THE FUTURE OF TRANSPORT

PUBLISHED PROJECT REPORT PPR2042

Road Condition Monitoring Data

Network Study

Alex Wright and Stuart Brittain

Report details

Report prepared for:	Department for Transport		
Project/customer reference	26295 Network Study		
Copyright:	© TRL Limited		
Report date:	26/03/2024		
Report status/version:	1.0		
Quality approval:			
Warsame Mohamed (Project Manager)	W Mohamed	Alex Wright (Technical Reviewer)	A Wright

Disclaimer

Table of Contents

Executive Summary

Road Condition Monitoring (RCM) data forms an important component of the information used by local authorities to understand and manage the condition of their networks. It can be used to report the network-level condition of road networks for both local and national reporting, to identify lengths in need of investigation for maintenance, to decide on the treatments required, and to enable central government to monitor the condition of the national road network.

National reporting of condition has been achieved since 2009 using dedicated survey vehicles collecting data to the SCANNER requirements. However, significant progress has been made in the development of new and alternative technologies for the collection of condition data. These developments provide an opportunity to reconsider the approach taken to the collection and reporting of condition on the local road network. This could lead to a transition to an approach that is less restrictive, in terms of the technologies deployed, offering greater choice to local authorities.

The transition to a new regime for road condition monitoring is to be achieved via the introduction of a new Publicly Available Standard (PAS 2161). In PAS 2161 condition categories will be reported (in the range 1-5) for each 10m sub-section of the network, so that the proportion of each local authority's network falling within each category can be determined. The PAS will help to ensure consistent reporting of these condition categories between highway authorities, whilst enabling a wider choice of technologies for the collection of RCM data.

Consultation undertaken during the initial development of PAS 2161 identified limited knowledge of how current RCM devices report condition categories, what the categories mean in terms of condition, and the level of consistency across the current devices. This suggested that technical challenges may be encountered during the implementation of the new regime. The network study described in this report has sought to provide further insight on the collection of RCM data. The study has drawn on a project stakeholder group that included representatives from DfT, Local Authorities, RCM providers and UKPMS. The group provided input, advice, and guidance throughout the period of the study.

Prior to the study, the PAS 2161 steering group concluded that the PAS should include an end-result requirement for reporting condition in the range 1-5 for each sub-section. The steering group stated that these categories should be defined in a way that does not require RCM data providers to measure specific defects. However, the PAS steering group did not establish definitions for the categories. Therefore, the study commenced by undertaking workshops and research to determine a suitable approach to define the condition categories. It was concluded that the condition categories should indicate the expected level of treatment that may be required to restore a sub-section length to a condition which would require no maintenance. This approach would enable the category definition to be separated from any requirement to report specific types of defects. A set of treatments to define categories 1-5 was therefore proposed.

The work then proceeded to undertake a practical study of the measurement of condition categories on a 100km study network located in Surrey. The study engaged with 11 RCM providers who provided their time freely to undertake a practical demonstration of the

collection of RCM data on this network. A methodology was developed to collect benchmark condition category data using local authority engineers, who also gave their time freely to support this study. Eight engineers collected condition category data using a dedicated App when driven slowly around the study network. The engineers' benchmark data was obtained as the mode of the ratings recorded by the individual engineers.

RCM data was also collected by the 11 RCM data providers using a range of technologies including mobile phone Apps, SCANNER, engineers' inspections, crowdsourcing and dedicated vehicle technologies. To simply the terminology the study has developed an approach to classify the wide range of RCM technologies by type, which may be useful for local authorities considering data collection regimes. The approach taken to calculating the condition categories from their data was determined by the individual RCM providers.

The category data provided by the participants was processed, aligned and compared with the engineers' benchmark data reported over both 10m and 100m subsection lengths. Due to significant differences observed between some of the datasets a calibration phase was undertaken, in which engineers' benchmark data from part of the study network was provided to the participants. This information was used by a number of participants to recalibrate (and resubmit) their data. The key conclusions from the study are:

- The RCM technologies participating in the study show potential for providing consistent condition categories, but there are some differences between systems. The calibration stage, in which providers were able to benchmark their data to a common level, reduced these differences. However, the consistency (across all RCM devices) may not yet be at a level that is comparable with the current devices applied in national reporting of pavement condition. Until this is achieved, greater variability may be expected in cross-year and cross-authority reporting.
- In addition to categories, RCM devices can provide additional data on the defects. However, this is reported in different ways by each device. Care would need to be taken interpreting this data, and in particular when switching between providers.
- The application of engineers' assessments as the benchmark data to support both calibration and performance testing appears to be robust, and this work has proposed a practical approach for the collection of this data. However, this does come with some risks. Any future process to demonstrate an approved set of RCM devices will need to ensure that the benchmark network is representative of the network as a whole, and that the engineers benchmark data is robust. The process should also ensure that the calibration process does not result in a fleet that is "over-tuned" to the network deployed for the demonstration. In addition, attention will need to be paid to ensure that comparisons with the benchmark are not adversely affected by locational errors. This may be achieved through longer reporting lengths (e.g.100m), or through the application of processing tools to better align the data.
- This study has investigated the comparability of current RCM technologies, and proposed a set of tools that could be used to quantify the performance. Whilst it has provided insight into the comparability of RCM data provided by those participating into the study, it did not aim to test the performance of the participants. Further work would be required to establish comparability thresholds, and to undertake approval tests.

1 Introduction

Road Condition Monitoring (RCM) data forms an important component of the information used by road authorities to understand and manage the condition of their networks. It can be used to report the overall condition of road networks, through the application of indicators and network wide statistics. It can also be used to identify lengths in need of investigation for maintenance and, where the data contains a sufficient level of detail, it can be applied to decide on the treatments required.

In England, each Local Authority is required to provide network-level RCM data to the Department for Transport each year, so that the DfT can publish annual official statistics for road condition in England. This process is called national reporting. Since 2009 national reporting of condition has been achieved using dedicated survey vehicles collecting data to the SCANNER[1] requirements. In SCANNER various aspects of condition are measured, and the condition parameters are combined within the Road Condition Indicator (RCI). The percentage of RCI values exceeding specific levels is used for national reporting of road condition in each local authority. However, since the introduction of SCANNER significant progress has been made in the development of new and alternative technologies with the potential to provide RCM data. A number of Local Authorities are already using these technologies to provide information on the condition of their networks. However, these Authorities are also required to undertake SCANNER surveys for national reporting.

The progress in the development of alternative RCM technologies provides an opportunity to reconsider the approach taken to the collection of condition data for national reporting on the local road network. It has been proposed that a new regime could provide Local Authorities greater choice in terms of the technologies deployed for RCM data collection. It could also enable Local Authorities to optimise the value of their RCM data, by removing the need to undertake different types of surveys for local and national reporting of condition.

The transition to a new regime for road condition monitoring is to be achieved via the introduction of a new Publicly Available Standard (PAS 2161). In PAS 2161 condition will be reported as a set of condition categories (in the range 1-5) for the purposes of national reporting. The PAS will help to ensure consistent reporting of these condition categories between highway authorities, whilst enabling a wider choice of technologies for the collection of RCM data. However, consultation undertaken during the initial development of PAS 2161 identified limited knowledge of how current RCM devices report condition categories, what the categories mean in terms of condition, and the level of consistency that is achieved across the currently available RCM devices. This suggested that technical challenges may be encountered during the implementation of the new regime.

This network study has sought to provide further insight on the collection of RCM data, to:

- Establish an understanding of how RCM condition categories should be defined for the PAS.

[1] The SCANNER requirements are published here: https://ukrlg.ciht.org.uk/ukrlg-home/guidance/road-condition-information/data-collection/scanner/

- Establish an understanding of the performance of current RCM measurement technologies in the reporting of these condition categories.
- Understand how the lessons learnt would influence the requirements specified in PAS 2161.

This report discusses this work and the conclusions found.

Acknowledgements

The work presented in this report would not have been possible without the support of the project stakeholder group, which provided input, advice and guidance throughout the period of the study. Members of the following local authorities and organisations supported the group.

Aisin
Amey
Bristol city council[a]
City of Doncaster council[a]
Cyclomedia[b]
DfT
Derby city council[a]
GAIST[b]
Hertfordshire county council[a]
Hochtief
Linhay Consultancy
Metricell[b]
Michelin[b]
NIRA Dynamics[b]
Norfolk County Council
Oxfordshire County Council
PTS[b]
Reading borough council[a]
Roadmetrics[b]
Route Reports
Surrey county council[a]
Vaisala[b]
Waterman Group[b]
WDM[b]
XAIS[b]

In particular, we acknowledge the time freely provided by those marked (a) above for the collection of engineer's benchmark data on the study network, and those marked (b) for undertaking data collection and providing survey data from the study network.

2 Approach and timeline

This study aimed to provide insight to support the development of a PAS that will facilitate the wider deployment of new technologies providing RCM data on the local road network. A network study stakeholder group was established to guide the work. Members of the group included representatives from local authorities, RCM providers and other members of the industry, as listed in Section 1. The goals of the stakeholder group were to guide and advise on:

- How the study would define "RCM condition categories"
- How the performance of current RCM devices in the measurement of condition categories could be assessed using a practical study carried out on the local road network. Note that, whilst the study sought to investigate the performance of current RCM technologies and how this could be quantified, it was not a test of the performance of the participants.
- Where the study network would be located, and its broad content.
- How any required benchmarking data would be collected.
  - Including any requirements for "calibration".
- The analysis and the results.
- The implications for the PAS.

A series of workshops, practical data collection, analysis and reporting was undertaken during the period of the study, as summarised in Figure 2.1. The sections of this report present the work carried out within each stage. Although the programme of work to deliver the objectives of the study commenced in the early summer of 2023, this study built on the recommendations of the group undertaking the development of the PAS (the PAS steering group), as shown in Figure 2.1, and discussed further in Section 3.

