5 and 0.9. The variation is similar to that of damage points, but the average is significantly higher.

4.2. Consistency of classification of flood defence defects

Based on the database of inspection registrations it can be determined how consistently damages were classified in the database. Consistency is here defined as the agreement between inspection reports of different inspectors. This is determined both by whether the damage was detected, and what parameter and severity they assigned. Theoretically, each damage point should have 1 correct parameter, although in some cases multiple parameters of the Digigids might be applicable. For instance, in many places overgrown vegetation consists of both weeds, woody vegetation and generally rough vegetation. Henceforth, when looking at consistency between parameters it is not analysed how many inspectors chose the 'correct' parameter, but rather at how many different parameters were used for the same damage point. To that end, a consistency index is defined:

$$C = \frac{(N/N_{\text{par}})}{N}, \tag{4}$$

where N is the number of records in the database, and N_{par} denotes the number of unique parameters (e.g. weeds, small burrowing, large burrowing) in the damage registrations by inspectors. Higher values for C mean that inspectors were more consistent in their parameter choice.

Figure 8 shows the consistency index for all damage points with $N > 3$. Damage point 1_4 was not included as this encompasses 2 damage parameters (bare spots and cover). It is found that especially for the damage to the transition (4_3 and 4_4) the registrations are very inconsistent: 4_3 was registered 11 times with 7 different parameters, 4_4 5 times with 3 different parameters. This indicates that there is no clear parameter to define damage to a transition. Note that in this case many of the parameters used by the inspectors are actually visible, as is shown in Figure 9. Here it can be clearly seen that there is rough vegetation, weeds, a bare