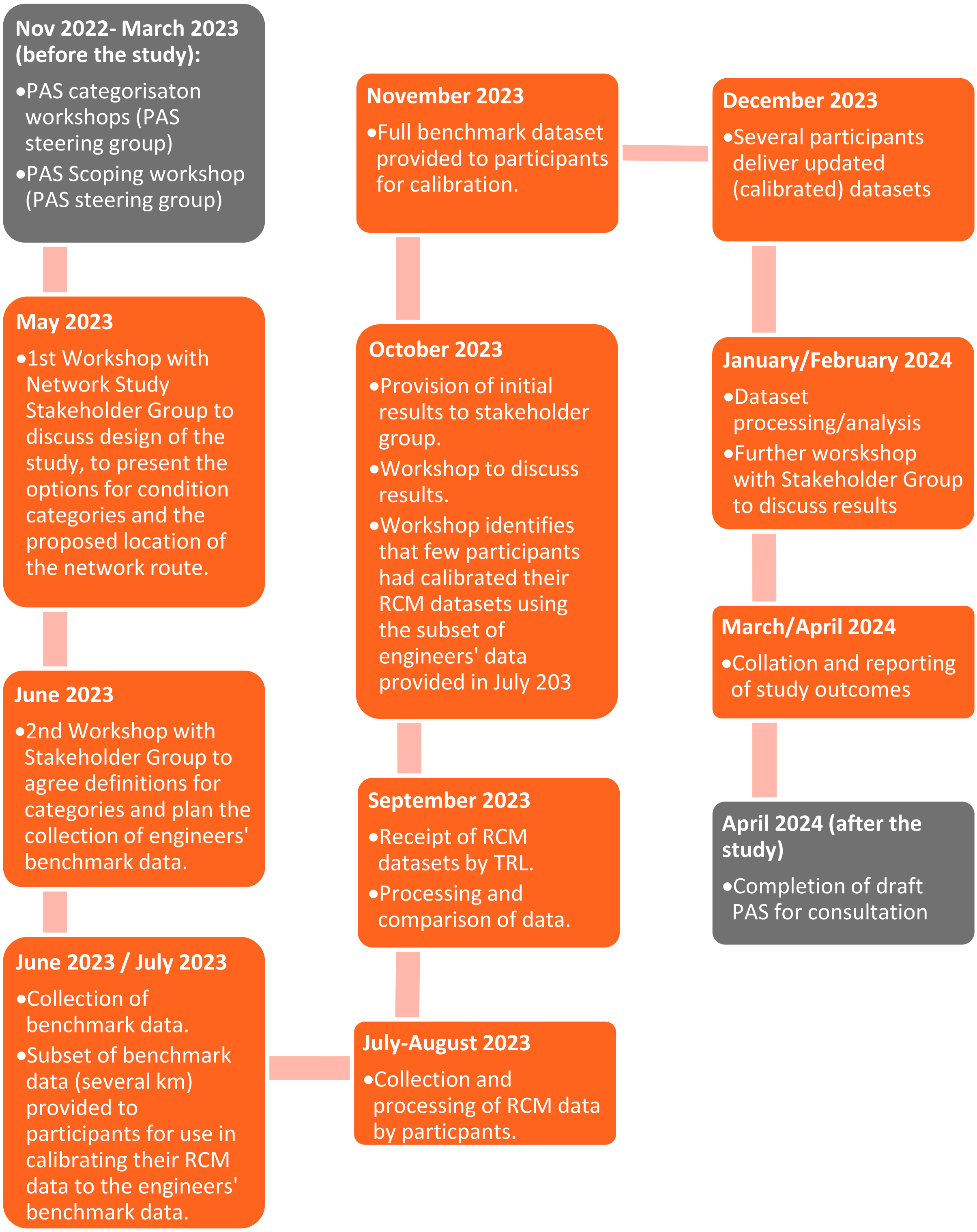


Figure 2.1: Timeline and stages of the network study

3 Preparation for the study

3.1 Condition categories for the PAS

Since the late 2000s Local Authorities have been required to report network condition using data collected by SCANNER survey vehicles[2]. The condition parameters measured by SCANNER (e.g. cracking, rutting) are converted into a Road Condition Indicator (RCI) value which is reported every 10m. SCANNER data is collected for national reporting on local classified roads. SCANNER data is not required to be collected on unclassified roads (although some Local Authorities do apply the survey on these roads). SCANNER is also not required on national (trunk) roads. For national reporting on local classified roads thresholds are applied to the RCI values, which enable the condition of each 10m length to be placed into one of three category bands - Red, Amber or Green. The percentages of the 10m sub-section lengths within the local authority's network that fall into each band are reported.

In the PAS, SCANNER data will be replaced by the collection and reporting of data using a wider range of RCM data collection technologies. However, the RCM data must still be reported in a way such that the condition of local road networks can be compared at a national level. Prior to the commencement of this study, consultation and research was carried out to establish how condition data would be reported for national reporting under the PAS. Workshops of the PAS steering group established that condition should still be reported on a category level, but this would be increased to a five- point scale (1-5, with 5 as the poorest condition rating). PAS steering group categorisation workshops also concluded that an approach such as the RCI (or similar) would be inappropriate to define the new RCM categories. This is because the RCI requires specific condition parameters (such as rut depth) to be measured, and a "ruleset" applied to determine the value of the condition indicator. A requirement for reporting specific condition parameters could lead to an implied requirement for certain types of RCM equipment to be used. This is one of the limitations that reduces the ability to innovate in the SCANNER survey. For example, because SCANNER specifies the measurement of rut depth to a defined resolution and accuracy, this leads to the selection and deployment of specific types of laser profiling technology by SCANNER providers. There was also an understanding that current RCM devices report many different defects, but this was vendor / technology specific, with no common defect "catalogue", which would further complicate the development of rule-based categories (an example is provided later in this report for the additional data discussed in Section 3.5). Therefore, it was concluded that the PAS should include an end-result requirement for reporting condition in the range 1-5 for each sub-section. An end-result requirement (i.e. defining the outcomes, not the method) would not require RCM data providers to measure specific defects, and would allow the providers themselves to determine the methodology used to determine the condition category.

However, the PAS steering group did not establish definitions for the five condition categories. The development of definitions was therefore considered by the network study

[2] The SCANNER specification is published at: https://ukrlg.ciht.org.uk/ukrlg-home/guidance/road-condition-information/data-collection/scanner/

stakeholder group. Different approaches, from the UK and elsewhere, were considered. Many of these surveys record specific defects, which are combined to report a condition category. For example:

- UKPMS Visual Inspection CVI[3]
- UKPMS Visual Inspection DVI[4]
- P-AMS Visual Surveys
- CROW 146a/b[5]

However, other surveys were identified in which assessors directly report condition categories, for example:

- (AEI) Annual Engineers Inspections[6]
- Pavement Surface Evaluation and Rating (PASER)[7]
- The National Highways Vehicle Restraint System condition assessment
- The General Asset Management approach[8]

Although these methods do not require detailed reporting of defects, they do provide guidance to assessors on the level of defectiveness that might be present when making decisions on the required treatment reported for that length, as shown in the examples of Figure 3.1 and Figure 3.2.

4 Fair	Severe surface raveling. Multiple longitudinal and transverse cracking with slight raveling. Longitudinal cracking in wheel path. Block cracking (over 50% of surface). Patching in fair condition. Slight rutting or distortions (½″ deep or less).	Significant aging and first signs of need for strengthening. Would benefit from a structural overlay (2″ or more).
8 Very Good	No longitudinal cracks except reflection of paving joints. Occasional transverse cracks, widely spaced (40′ or greater). All cracks sealed or tight (open less than ¼″).	Recent sealcoat or new cold mix. Little or no maintenance required.

Figure 3.1: Example PASER ratings showing treatment (right) and guidance on the defects associated with that treatment (from University of Wisconsin PASER Manual)

[3] UKPMS CVI Manual: https://ukrlg.ciht.org.uk/media/11969/ukpms_manual_02_07v09.pdf

[4] UKPMS DVI Manual: https://ukrlg.ciht.org.uk/media/11970/ukpms_manual_02_08v09.pdf

[5] Manual can be purchased at: https://www.crow.nl/publicaties/handboek-visuele-inspectie-2011

[6] UKPMS Visual Survey Manual 2017, Volume 2, Chapter 10, Annual Engineer's Inspection (AEI), April 2019

[7] The PASER manual can be found at: https://interpro.wisc.edu/tic/documents/paser-manual-asphalt-pubpas01/

[8] https://openlibrary.org/books/OL21245688M/International_infrastructure_management_manual.

Treatment Recorded	Major Cracking		Major Fretting		Settlement / Rutting		Wheel Track Cracking		Surface Detn		Joint Detn		Edge Detn
Resurface Binder	>20	And / Or	>20	with	>5	And / Or	>10	with	N/A		N/A		N/A
Resurface	>20	And / Or	>20	with	<5	And / Or	<10	with	N/A		N/A		N/A

Figure 3.2: Example AEI ratings showing the treatment that is recorded by the survey (left), and the guidance provided on defectiveness (from RCMG AEI Visual Survey Manual)

These were reviewed and discussed by the network study stakeholder group. It was concluded that an approach to assessing condition that is based on the expected level of treatment would be a practical route to reporting condition in the PAS, in a way which could be separated from the requirement to report specific types of defects. In addition, reference to specific defects provided in the PAS "guidance" (e.g. like the example defects given in Figure 3.1 and Figure 3.2) should be minimised, in preference for a more general description of the condition.

Therefore, the definitions and defects provided in the above examples were reviewed, and a draft set of treatments to define categories 1-5 was proposed. The category names and definitions were reviewed and refined by the network study stakeholder group and finalised ahead of the commencement of the network study data collection. These are shown in Table 3.1, and are subsequently referred to as **RCM condition categories**. Note that the RCM condition category definition asks what level of treatment may be required to restore the pavement to a level not requiring treatment[9]. This takes into account the variations in condition expected for different road classes, as different levels of restoration may be considered appropriate for (e.g.) an A road than a C road. For the deterioration, references are not made to specific defects, but to a level of defectiveness, although potholes are included due to their significant influence on the condition of local roads.

[9] Note that the original definition was for the required to restore the condition to "as new". This was later revised to a condition where no treatment is required.

Table 3.1: RCM condition categories

	Category*	Example deterioration	Action to restore to a level of condition not requiring treatment
1	**No deterioration.**	Pavement appears in good condition and provides a smooth running surface.	Pavement not considered for maintenance.
2	**Minor (and/or Aesthetic) deterioration.**	Aesthetic and/or minor deterioration, associated with very early signs of ageing. Pavement presents superficial defects. Occasional/minor ride quality issues.	'Light touch' maintenance. (e.g. minor patching).
3	**Moderate deterioration.**	Pavement is showing notable signs of ageing. Non-structural defects visible Evidence of minor potholes or development of potholes. Ride quality is deteriorating.	Localised intervention or mid-life preventative maintenance. (e.g. surface dressing, patching, crack sealing, patching of anti-skid surface).
4	**Moderate to severe deterioration.**	Pavement is substantially aged. Defects across a significant extent of the surface. Evidence of failed maintenance/ patches. Indications of structural failure/onset of structural failure. Major potholes, Ride quality is compromised.	Preventative, perhaps full carriageway maintenance. (e.g. resurfacing, with thin overlay/surface dressing, multiple patching, edge haunching, renewal of anti-skid surface).
5	**Severe deterioration.**	Extensive deterioration/defects. Structural failure. Serviceability may soon be compromised, or is no longer serviceable. Very poor ride quality.	Significant maintenance is required, which is likely to include full carriageway resurfacing or reconstruction.

3.2 The study network

Surrey County Council offered to assist in establishing a study network on their network, and provided recent condition data (SCANNER), which was viewed alongside online maps and images to identify sections which would be appropriate to include in the study network. For practicality, it was decided that the study network would be structured around a single main route. This would simplify the collection of data for both RCM and benchmark technologies. However, in addition to this single main route, two further “loops” would be included on which to collect repeat data. To optimise practical collection, RCM data could be collected on the main route and then two further sets of collection undertaken on the repeat loops. As the repeat loops would be a subset of the main route, this would provide a full set of data from the main route and three datasets from the repeat routes.

An initial study network was established around Guildford in Surrey and discussed with the study stakeholder group. It was agreed that the network should include more U road lengths, which were therefore added to the study network. TRL undertook a survey of the study network using the HARRIS3 survey vehicle (owned by National Highways) to confirm that it was suitable for a driven survey, and to obtain images of the network for later reference. The study network is shown in Figure 3.3. The network contained A, B, C and U roads at an approximately ~35%;25%:25%:15% split (Red, Amber, Green and Black in Figure 3.3), and was approximately 100km in length. The repeat loops, totalled around 30km. Example images collected on the study network are provided in Appendix B to demonstrate the range of road types included. The network was defined using the OSGR coordinate centre line (obtained from the HARRIS3 survey) and a set of sections, defined by their start and end coordinates and section lengths.

Note that, although the study network was selected to cover as many different types of roads/features as practical, the network was probably not representative of all roads and features that may be seen on the wider English road network. The steering group agreed that a robust exercise would ideally cover a range of authorities, so that a more representative sample of the roads is included. This was not practical within the timescale and resources of this study, but should be considered for future exercises, such as when formally approving/demonstrating the performance of RCM technologies for the purposes of national reporting.

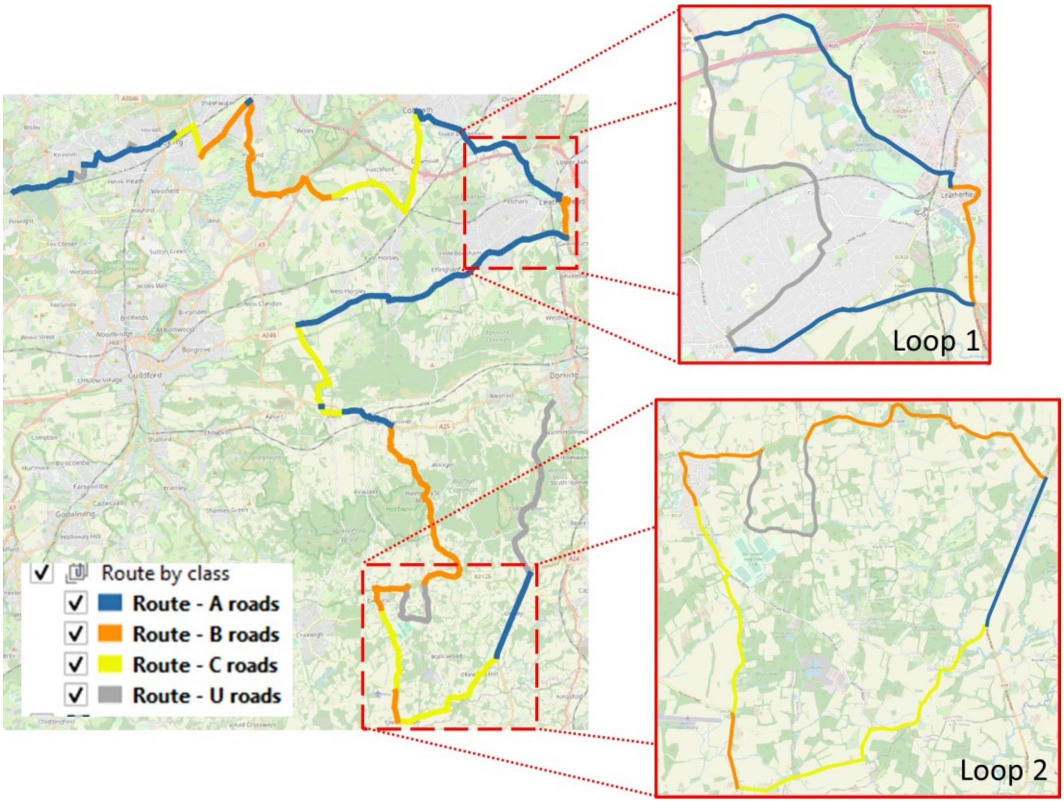


Figure 3.3: The study network, including repeat loops 1 and 2[10]

3.3 Benchmark surveys

3.3.1 Benchmark methodology

To undertake the study there was a need to establish "benchmark" condition categories for each sub-section. The study aimed to compare these benchmark categories with the categories reported by the individual participants to understand comparability and consistency. Consequently, there was a need to agree the method that would be used to provide the benchmark RCM condition categories. The study stakeholder group considered the following options:

1. Engineers' assessments could be carried out to on the network to establish the benchmark condition category values.
2. The consensus of the condition category values reported by RCM devices themselves used to establish the benchmark category values.
3. A hybrid of the above options.

[10] OpenStreetMap contributors. (2015). Map data copyrighted OpenStreetMap contributors and available from https://www.openstreetmap.org

The group anticipated that, without a calibration exercise to bring RCM data onto the same scale/distribution, there would be a large spread of condition categories reported by different RCM technologies. This might make it difficult to establish a consensus from these technologies, due to difficulties determining which (if any) that should be excluded as outliers. Also, even if a consensus was achieved, there would be no basis to show how the RCM categories related to the categories that would be provided by maintaining engineers (and which would be more familiar/acceptable to local highway authorities). This would present the risk of RCM condition category ratings differing from local authority assessments, undermining confidence in the RCM data.

Therefore, it was decided that Engineers' assessment would be used for the network study. However, the potential to apply option 3 (hybrid of engineers' and RCM data) would remain, with the ultimate decision on this option depending on the comparison (consensus) between the categories reported by the RCM providers and between the categories reported by RCM and the engineers' categories.

3.3.2 Design of the benchmark data collection

Although the steering group had identified a risk that the categories reported by RCM devices may differ, it was recognised that engineers' assessments also often differ. Therefore, the design of the benchmark data collection adopted aspects of the second of the options discussed in Section 3.3.1, in that the collection would be carried out by several engineers, and the consensus opinion on each length would be used as the benchmark condition category. This would (at least partially) overcome the differences between engineers, and (if the engineers are representative from local authorities) provide a reasonable level of confidence that the benchmark data is representative of the level of condition that would be reported by local authorities. Therefore, the design of the benchmark data collection was as follows:

- An Application (App) was developed for a tablet computer that would enable Local Authority engineers to report condition categories when being driven (slowly) along the study network (Figure 3.4).
- Engineers would report their condition category rating as a continuous assessment. Every time that the engineers considered the category to have changed, they would record a new rating on the tablet.
- The tablet would continuously record position (grid coordinate) and the current condition category rating. This data could be processed after the survey to obtain a continuous benchmark dataset for each engineer (see Section 4.2).

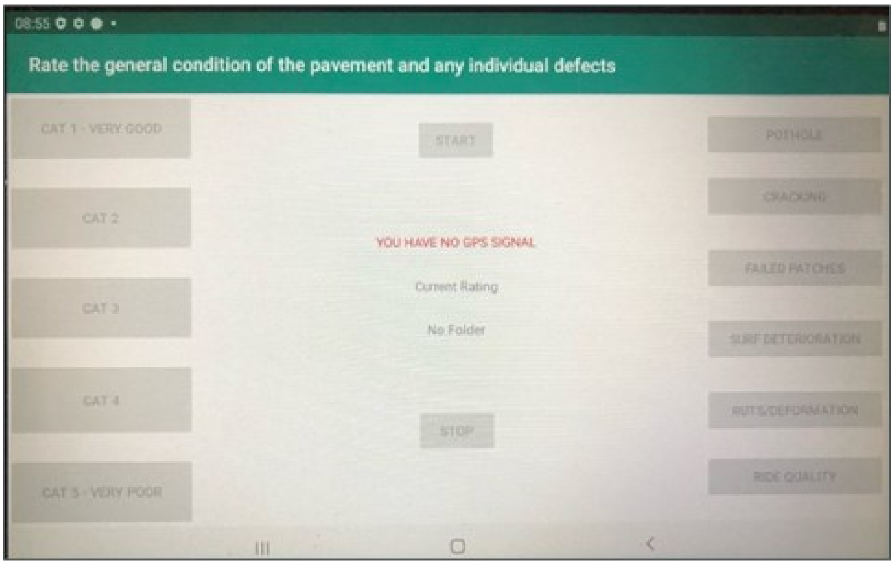


Figure 3.4: Tablet application used for engineers' assessments[11]

To reduce the risks of inconsistency between engineers' assessments, attention was given to the advice/training which would be provided to the engineers undertaking the benchmark data collection. The following was noted:

- Different road classes are typically maintained to different quality levels. If the same criteria for assessing condition was applied to all roads, then A roads could tend to be rated higher, with B, C and U roads tending to score progressively lower on the scale. Therefore, engineers should be asked to take into account the expected condition for different road classes when reporting condition categories. The condition categories should be based on the difference between the current and expected ('new') condition for that class of road. For example, an A road might have had better 'new' ride quality than a U road.
- Condition categories should indicate/rate the 'need' for treatment regardless of context or priority of the specific sub-section of road under assessment. However, this might not reflect what engineers do in their day-to-day activities. Therefore, engineers should specifically be advised to disregard the following aspects when determining the condition category:
    - Strategic priority of the road i.e. a main artery A road of high strategic importance should be considered to require the same treatment as a low strategic importance A road.
    - Location or environment of the road i.e. urban and rural roads are treated the same.
    - The level of traffic that the road is taking – higher trafficked roads would not be given better treatment or have higher priority for treatment.