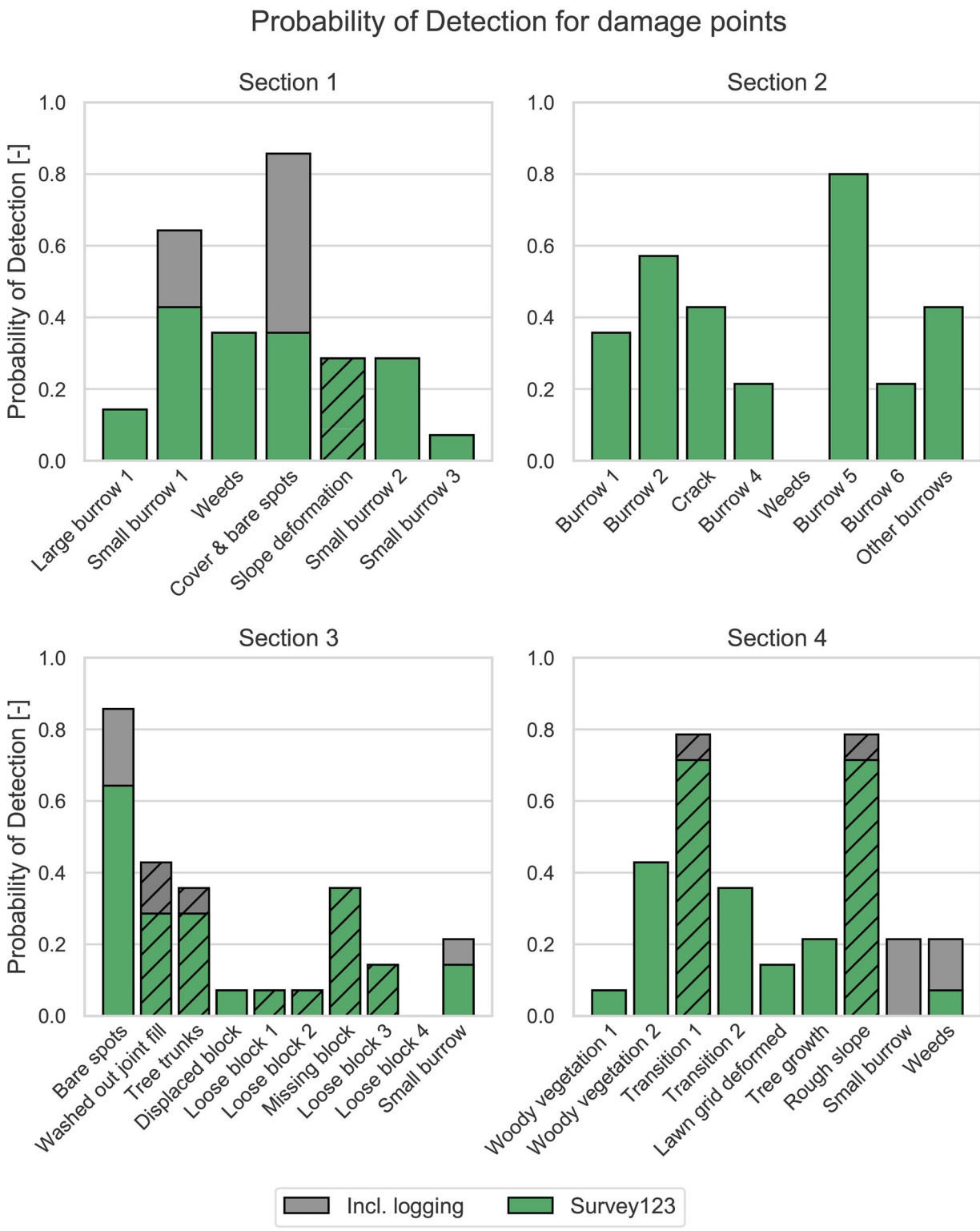


Figure 5. Probability of Detection for all damage points per section. Hatched bars denote damage points that were categorised as essential damages. Green indicates registrations in Survey123, whereas grey includes registrations added based on the supervisor logging. Note that Burrow 5 at section 2 was only present in the last field test session and was therefore only observable during 5 inspection rounds.

spot, and possibly also burrowing. All these parameters were used by different inspectors and are at least partially representative for the situation.

For burrowing the consistency is 0.5 in many cases, as inspectors used both small and large burrowing as parameters. This indicates that distinguishing these in practice is difficult. Other cases with low consistency (e.g. 2_3, 3_1 and 4_7) are also typically damage points where multiple damage types are present. In practice asset managers always use the photographs to review the actual situation (or do a field visit) before deciding what maintenance is to be done. This procedure likely ensures that, even though the parameters might be inconsistent or incorrect, at least the correct maintenance action is taken.

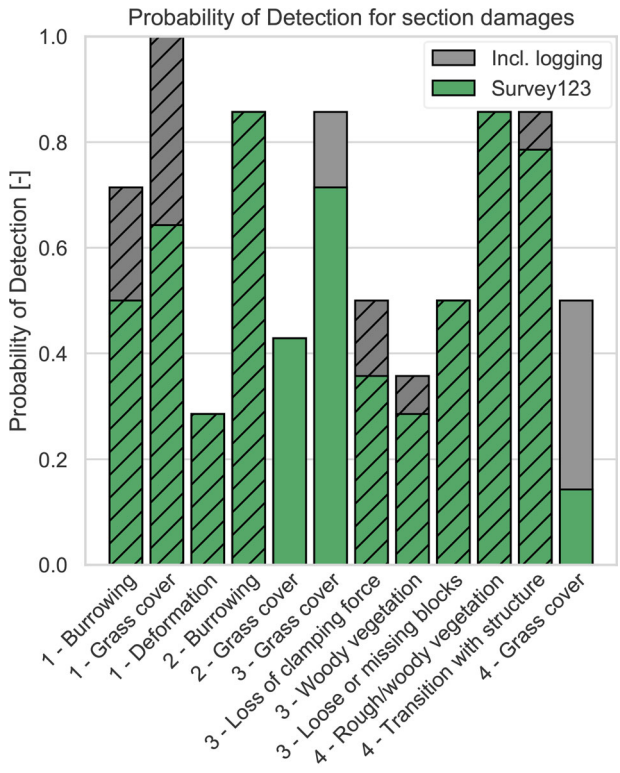


Figure 6. Probability of Detection for all section damages (numbers at horizontal axis denote the section). Hatched bars denote section damages that were categorised as essential. Green indicates registrations by inspectors in Survey123, whereas grey includes registrations added based on the supervisor logging.