[11] Note that the Application also enabled engineers to report the presence of any significant defects that might have influenced the category rating they reported. This data was not used in the comparison process.

- Potential maintenance prioritization or other asset management considerations that would form part of treatment selection or prioritization.

Drawing on these observations, a guidance pack was developed for all engineers undertaking the benchmark condition data collection. This is shown in Appendix A.

Note that, for local road asset management, considerations such as context, environment and priority are factors that influence maintenance decisions. However, information on these factors is unlikely to be available to RCM technology when determining a condition category. It is anticipated that the Local Authority would consider these (and other) contextual factors when applying the RCM condition categories in asset management decisions,

3.3.3 Benchmark data collection on the study network

Benchmark data collection was undertaken by eight Local authority engineers (as listed in Section 3.3). For these surveys the engineers were provided with advice/training by TRL before being driven over the study network to record condition categories using the tablet App. After the survey engineers were asked to provide feedback on their experience so that any refinements required for the remaining surveys could be identified and considered for the next set of engineers' surveys.

Following the first two engineers' surveys (which were conducted on the same day) it was concluded that the advice being provided in the training was not sufficiently clear with regard to the need to focus on the treatments that may be required, and that the data collection should not be considered to be a 'defect survey'. Therefore, closer attention was given in the training / briefing sessions for the engineers for subsequent surveys.

3.4 RCM data collection

3.4.1 RCM technology types

There are many different technologies/approaches used for the collection of RCM data. To assist in understanding the general approach to data collection taken by any RCM technology the project developed definitions of "RCM technology type". Although it is noted that new technologies may be developed in future, the current RCM technologies were identified to fall within five categories, with a sixth category for future devices. These are presented in Table 3.2, with examples shown in Figure 3.5- Figure 3.8 (these examples are taken from the literature of the participants in this study)

Table 3.2: RCM technology types

Type	Description	Notes
S Dedicated survey vehicles (Figure 3.5)	Individual survey vehicles either permanently or temporarily equipped with dedicated high resolution measurement equipment for the specific collection of RCM data.	Typically deploying dedicated distance measurement devices and inertial GNSS to provide high locational accuracy, and industrial image, laser profile/LiDAR or other dedicated measurement devices.
M Dedicated mobile technologies (Figure 3.6)	Individual, mobile, portable measurement equipment typically installed in a vehicle on a temporary or permanent basis for the specific collection of RCM data.	Typically deploying small scale image collection, such as mobile phone or dash-cam technologies supported by GNSS, running bespoke Applications. Collection is managed/controlled by data collector operatives, or operatives trained in the use of the system.
A Application-based mobile technologies	The use of data obtained from mobile phone applications, but using sources not specifically installed for RCM.	Typically obtained via remote access to data collected by applications running on mobile phones, or other similar devices
V Vehicle telematics (Figure 3.7)	The use of vehicle telematics data (provided via instrumentation not specifically installed for RCM)	Typically obtained via remote access to vehicle on-board telemetry data.
E Engineers' inspection (Figure 3.8)	Driven or walked engineers' inspections.	Typically obtained via measurement of condition using dedicated walked or driven engineers, inspections.
O Other technology	Data collected using technology not able to be described by one of the other types listed.	This can include technologies such as remote sensing, aerial, satellite inspections, etc.

Figure 3.5: Dedicated survey vehicles (e.g. bespoke vehicle installations)

Figure 3.6: Dedicated mobile technologies (e.g. dedicated mobile phone App)

Figure 3.7: Crowdsourcing (e.g. using vehicle telematics)

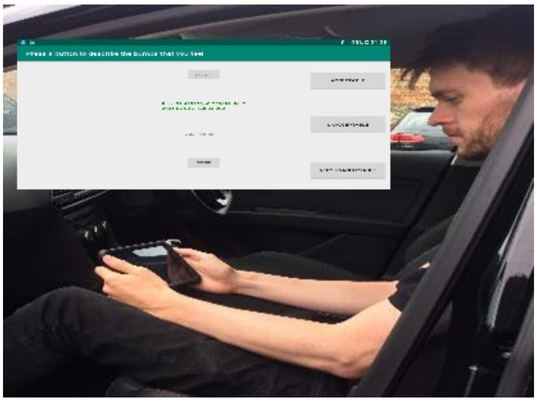


Figure 3.8: Engineers inspection (e.g. driven using a data collection device)

3.4.2 Participants in the study

Eleven RCM providers participated in the study. Each was invited to join the study stakeholder group and participated in the discussion on the development of the category definitions, benchmarking process and the design of the study network. The participants are listed in column 1 of Table 3.3, along with the RCM technology type deployed by that participant (column 3). The additional content of Table 3.3 is discussed further in Section 4.

Note that in the remaining Sections of this report the datasets have been reported using a letter-based referencing system so that the provider of each dataset is anonymised. Anonymisation has been applied because the intention of this work was to provide a general understanding of the current situation regarding current RCM devices, and not report the capability of any specific current system. However, it should also be noted that the study aimed to provide insight into the comparability of RCM data provided by those participating, it did not aim to test the performance of the participants.

Table 3.3: RCM participants

Company	Dataset	RCM type	Additional data provided?	Comments
Cyclomedia	RSA	S	Yes	
	PASER	S		
Gaist	Gaist	S	Yes	
Metricell	Metricell	M		
Michelin	Roadbotics	M		
Nira Dynamics	Nira Dynamics	V		
PTS	MFV – RCI method	S		
	MFV – PCI method	S		
	Tempest – RCI method	S		
Roadmetrics	Roadmetrics	M		
Vaisala	Vaisala	M	Yes	
WDM	WDM Method 1	S		Based on SCANNER RCI using the average condition for each 10m subsection.
	WDM Method 2	*S*		*Method 1 but using worst condition. Similar to method 1 - excluded from the analysis.*
	WDM Method 3	S		Similar to method 1 but using additional bespoke measures.
	WDM Method 4	*S*		*Method 3 but using worst condition. Similar to method 3 - excluded from the analysis*
Waterman group	Waterman	M		
XAIS	AEI	E		
	EyeVi	S		

3.5 Additional data

Most participants in the study are able to provide additional data, as well as condition category values. This additional data typically takes the form of parameters describing the location, extent and severity of defects (e.g. cracking, pavement unevenness). These parameters are typically applied by the participant to determine the condition category rating that they report. As shown in Table 3.3, some RCM providers included additional data when providing their datasets. However, additional data was not a core component of the study and was not used in quantitative assessments.

Examination of the additional data provided showed that:

- The additional provided covered similar types of defect parameters – such as Cracking, Surface deterioration/potholes and Fretting/Ravelling
- The additional data is reported in different ways, this included reporting as a simple "present/absent" value for each defect type, or a "present/absent" value for different severities of the defect type, or the specific extent of the defect (as a percentage)
- Some additional data datasets use the name of the defect to differentiate between the type and severity of the defect e.g. minor longitudinal cracking, moderate longitudinal cracking.

Because the additional data was provided in different ways there was a need to undertake some manipulation to enable the additional data provided by different RCM providers to be compared. In this case we have attempted to present the additional data datasets on a common scale of 0-100. To achieve this the datasets that provided simple present/absent indicators were converted into 100 for present and 0 for absent. Datasets which provided a present/absent indicator for different severities of defect required some judgment to select values that could be assigned to each severity so that they could be compared with defect data from other RCM providers on comparable scales. An example of the outcomes of the data manipulation is presented in Figure 3.9 for the additional data "cracking".

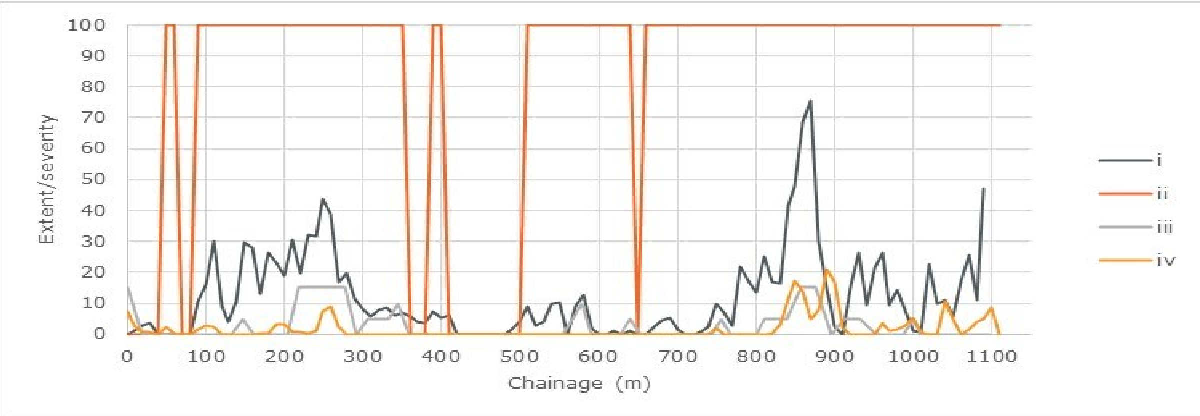


Figure 3.9: Example of comparison of cracking data

From these simple plots it can be seen that the additional data (cracking) provided by the four participants generally showed where greater amounts of this defect were present on the example length (and this was typically the case for the data examined, although we have not attempted quantitative analyses). However, as can be seen from Figure 3.9, the way in which the data is delivered differs by device. The intensities reported (or interpreted though our analysis) differ significantly between devices. This is because there is no standardisation regarding the way the additional data is reported. This leaves interpretation (what is high, medium, low) to the local authority, or (more likely) to a discussion between the local authority and the RCM provider.

Clearly, additional data could be useful to better understand condition category values, assist in maintenance scheme development and programming, and the steering group agreed that it should be included as desirable, but not compulsory, data in the PAS. However, the PAS will not include any requirement to demonstrate performance or consistency in the reporting of additional data. End users would need to understand that additional data provided by different technologies may not be directly comparable. It may not be straightforward to use this data interchangeably (for example switching between the data supplied by different RCM providers in different), without additional work to understand the differences between datasets.

4 Approach to analysis

4.1 Two datasets delivered for analysis

At the commencement of this study it was anticipated that the current RCM devices might provide very different interpretations of the condition categories. This is because there has been no previous attempt to standardise the classification of condition reported by such a wide range of technologies on English Local Roads. To reduce the potential level of variation across devices it was proposed that the participants would be provided with sample data. This would enable them to undertake a calibration exercise to place their data into the context of the benchmark, and (hopefully) reduce large scale differences. Therefore, at an early stage in the processing phase (June 2023), the participants were provided with data, as described in Section 4.2 below, which we refer to as the *initial calibration data*.

Following the delivery of the RCM category data by the participants in September 2023, significant differences were observed between some of the datasets and the benchmark. A workshop was held with the study stakeholder group to discuss the initial results in October 2023. The workshop identified that a range of approaches had been taken by the participants, some applying the initial calibration data and other using their own benchmarking methods to calibrate their RCM data. It was agreed that a further calibration dataset would be provided to the participants. as described in Section 4.2 below, which we refer to as the *further calibration data*.

Hence, two datasets were considered in the analysis – the *initial RCM data* (the RCM data following provision of the initial calibration data) and the *further RCM data* (the RCM data following provision of the further calibration data).

4.2 Benchmark and calibration data

4.2.1 Obtaining the benchmark

As discussed in Section 3.3, benchmark data was obtained using surveys undertaken by eight local authority engineers. As a condition category rating was reported by the engineers at each position along the network where they identified a change from the current category, the categories reported in the engineers' data were provided over un-equal lengths. Therefore, the condition categories recorded using the tablet App were resampled to report a condition category for each 1m length of the study network, along with its OSGR grid coordinate, for each engineer. The definition of the centre line of the study network (OSGR grid coordinates and chainages within each section provided by the HARRIS3 survey) was then applied to each engineers' dataset to fit their data the network and report 1m engineers' ratings relative to section and chainage.

The benchmarking dataset took the form of a single set of ratings, for each 10m length and for each 100m length of the network. These were obtained as the mode[12] of the 8x1m

[12] See Appendix C for a note on the application of the Mode in this work.

engineers' ratings reported in each 10m (or 100m) length. It should be noted that, during the benchmark surveys the traffic, environment and other practical constraints led to some lengths of the network not being surveyed by all 8 engineers. To maximise the robustness of the data, only lengths where at least 6 engineers' ratings were available were used in the analysis. Further, on review of the engineers' ratings it was found very little of the network had been reported as category 5. Therefore, for the analysis the number of categories was reduced to 4 by merging categories 4 and 5 into one rating "category 4/5".

4.2.2 Calibration data

The initial calibration data provided to participants contained:

- The distribution of the 100m ratings for the whole network (i.e. the percentage of the network that fell within categories 1, 2, 3, 4 and 5.[13]
- A sample of the modal 100m engineers' category ratings (i.e. the benchmark ratings) for 5.8km of the study network.

The further calibration data provided to participants contained the entire set of modal engineers' category ratings for the whole of the study network.

4.2.3 "Cleaning" the mode of the benchmark data (for future assessments)

During the preparation of the benchmark data it was noted there is potential for there to be lower levels of consensus on some lengths. The question was raised as to whether lengths with lower levels of consensus should be excluded. Two parameters were proposed to quantify the level of consensus - the overall confidence factor (OCF) and lower confidence factor (LCF).

- The overall confidence factor measures the percentage of individual engineers' condition category ratings that match the reported benchmark condition category rating.
- The lower confidence factor measures the percentage of individual engineers' condition category ratings that fall within one condition category of the reported benchmark condition category rating.

These are calculated for each sub length (10m or 100m) in the reference dataset using equations 1 and 2.

$$OCF_x = \frac{\sum_{BCC_x}^{BCC_x}(BTCC_x)}{BCC_x} \qquad [1]$$

[13] For the initial calibration data an exercise was carried out to removal of outlier engineers' ratings from each 100m length before calculation of the mode. However, this was found to have little effect on the reported modes and therefore was not applied in the general analysis or in the provision of the further calibration data.

$$LCF_x = \frac{\sum_{BCC_x-1}^{BCC_x+1}(BTCC_x)}{BCC_x} \qquad [2]$$

Where:

1) BCC_x is the number of condition category ratings provided for sub-section x
2) $\sum_{BCC_x}^{BCC_x}(BTCC_x)$ is the number of condition categories reported by the assessors for sub-section x which were the same as the benchmark condition category for sub-section x.
3) $\sum_{BCC_x-1}^{BCC_x+1}(BTCC_x)$ is the total number of condition categories reported which are within one condition category of the benchmark condition category for sub-section x.

Visual examination of the data suggested that these could be applied to remove lengths where there appeared to be less agreement/consensus between engineers' ratings. A threshold of 0.5 for the OCF and 0.8 for the LCF appeared to be suitable. These thresholds can be read to require that:

- 50% of the ratings should match the mode for the length to be included.
- 80% of the rating should be within one of the mode for the length to be included.

However, note that:

- Application of the above thresholds to the 10m benchmark dataset showed that ~15% of lengths would be remove as having low consensus.
- As noted in Section 3.3, following the first two engineers' surveys updates were made to training/briefing to make it clearer that it was not to be a 'defect survey'. Application of the above thresholds to the 10m benchmark dataset following the removal of the first two engineers' datasets showed that only 6% lengths would then be removed as having low consensus.
- The outcomes in Section 5 have used data *without* the removal of potential low consensus lengths.

4.3 Processing of RCM datasets

4.3.1 RCM condition category requirements

To assist them in determining the category ratings the participants were provided with the table defining the condition categories, a copy of the advice that was provided to the engineers ahead of the collection of benchmark data (Appendix A), and the initial calibration data discussed above. The determination of the category value (e.g. from the "raw" data collected by each participant's equipment) was the responsibility of the participant, as was ensuring that the reported category values were correctly fitted to the study network.

The participants in the study provided condition category ratings on the 1-5 scale for each 10m length of the study network. On receipt, the data was reviewed to confirm that it was reported in 10m sub-section lengths, that it was fitted to the network provided for the study and that the direction of the data matched the section direction. If the data was unsuitable, then this was discussed with the participant and an updated dataset provided.

Fitting to the network was achieved using co-ordinate data, which was estimated to have errors of at least a few m for each device. It was noted that there could be significant misalignment between 10m lengths, and concerns were raised over the ability to robustly compare the 10m datasets spatially (i.e. length for length). Hence, it was decided that comparisons between engineers' and RCM ratings would also be carried out over 100m lengths, to minimise the influence of location referencing. The 10m category data provided in each RCM dataset was converted into 100m ratings by resampling the 10m category data into 1m lengths (so that shorter sub lengths could be weighted appropriately) and then taking the mode of these values for each 100m length. This approach was consistent with the method applied to convert the engineers' ratings into 100m lengths. As for the benchmark data, the number of categories was reduced to 4 by merging categories 4 and 5 into one rating "category 4/5" for the analysis.

Note that some participants decided to apply more than one approach when determining the condition category from their raw data. Participants that provided multiple datasets (and the data provided) are shown in Table 3.3. The analysis showed that some of these datasets were very similar to each other, and therefore some of these were excluded from the presentation of the results. Excluded data is shown in grey italic text in Table 3.3.

4.4 Comparison measures

A number of methods were considered as tools to quantify the consistency between the datasets. These were considered in context of the future need to establish objective methods to demonstrate the performance of RCM technologies (which could be specified in the PAS). The comparisons were separated into two types. The comparison of the distributions (Section 4.4.1) quantifies the agreement between the proportions of the study network reported in each condition category by the benchmark and by the participant. This is a non-spatial assessment, which does not compare the data on a length for length (or sub-section by sub-section basis). It provides an indication of the consistency at the network level. The spatial comparisons (Section 4.4.2) align the benchmark and participants datasets and directly compare the condition categories reported in each sub-section. This provides an indication of the consistency at the sub-section (spatial) level.

4.4.1 Comparisons of distributions

The first level of comparison considers the reporting of condition categories in terms of the number (or percentage) of lengths reported in categories 1-5 by the RCM technology and by the benchmark.

4.4.1.1 Chi-square

The Pearson's Chi-square test can be applied to compare distributions of categorical data (i.e. where the data falls in categories such as our condition categories 1-5). It tests whether a measured distribution (in this case the RCM distribution of condition categories reported on the study network) is representative of the expected distribution (in this case the benchmark distribution of condition categories reported on the study network). The Chi-square test statistic is relatively simple to calculate by comparing the number of values reported in each category in the test and benchmark distributions (Equation 3):

$$X^2 = \sum \frac{(O - E)^2}{E} \qquad [3]$$

Where

- *O* is the test frequency and E is the benchmark frequency for each category.

The Chi-square test statistic can be compared with a critical value, which is determined by the level of confidence required and the degrees of freedom (which can be calculated from the number of categories). If the value is less than the critical value then the difference between the test and benchmark distributions is not statistically significant.

Instead of/in addition to comparing with the critical value from a look up, the test can also be undertaken through direct calculation of the P-value (essentially, quantifying the distribution used to generate the critical values).

4.4.1.2 Pearson

The Pearson correlation coefficient describes the strength of the relationship between two datasets that are assumed to be linearly related. Here we assume that the relationship between the RCM and benchmark distributions of condition categories reported on the study network should be linear (with the ideal case being a 1:1 linear relationship) and assess this using the Pearson correlation coefficient.

4.4.1.3 Delta value

The delta value is a further test of consistency which compares the numbers of sub-sections reported in each condition category. It calculates the difference between the number of sub-section lengths (e.g. 10m) reported in that category by the benchmark and by the RCM, and reports this as a proportion (fraction) of the total number of lengths reported as that category by the benchmark. Lower values suggest higher levels of agreement. The delta value for each condition category is hence calculated in accordance with Equation 4:

$$\Delta C = \frac{|A_C - B_C|}{B_C} \qquad [4]$$

Where:

- ΔC is the Delta value for condition category c.
- A_c is the number of sub-sections in the RCM data reported as condition category c
- B_c is the number of sub-sections in the benchmark data reported as condition category c

4.4.1.4 *Difference between the total percentage recorded by the benchmark and RCM dataset in condition categories 4 and 5*

This is a broad assessment of the agreement between devices when the reporting of the proportion of the study network that falls in the poorer condition categories. It does not provide a detailed assessment, but does reflect the approach to national reporting that has traditionally been used – “the percentage red”. Therefore, this assessment compares the (combined) percentage reported in categories 4 and 5 by the benchmark and RCM.