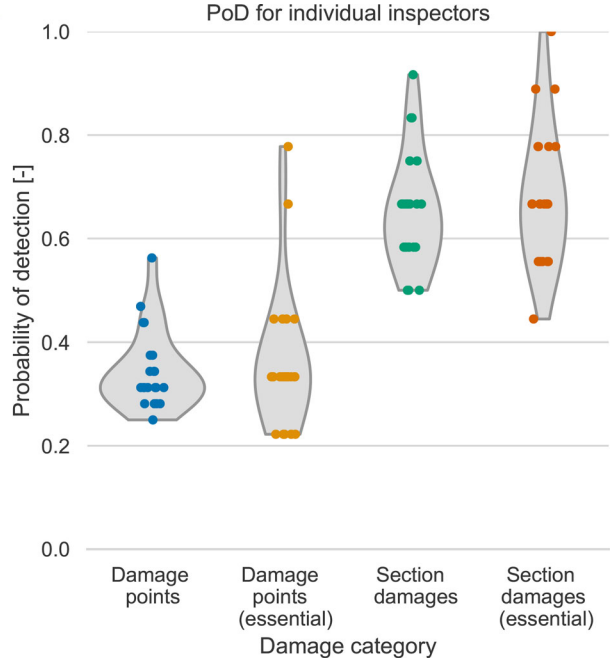


Figure 7. Probability of Detection (PoD) for individual inspectors. Grey shaded areas provide a density estimate capped at the highest and lowest PoD encountered in the test. Coloured dots provide results for individual inspectors for (essential) section damages and damage points.

However, it has to be noted that the reported severity does play an important role in an asset managers decision to review a registered damage, as they mainly focus on reviewing with severity 'bad'. Figure 10 shows the fraction of records that was categorised in the different severity categories, including the fraction of inspections where a damage was not registered. The latter consists of damage points where inspectors failed to detect a damage, or where they found it not severe enough to register. For damage points with severity 'bad', more than half of the inspectors that registered such a damage classified it as less severe. For damage points with reference mediocre or reasonable, there is also a large variation in reported severity. It should be noted that this figure is based on the entire database, including points added based on the supervisor logging. If the same figure is made just for the points registered in Survey123, the fraction classified as 'good' is 0 in most cases, which emphasises that the grey bars contain many detected damage points classified as good.

4.3. Influential variables for flood defence inspection accuracy

As was mentioned in Section 3, the overall number of samples in this test limits the extent to which conclusions can be drawn with regard to influential variables. However, a Bayesian Parameter Estimation can give some directions for future research. In general there are two main types of variables: categorical variables that are split into 2 to 4 categories, and numerical variables that are either numerical variables (such as age) or answers given on a 1 to 5 scale. The majority of the analysed variables are based on the questionnaires in Table A1. All variables are listed in Table A2, including the origin of the data (question or other data) and whether these have been included in the analysis or not. The influence of variables on inspection performance was estimated using Bayesian Parameter Estimation as described in Section 3.4.

Figure 11 displays the Highest Density Intervals (HDI) for all numerical variables (green), and all categorical variables (orange) where each group consists of 5 or more participants. The HDI indicates that 95% of the probability density of the posterior distribution falls within this range (grey bar). For numerical variables the HDI for the slope of the regression line is shown. A negative slope means that higher values for the numerical variable relate to less detected damage points. To compare variables with different ranges all values have been rescaled to a 1 to 5 scale based on the minimal and maximal values in the dataset. For categorical variables the HDI for the effect size is given. If the value is positive it means that the listed category performs better. E.g. inspectors who are also asset manager perform slightly better than those who are not. If the HDI is (almost) entirely negative or positive, this indicates that a variable has effect on inspection performance (number of damage points detected). Additionally the coloured lines (orange for categorical, green for numerical variables) indicate the mean and interval $\mu \pm \sigma$.

The variables are grouped in 3 categories. Experience & training variables are often related: in practice asset managers are involved in maintenance planning and execution,