4.4.2 *Spatial comparisons (sub-section vs sub-section)*

Whilst the above measures consider the distributions of categories reported by the benchmark and the RCM, the next stage of the comparison directly compares the datasets on a sub-section by sub-section basis. This asks the further question – *“do the RCM condition ratings reported for a specific location match the benchmark rating?”* The study has used confusion matrices to answer this question.

4.4.2.1 *Confusion matrices*

Confusion matrices quantify the number of corresponding lengths reported in each category by the RCM that are also reported in the same category by the benchmark. Conversely, it also quantifies how many of these corresponding lengths were reported in a different category. For the network study we consider 4 condition categories (1, 2, 3 and 4/5 (combined)). The confusion matrix is a four-by-four table (Table 4.1).

Table 4.1: Example layout of a confusion matrix

		RCM (Assessed)			
		1	2	3	4/5
Ref (Bench)	1	a	b	c	d
	2	e	f	g	h
	3	i	j	k	l
	4/5	m	n	o	p

In Table 4.1:

- a = the percentage of lengths reported as condition 1 in the benchmark (reference) dataset that was also reported as condition 1 in the RCM (assessed) dataset
- b = the percentage of the data reported as condition 1 in the benchmark dataset that was reported as condition 2 in the RCM dataset
  - etc
- High percentages on the diagonal (a, f, k and p) correspond to the two datasets reporting similar condition categories for the corresponding lengths in each dataset (i.e. length x was reported as category 1 by both systems).
- High values in the cells located above the diagonal suggests that the RCM reported a significant proportion of lengths to be in a worse condition than the benchmark. High values in the cells located below the diagonal suggests that the RCM reported a significant proportion of lengths to be in better condition than the benchmark.

In a confusion matrix the values can be shown as counts (number of lengths) or as percentages (percentage of the total number of lengths reported as that category). In this study we present percentages. Therefore, the rows in each table should add up to 100 (or close, as the data is reported to zero decimal places). This enables direct comparison between the matrices obtained on different classes of road (as there were a different number of lengths in each class). However, the percentages can confound the interpretation of the matrix where there were very few lengths in that category. E.g. if there were only 5 lengths reported as category 4/5 by the benchmark and 4 were misreported by the RCM, the value of p would be only 20%, but the test was actually very demanding and only a few (4) lengths were "missed" by the RCM device.

4.4.2.2 Precision values

Two "precision" values have been adopted to assist in collating the data from the confusion matrix into a single number. The precision values are:

- "precision 0": The percentage of lengths for which the two datasets provide the same value (the total % of cells a, f, k, p) – e.g. RCM reported "3" benchmark reported "3"
- "precision 1": The percentage of lengths where the category reported by the RCM was within one category of the benchmark (the total % of cells b, g, l, a, f, k, p, e, j, o) – e.g. RCM reported "3" benchmark reported "2", "3", or "4".

5 Analysis of the category data

The study collated an extensive set of data comparing the RCM and benchmark condition category ratings. For succinctness, we present the collated 100m updated data in Appendix D .The data are separated into A, B/C and U roads, which reflects the way in which condition data is currently separated for national reporting. Within this chapter we present selected data and observations from the study, including comparisons between the initial and updated data where appropriate. Also, as discussed above, the study obtained condition category values reported over both 100m and 10m sub-section lengths. Whilst Appendix D presents the 100m data, this chapter discusses the relevant differences observed between the 100m and 10m datasets.

5.1 The participants providing updated RCM data

Participants B, H, I, K and O provided updated RCM datasets following the circulation of the further calibration data, as discussed in Section 4.1. These participants took different approaches when providing their updated RCM data:

- One participant stated that they had not utilised the information contained within the further calibration datasets. Instead, they had undertaken a general reassessment of their approach for categorising the network in the light of the overall levels and distribution of categories, and used this to refine their RCM data.
- One participant stated that they had applied a “global optimisation algorithm” to their data following the second workshop.
- One participant stated that they had investigated the approach taken to obtain categories from their raw data and identified that better agreement with the benchmark could be obtained by revising the way they in which classified they severity of some defects (before determining the category).
- One participant utilised the further calibration dataset to train a machine learning tool. The machine learning tool was then applied to determine the updated categories from the data collected using their technology. This participant provided three new RCM datasets, that were trained using 8%, 50% and 80% of the further calibration data. The results presented in this work are from the dataset trained using 50% of the further calibration data[14].

5.2 Comparison of the distributions

Figure 5.1 shows the distribution of 100m categories reported on the A roads of the study network by the benchmark and the participants. The upper plot shows the initial RCM data, and the middle plot shows the updated data. As only participants B, H, I, K and O provided updated data, only the data from these participants has changed. Corresponding plots are also provided for B/C roads in Figure 5.2. It can be seen that the benchmark reported very

[14] As the further calibration data contained all of the benchmark condition category ratings from the study network, this therefore used 50% of the study network for training.

few lengths as category 4/5. This is a noted limitation of the study and constrains the ability to undertake robust numerical comparisons of the benchmark and RCM datasets.

In the initial data (upper plots) the RCM data provided by several of the participants follows a similar pattern to the benchmark. However, there were some notable outliers. For example, datasets A, C, K and P reported a noticeably greater extent of the network in category 1 in comparison to the benchmark. In addition, dataset I reported a large percentage of the network in category 3 for all road classes. Furthermore, although some participants' datasets reported visually similar distributions to the benchmark for A roads, there were noticeably greater differences in the distributions reported on B and C roads. An example of this is dataset O, which identified very little of the B and C roads as category 1 (whereas category 1 was the largest category for all road classes). Following the recalibration and the provision of updated RCM data for participants B, H, I, K and O there is improved agreement. The two clear remaining outliers (e.g. A and P) did not provide updated data. The visual agreement also seems to decrease on B/C roads.

The above observations are for the 100m sub-section data. In general, the observations applied equally to the 10m data, in terms of the change in the shape of the distribution with change in class. When the 10m and 100m distributions are compared on the same class they are visually similar (compare the middle and lower plots in Figure 5.1 and Figure 5.2). There are differences of a few percent between the proportions of the network reported in each category when reported over 10m and over 100m lengths. However, there a few instances where the differences are larger (around 10%) – for example the increase in the percentage in category 2 reported by participant M on B/C roads when the data is reported over 10m lengths. It is possible that larger changes reflect greater levels of variability in condition along the length of the road, to which some devices are more sensitive than others.

On U roads (Figure 5.4) there were much greater differences between the participants' data and the benchmark, and between the participants themselves. However, it is noted that on these roads the distribution of engineers' category data is "unusual", as it reports increasing proportions of the network to fall in categories 2, 3 and 4.

The comparison measures discussed in Section 4.4 were applied to provide a quantitative indication of the consistency between the distributions. The rather broad measure – the percentage in category 4/5 - suggests that the participants datasets somewhat agree with the benchmark (Figure 5.3) in terms of the proportion of the network that is "poor". All except P fall within 5% (absolute) of the benchmark on A roads, and all except O and P achieve this on B/C roads. It was found that the updated RCM data provided a greater level of agreement with the benchmark across nearly all participants for this measure. However, the updated RCM data did not overcome the large differences observed on U roads – as can be seen in Figure 5.5. Although not shown here, broadly similar behaviour was seen when the 10m distributions were assessed. This aligns with current national reporting for U roads, where the data shows greater variability than for classified roads. This suggests further work may be needed to improve the comparability of RCM data for U roads.

For the Chi-square test, higher values of the test statistic indicate lower levels of agreement between the benchmark and the participants' distribution. Table 5.1 shows that quite high values of the statistic were obtained for many of the comparisons between the RCM and benchmark distributions. The test statistic falls below the critical value[15] (i.e. shows high level of agreement between the benchmark and participant) for around half the participants on A roads, less than half on B/C roads and less than a fifth on U roads (these are highlighted green). The same systems do not all perform "well" on each class (only system H). For the Pearson correlation test higher values indicate higher levels of agreement between the benchmark and the participants' distribution. High levels (>0.8) of the correlation test value are obtained for most systems on A and B/C roads (highlighted green), but by few systems on U roads. This test does not seem to be as strong/sensitive an indicator of the visual agreement between the distributions as the Chi-square test.

Whilst the above comments relate to the 100m datasets, it was found that applying either the Chi-square or the Pearson test to the 10m datasets resulted in only a slightly reduced performance. However, the Chi-square test was more sensitive to the effects of changing from 100m to 10m than the Pearson test.

The tables in Appendix D show the values of Delta obtained for each category and each class for the updated datasets. Lower values of Delta are associated with greater agreement between the benchmark and the participant data. For A and B/C roads we see low Delta values for nearly all the participants' datasets for the for the lower (better) condition categories. In general, higher values of Delta (i.e. poorer agreement with the benchmark) are only obtained for category 4/5. However, this test is significantly influenced by the low numbers of lengths falling into category 4/5, which inevitably result in higher Delta values.

Table 5.1: Chi-square statistic and Pearson correlation values (updated 100m data).

Participant (dataset)	Chi Sq (A roads)	Chi Sq (B/C roads)	Chi Sq (U roads)	Pearson (A roads)	Pearson (B/C roads)	Pearson (U roads)
A	168.9	336.74	22.29	0.77	0.69	0.92
B	49.19	32.38	2.7	1	0.97	0.9
C	59.56	152.5	16.03	0.91	0.84	0.74
D	1.05	4.99	34.75	1	0.99	-0.69
E	172.27	85.17	19.79	0.98	0.86	0.49
F	9.57	30.43	10.98	0.99	0.93	0.84
G	13.69	7.97	37.49	0.99	0.98	-0.75
H	1.31	2.54	4.57	1	1	0.93
I	1.62	23.66	33.53	1	0.96	0.14
J	34.5	58.9	10.67	0.98	0.86	0.68
K	52.83	115.76	31.68	0.96	0.9	0.86
L	4.58	1.36	12.49	0.99	1	0.55
M	0.12	31.39	46.58	1	0.94	0.15
N	4.01	2.68	23.88	1	1	-0.11
O	38.75	289.05	62.48	0.95	0.53	-0.97
P	2400.64	732.45	16.57	0.63	0.56	0.9

[15] Estimated to be 7.8 for a confidence level of 0.05 and 3 degrees of freedom

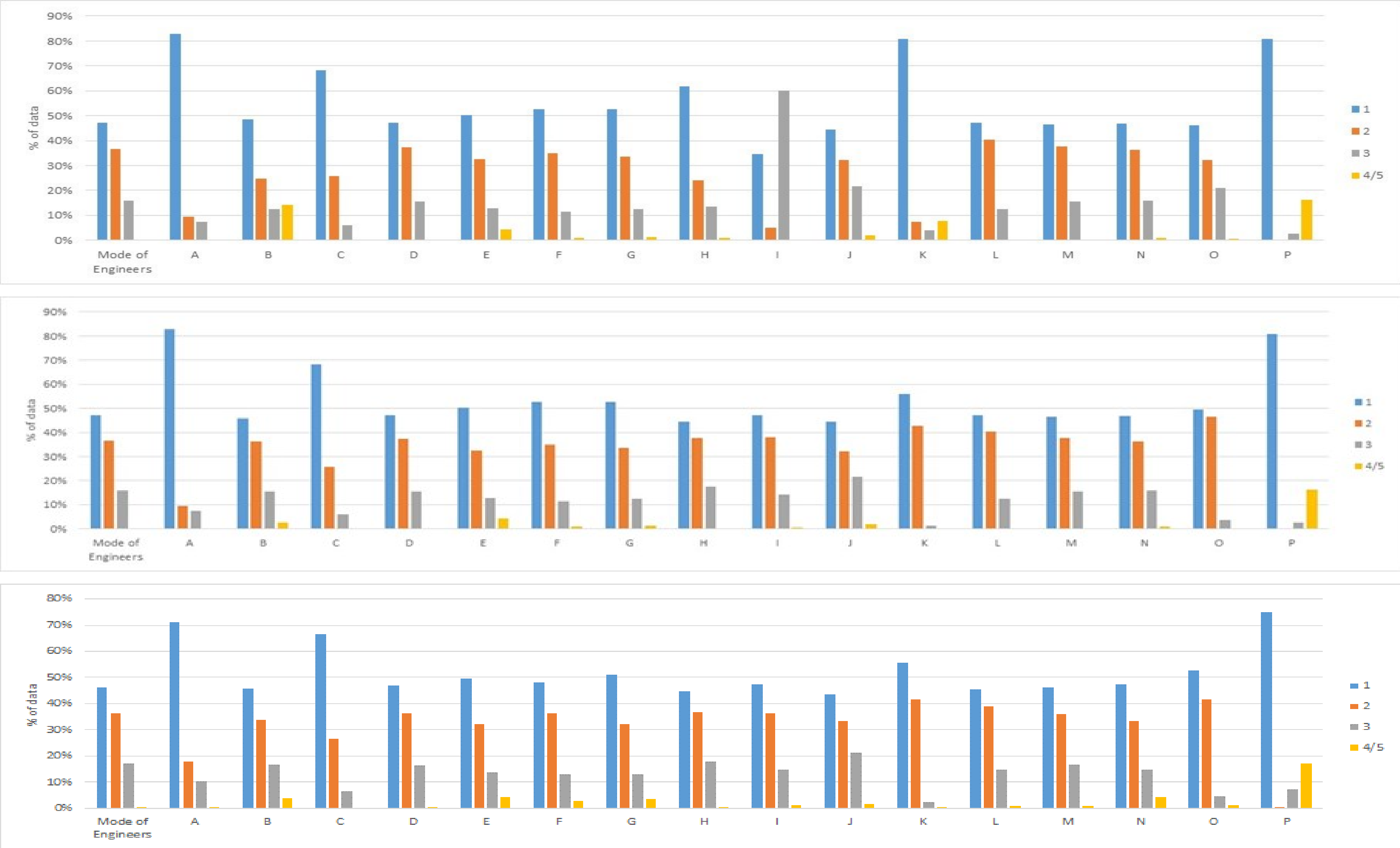


Figure 5.1: Distributions of 100m condition categories on A roads for participants A to P. Upper – 100m initial condition categories. Middle 100m updated condition categories. Lower – 10m updated condition categories.

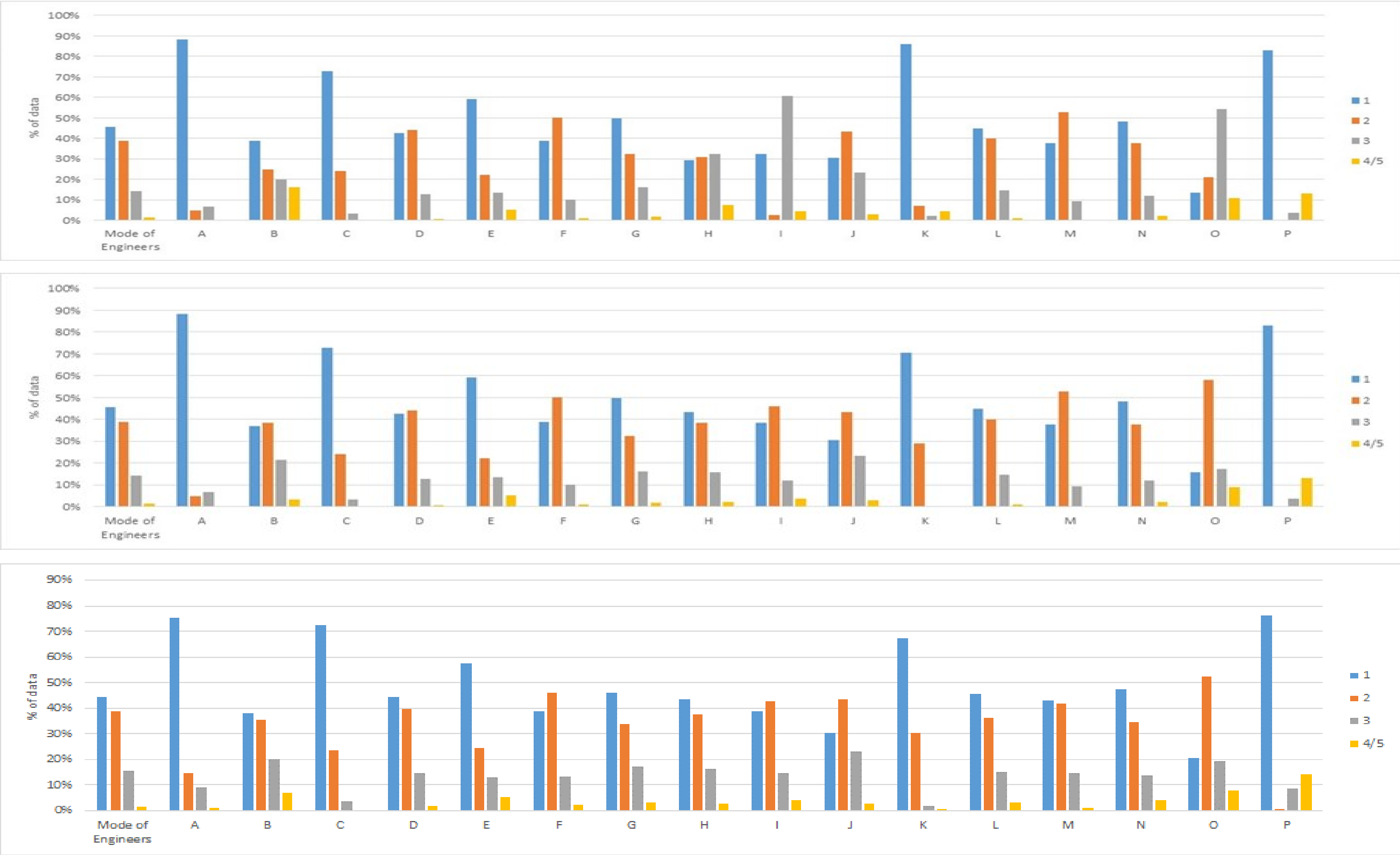
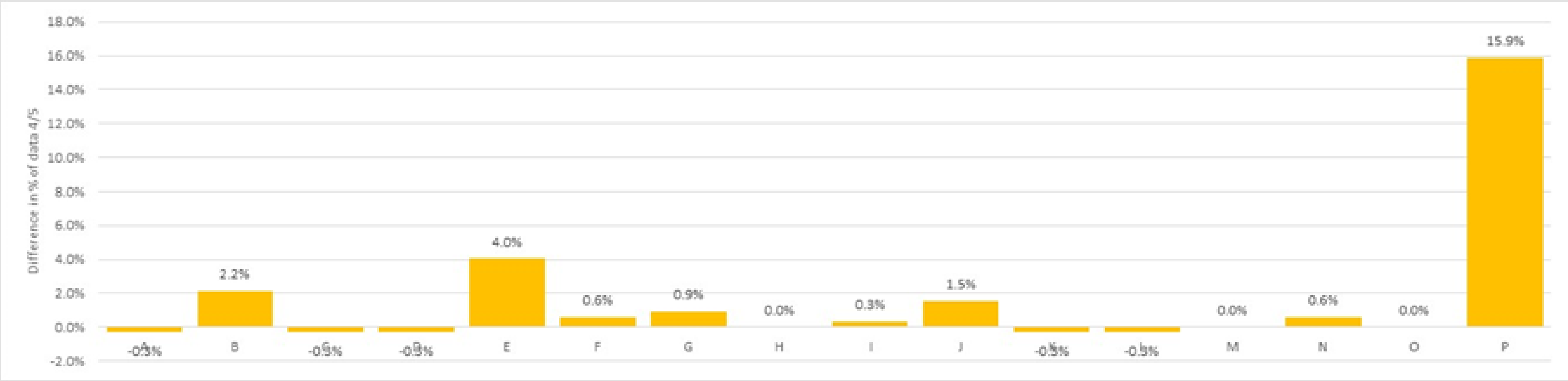


Figure 5.2: Distribution of categories on B/C roads for participants A to P. Upper – 100m initial condition categories. Middle 100m updated condition categories. Lower – 10m updated condition categories.