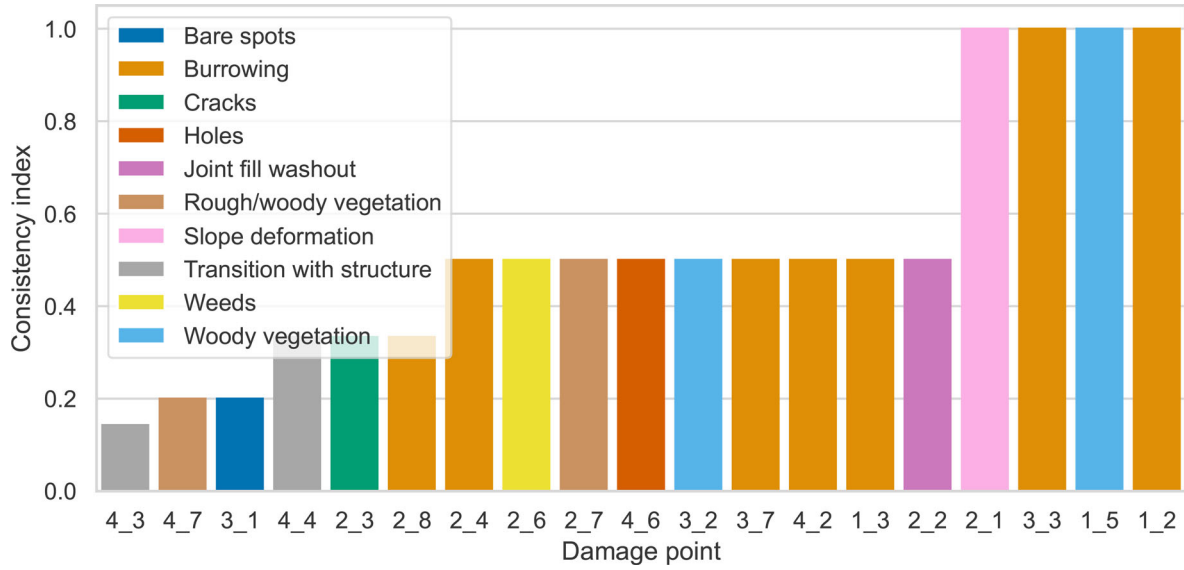


Figure 8. Consistency index of different damage points with more than 3 registrations in the database. Colours indicate the reference classification parameter (see Table 1) except for 'Transition' where no reference parameter could be determined, and burrowing where small and large burrows are combined.

Figure 9. Example of the presence of multiple damage parameters at a single damage point (4_3). At least bare spots, weeds and rough vegetation could be applicable here, while the cause of the bare spot (just right of the reference marker) could be burrowing.

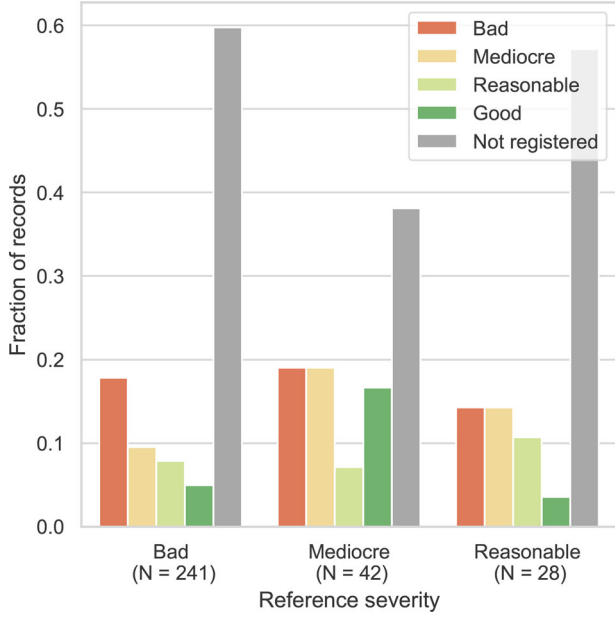


Figure 10. Fraction of records in the Survey123 database for different severity classifications compared to the reference. *N* denotes the number of possible registrations for each category.

and both daily and emergency inspection. All these variables are found to relate to a somewhat better inspection performance, although in none of the cases the HDI is entirely positive or negative. Logically, more experience in years also relates to better performance. For personal characteristics no clear relations are found. For the test circumstances and approach, it is found that the (negative) effect of inspecting solo at section 3 is relatively large.

5. Discussion

This paper presents a field test where the accuracy and consistency of visual flood defence inspections is investigated. The goal of this field test is to answer three main questions.

The first main question concerns the Probability of Detection (PoD) of flood defence inspections. The field test shows that there is a large variation in the PoD, both between inspectors as well as between different (types of) damage points. The variation between inspectors cannot be explained by the parameters elicited in the various questionnaires, and is likely due to the nature of visual inspection (of flood defences) itself, and the general method used in the field test. This is in line with findings from literature (e.g. Dirksen et al., 2013; Graybeal et al., 2002), where also large variations between inspectors and damage points was found. It should be noted that damage points with severity 'good' were not registered: therefore some of the non-detections might have been detected but classified too leniently. However, most of the cases where this was observed concerned minor damages, and there is a general agreement among inspectors that essential damages (e.g. loose blocks and slope deformations) should be registered.

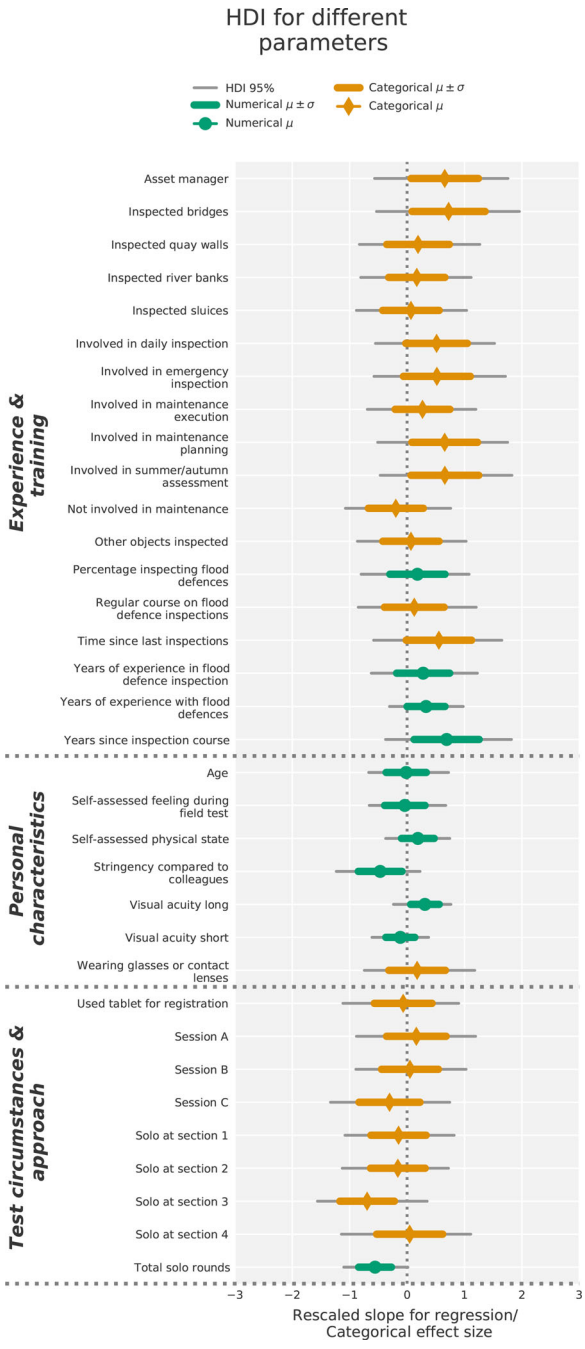


Figure 11. Results of Bayesian Parameter Estimation for categorical (orange) and numerical (green) variables. For categorical variables the effect size is given, for numerical variables the estimated slope of the linear regression. Diamonds and circles indicate mean values, coloured lines indicate the interval $\mu \pm \sigma$. Grey lines indicate the 95% Highest Density Interval. Note that the slope has been rescaled to the interval 1–5 to make numerical variables with different ranges comparable.

The results clearly indicate that the block revetment at section 3 was more difficult to inspect. Although circumstances during the field test were difficult due to the high water levels, the fact that multiple damage points were also not detected in the pre-assessment emphasises the difficulty of detecting flaws in block revetments. For other types of damage points, such as animal burrows, the PoD varies significantly per damage point. In some cases this is due to the method of registration: for instance at section 2, there is a large number of burrows and it is time-consuming for inspectors to register all individual burrows. Inspectors indicated that they often register only 1 or 2 burrows, as the commonly applied maintenance method will ensure that all burrows over a longer section are repaired. This might explain the variance between the PoD for different burrows at this section. It should be noted that in this particular case a more consistent method of registering damage would be to assess the number of burrows at a section level, rather than at individual points. To improve consistency it is therefore recommended to align the spatial level (section or point) of the damage registration with that of the commonly applied maintenance measure.

The second goal of this field test was to investigate the consistency of damage registrations. Here it is found that the parameter registrations for damage are generally quite consistent except for transitions, where a parameter is lacking, and for animal burrowing where the distinction between small and large burrowing is hard to make in practice. As it is found from overflow tests (Aguilar-López, Warmink, Bomers, Schielen, & Hulscher, 2018; Steendam, Van Hoven, Van der Meer, & Hoffmans, 2014) that transitions between structures are often places where erosion initiates, a specific parameter for transitions between revetment types and/or structures will be a valuable addition for risk-based inspections. In line with the aforementioned variation in PoD for animal burrowing, it is questionable whether distinguishing burrows with two parameters adds any value.

In general the number of parameters in the applied inspection guidelines is large, while literature shows that inspections with less parameters are generally more reliable (see e.g. Dalton & Drury, 2004; Dirksen et al., 2013; Gallwey & Drury, 1986; van der Steen et al., 2014). Given the considerable overlap between parameters, a valuable improvement towards improving both accuracy and consistency of inspections would be to reduce the number of parameters in the inspection guidelines. However, it has to be noted that, as asset managers typically check the severe damage points themselves, in many cases a suboptimal parameter choice will not have a major impact on maintenance. Asset managers do however consider the reported severity in prioritising damage points for maintenance. In that sense, the inconsistency in severity encountered in the field test is more worrisome: for damage points with reference severity 'bad', only 18% of inspectors registers such a point as 'bad', while approximately 22% registers such points as mediocre or reasonable, and 60% of the inspectors fails to register the point at all (either due to underestimating the severity, or failure to detect the damage point at all). Practically such inconsistencies lead to inadequate maintenance.

The third goal of this field test is to identify potentially important factors for higher or lower inspection accuracy. Due to the relatively limited sample size (22 inspectors) drawing any major conclusions on this is not possible. Using Bayesian Parameter Estimation (BPE) some parameters were identified that might have influence on inspection performance. In general, most of the factors that relate to the work of flood defence asset managers relate to higher

inspection performance. A possible explanation is that such inspectors have more practical experience with inspection and maintenance as a whole, which enables them to better assess different types of damage points and their potential consequences. Another major finding from the BPE is that there was a significant difference between inspectors who inspected section 3 as a pair, versus those who had to do it alone. Initially all inspection rounds were planned to be done by pairs of inspectors, but as some inspectors were not present during the field test it was decided to maximise the number of inspection rounds for each section, rather than reschedule such that there would only be pair inspections (but less inspections overall). For future tests it would be better to have all inspections done by the same number of inspectors to ensure consistency of the data. Additionally, from literature it is found that vigilance and tiredness can also be factors that influence inspection accuracy and consistency. Due to the relatively short inspection times, these factors could not be investigated in this experiment, but might be relevant for future tests.

6. Conclusions

In the field test described in this paper 4 dike sections of 200 metres have been inspected by 22 inspectors in order to estimate the accuracy and consistency of visual flood defence inspections. Approximately half of the inspections was done by a single inspector, the other half by a pair of inspectors. The inspected sections are different, but representative for the variety of flood defences encountered in practice. Three of the sections inspected consist of grass revetments, 1 of the sections contains a block revetment. Each section contains approximately 8 damage points that should be detected by inspectors.

The Probability of Detection (PoD) differs significantly for different damage points as well as between inspectors. For different damage points the PoD in the field test ranges between 0 and 0.9. The PoD for damage to the block revetment is found to be lower, both because block revetments are generally more difficult to inspect, and as high water levels during the test reduced accessibility. For section damages, which are subsets of similar damage points at the same dike section (e.g. all burrows at a certain section), the PoD ranges between 0.3 and 1. There is significant variability between inspectors: for damage points the average PoD of different inspectors ranges between 0.25 and 0.6, while for section damages it ranges between 0.45 and 1. Combined with the large variation between different damage points it can be concluded that defining a single PoD for flood defence inspections is difficult.

It should be noted that the estimated PoD is a lower bound, as in some cases inspectors might have detected a damage but decided not to register a damage point, especially for smaller and less important damages. The registrations by inspectors have therefore been supplemented by observations from supervisors during the test. In future tests it is advised to try and distinguish more clearly between non-detections and non-registrations. It should however be noted that the practical implication is the same: a damaged spot remains unknown to the flood defence asset manager.

The consistency of damage registrations was evaluated based on the registered damage parameter and severity. Although there is some inconsistency in damage parameters, the encountered inconsistency in damage severity is of more importance for risk-based maintenance. For damage points with the highest severity (bad), only 18% registered such damage as bad, while 22% registered it as less severe. 60% did not register the point at all, either due to not detecting it, or due to not classifying it as damage. As asset managers often use the reported severity for maintenance prioritisation, this can have significant influence on the effectiveness of risk-based maintenance. Hence, improving the consistency of severity classification should be an important point of future attention.

A variety of variables that could influence inspection accuracy has been investigated using Bayesian Parameter Estimation. Some indication is found that inspectors who are also asset manager, participate in other types of inspections, and are involved in maintenance tasks, perform slightly better. The fact that none of the investigated variables explains the large variability indicates that variability originates from other sources. Some likely ones are the structure of the currently used inspection guidelines, and general variability among different dike sections. Based on literature (Dalton & Drury, 2004; van der Steen et al., 2014, e.g.), it is likely that simplifying inspection guidelines and tasks will lead to more consistent and more accurate inspections. Concrete improvements based on this test are to reduce the number of (sometimes overlapping) parameters, introducing a consistent parameter for transitions between structures and different revetments, and carrying out specific inspections for damage types with potentially high consequences, such as damage to block revetments or large animal burrowing. Other types of flood defence inspections (e.g. after high water) can likely be improved in a similar way, although their current PoD might be different.

In general terms, the average PoD found from this field test is in line with the PoD reported in literature on other types of infrastructure inspections such as sewers and bridges. Given the stringent reliability requirements for flood defences in most countries, it is doubtful whether it is sufficiently high to ensure these requirements are met. This is something that should be further considered jointly with knowledge on the influence of damage on flood defence reliability. Irrespective of the precise influence of damage on reliability, improving inspection accuracy and consistency leads to better maintenance planning, and is thus likely an effective means to decrease overall flood risk.

Acknowledgements

This work is part of the research programme All-Risk with project number P15-21, which is (partly) financed by NWO Domain Applied and Engineering Sciences. The authors would like to thank Ruben Bruijning for providing help in setting up the test environment in ArcGIS Survey123. Waterschap Rivierenland is gratefully acknowledged for their collaboration in this field test, specifically all inspectors who

participated. We thank the test supervisors for their invaluable contribution during the test.

Data availability statement

All data and code generated or used during the study are available from the corresponding author upon reasonable request.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was funded by Stichting voor de Technische Wetenschappen.

ORCID

W. J. Klerk http://orcid.org/0000-0002-6777-2705
W. Kanning http://orcid.org/0000-0002-9096-3358
M. Kok http://orcid.org/0000-0002-9148-0411

References

Aguilar-López, J. P., Warmink, J. J., Bomers, A., Schielen, R. M., & Hulscher, S. J. (2018). Failure of grass covered flood defences with roads on top due to wave overtopping: A probabilistic assessment method. *Journal of Marine Science and Engineering*, *6*(3), 1–28.

Bakkenist, S., van Dam, O., van der Nat, A., Thijs, F., & de Vries, W. (2012). *Principles of professional inspection - Organizational part (Tech. Rep.)*. Amersfoort, the Netherlands: STOWA. Retrieved from https://www.stowa.nl/sites/default/files/assets/PUBLICATIES/Publicaties2012/STOWA2012-3PIWorganizationalpartUK.pdf

CIRIA. (2013). London, UK: International Levee handbook. ISBN 9780860177340.

Crespo Márquez, A. (2007). *The maintenance management framework: Models and methods for complex systems maintenance* (1st ed.). London: Springer.

Dalton, J., & Drury, C. G. (2004). Inspectors' performance and understanding in sheet steel inspection. *Occupational Ergonomics*, *4*(1), 51–65.

Dirksen, J., Clemens, F. H., Korving, H., Cherqui, F., Le Gauffre, P., Ertl, T., Plihal, H., Müller, K., Snaterse, C. T. M. (2013). The consistency of visual sewer inspection data. *Structure and Infrastructure Engineering*, *9*(3), 214–228. doi:10.1080/15732479.2010.541265

Drury, C. G., & Fox, J. G. (1975). The imperfect inspector. In C. G. Drury & J. G. Fox (Eds.), *Human reliability in quality control* (1st ed., pp. 11–16). London, UK: Taylor and Francis.

Drury, C. G., Spencer, F. W., Schurman, D. L. (1997). *Measuring human detection performance in aircraft visual inspection*. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Santa Monica, CA (pp. 304–308). doi:10.1177/107118139704100168

Flikweert, J., & Simm, J. (2008). Improving performance targets for flood defence assets. *Journal of Flood Risk Management*, *1*(4), 201–212.

Gallwey, T. J., & Drury, C. G. (1986). Task complexity in visual inspection. *Human Factors*, *28*(5), 595–606.

Graybeal, B. A., Phares, B. M., Rolander, D. D., Moore, M., & Washer, G. (2002). Visual inspection of highway bridges. *Journal of Nondestructive Evaluation*, *21*(3), 67–83.

Harris, D. (1966). Effect of equipment complexity on inspection performance. *Journal of Applied Psychology*, *50*(3), 236–237. doi:10.1037/h0023419

Jonkman, S. N., Voortman, H. G., Klerk, W. J., & van Vuren, S. (2018). Developments in the management of flood defences and hydraulic infrastructure in the Netherlands. *Structure and Infrastructure Engineering*, *14*(3), 1–16.

Keprate, A., & Chandima Ratnayake, R. (2015). Probability of detection as a metric for quantifying NDE reliability: The state of the art. *Journal of Pipeline Engineering*.

Klerk, W. J., & Adhi, R. A. (2021). Degradation of grass revetments: A comparison of field observations and structured expert judgement. In *Science and practice for an uncertain future*. Proceedings of the FLOODrisk 2020-4th European Conference on Flood Risk Management, Budapest, Hungary, 21-25 June 2021. Budapest University of Technology and Economics. doi:10.3311/FloodRisk2020.8.2

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573–603. doi:10.1037/a0029146

Long, G., Mawdesley, M., & Simm, J. (2006, June). *Improved approaches to condition assessment - Volume 2: Detailed technical report* (Research Report UR10). Manchester, UK: FRMRC. Retrieved from http://www.fcerm.net/sites/default/files/resources/ur10-vol2.pdf

Moore, M., Phares, B., Graybeal, B., Rolander, D., & Washer, G. (2001). Reliability of visual inspection for highway bridges (Technical Report No. FHWA-RD-01-020). McLean, VA: US Department of Transportation - Federal Highway Administration.

Quirk, L., Matos, J., Murphy, J., & Pakrashi, V. (2017). Visual inspection and bridge management. *Structure and Infrastructure Engineering*, *2479*, 1–13.

Sanchez-Silva, M., Klutke, G. A., & Rosowsky, D. V. (2011). Life-cycle performance of structures subject to multiple deterioration mechanisms. *Structural Safety*, *33*(3), 206–217.

See, J. E. (2012, October). Visual inspection: A review of the literature (Tech. Report SAND2012-8590). Albequerque, NM: Sandia National Laboratories.doi:10.2172/1055636

See, J. E., Drury, C. G., Speed, A., Williams, A., & Khalandi, N. (2017). The role of visual inspection in the 21st century. *Proceedings of the Human Factors and Ergonomics Society*, *61*(1), 262–266. doi:10.1177/1541931213601548

Spencer, F. W. (1996). Visual inspection research project report on benchmark inspections (Tech. Rep. DOT/FAA/AR-96/65). Washington DC, USA: Office of Aviation Research. doi:10.21949/1403546

Steendam, G. J., Van Hoven, A., Van der Meer, J., & Hoffmans, G. (2014). Wave overtopping simulator tests on transitions and obstacles at grass covered slopes of dikes. *Coastal Engineering Proceedings*, *1*, 79. doi:10.9753/icce.v34.structures.79

Ter Berg, C. J. A., Leontaris, G., van den Boomen, M., Spaan, M. T. J., & Wolfert, A. R. M. (2019). Expert judgement based maintenance decision support method for structures with a long service-life. *Structure and Infrastructure Engineering*, *15*(4), 492–503.

van Bergeijk, V. M. v., Verdonk, V. A., Warmink, J. J., & Hulscher, S. J. M. H. (2021). The cross-dike failure probability by wave overtopping over grass-covered and damaged dikes. *Water*, *13*(5), 690. doi:10.3390/w13050690

van der Steen, A. J. v. d., Dirksen, J., & Clemens, F. H. (2014). Visual sewer inspection: Detail of coding system versus data quality? *Structure and Infrastructure Engineering*, *10*(11), 1385–1393.

Waterschapshuis, H. (2016). *Digigids 2016*. Retrieved from http://digigids.hetwaterschapshuis.nl/

Wiener, E. (1984). Vigilance and inspection. In J. Warm (Ed.), *Sustained attention in human performance* (pp. 207–246). Chicester: Wiley.