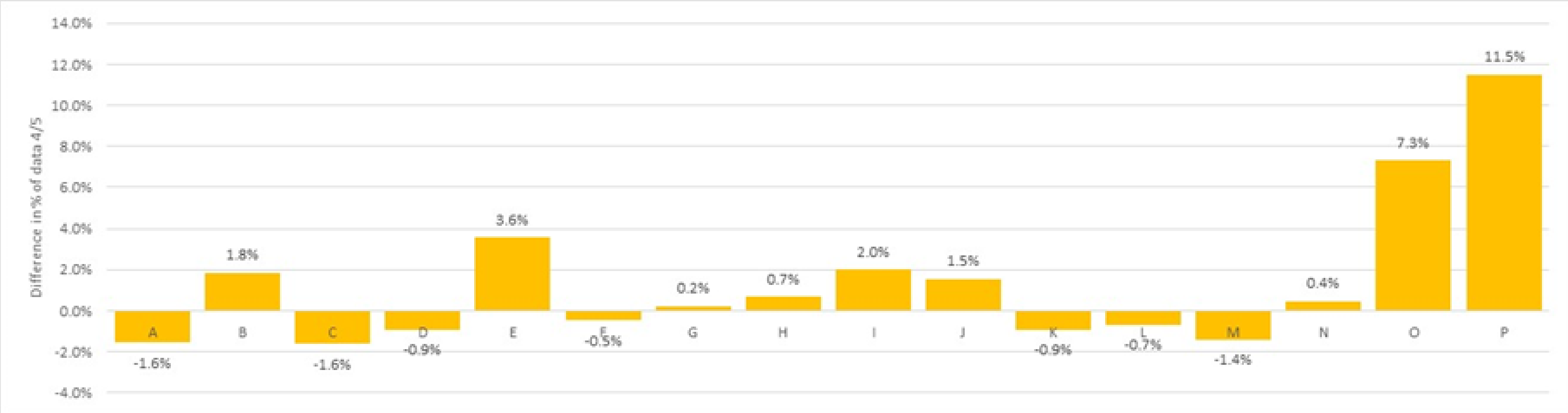


Figure 5.3: Differences between percentage reported in categories 4 and 5 by the benchmark and the participant data for participants A to P. Updated condition category data (A roads Upper, B/C Lower)

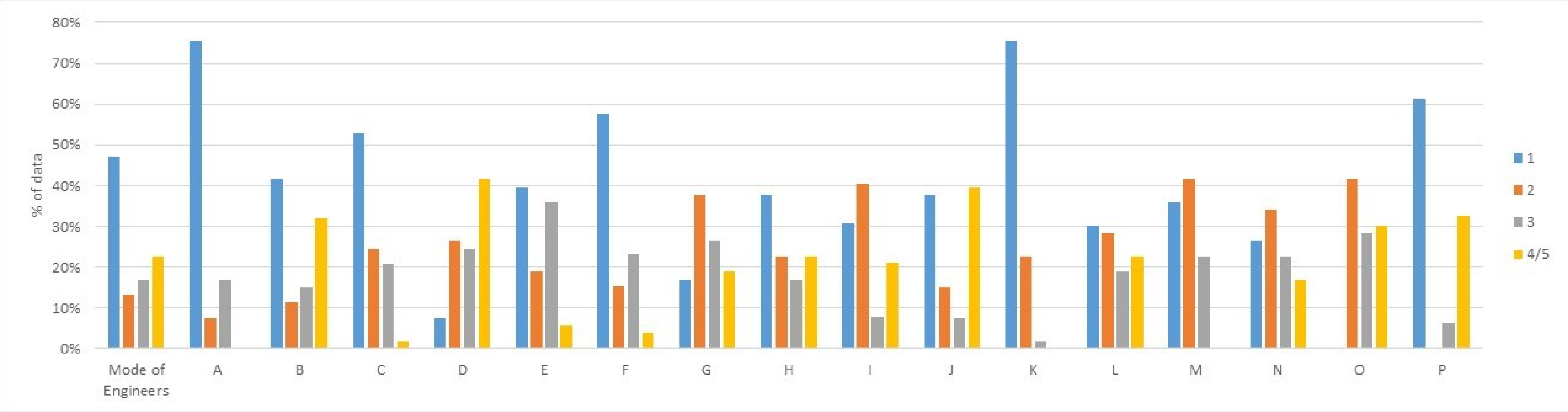


Figure 5.4: Distribution of ratings for participants A to P on U roads (100m updated data)

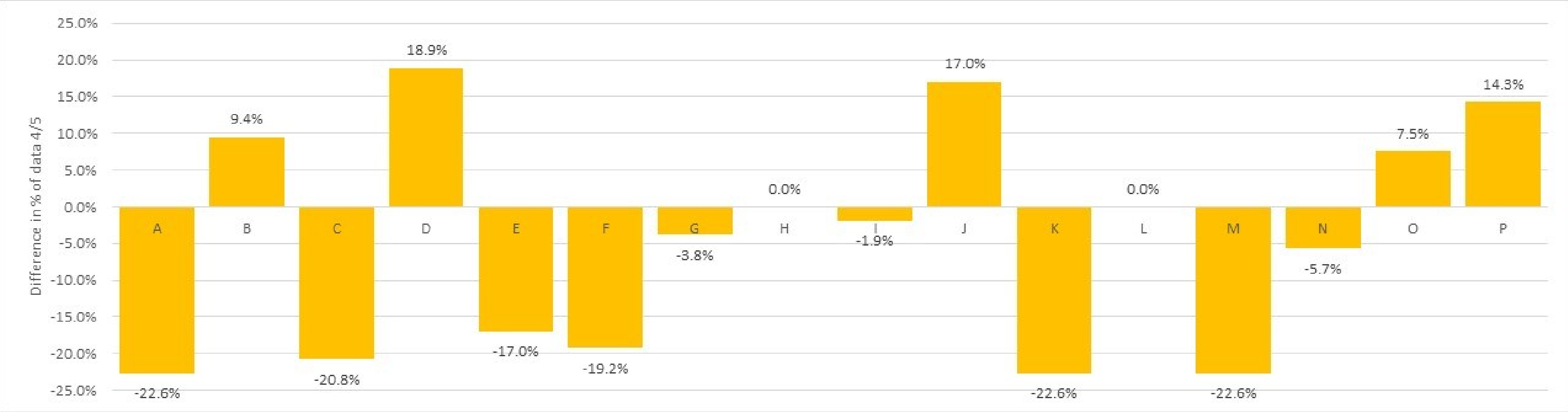


Figure 5.5: Difference between the percentage recorded by the benchmark and participant data in categories 4 and 5 for participants A to P on U roads (100m updated data)

5.3 Spatial comparisons

5.3.1 100m sub-section lengths

The spatial comparisons take a deeper dive into the reporting of condition categories on a sub-section by sub-section basis. To undertake this comparison the benchmark data was fitted to the study network as discussed in Section 4.2 so that categories could be reported in relation to section and chainage (distance). The participants deployed similar fitting processes to align their RCM data to the study network. The spatial comparison assumes that fitting delivers accurately aligned datasets. However, the fitting utilises coordinate values recorded during the data collection (e.g using GPS) to fit the categories to the network. The location measurement system is likely to have an accuracy of a few metres. Inaccuracies in the measurement of coordinates will introduce misalignment into the data that will inevitably affect the ability to directly compare the datasets at the sub-section level. This should be taken into consideration when reading this Section of the report.

At the simplest level we can visually compare the plots of condition categories reported by the benchmark and participant datasets. Figure 5.6 provides an example for the initial and updated datasets provided by participant B, along with example images recorded by the HARRIS3 survey at a few locations along the study network. The “shape” of the participant RCM condition category data broadly matches the benchmark, with the RCM data generally reporting higher values when the benchmark reports higher values. However, there are places where the RCM data reported spikes (e.g. category 5) when the benchmark did not, and locations where there are differences of one or more category (e.g. where the benchmark reports 2, RCM reports 3). These differences were reduced in the updated dataset provided by this participant. It is these visual differences that give rise to differences between the benchmark and RCM condition category distributions discussed in the previous Section. So, where the RCM shows a tendency for the orange line to be greater than the blue line in a high number of locations we will see larger bars in the distribution than the benchmark for the higher categories. For this example of participant B the upper bar chart in Figure 5.1 shows a tendency for the device to report more 4/5s and less 2s on A roads. A visual inspection of Figure 5.6 suggests that this is what you might expect.

The plots comparing the condition categories reported by the benchmark and each of the other participants are shown in Appendix D, each showing the various characteristics of that participants’ data to report higher (e.g. participant P) or lower (e.g. participant A) than the benchmark at various locations on the network.

A quantitative assessment of the spatial comparison can be achieved through the confusion matrices described in section 4.4.2. These are presented for all participants in Appendix D. Although the confusion matrices are not always intuitive to interpret, they reflect the comparisons that can be seen visually in the graphs of the condition category data. For example, Table 5.2 shows the confusion matrix for participant B on A roads for both datasets. The diagonal shows the percentage of sub-sections where both the benchmark and the reference reported the same category. The cell highlighted blue shows that there was quite a high percentage of lengths in the initial data where the RCM device reported category 4/5 when the benchmark reported category 3. In agreement with this, examination of the regions of Figure 5.6 (upper plot) shaded red (the A roads) shows a number of peaks where the RCM data reported higher categories in the A road sections. However, these differences were reduced in the updated dataset, and this is reflected by the greater agreement both along the diagonal and in the blue cell.

Table 5.2: Confusion matrix for participant B, initial data (left) and updated data (right)

Initial data:

Data	Precision 0	Precision 1
B	44.1%	84.9%

		RCM			
		1	2	3	4/5
Bench.	1	67	24	5	4
	2	39	24	18	18
	3	12	29	22	37
	4/5	100	0	0	0

Updated data:

Data	Precision 0	Precision 1
B	52.5%	94.8%

		RCM			
		1	2	3	4/5
Bench.	1	59	30	10	1
	2	26	49	22	4
	3	3	40	48	8
	4/5	0	0	86	14

Examining the confusion matrices for the updated datasets (Appendix D), suggests that:

- For A roads, the confusion matrices for the majority of the participants have high values on the diagonal (i.e. there is spatial agreement between the categories reported by the benchmark and the RCD data) for the better condition categories (1,2). However, the agreement decreases for category 3 for most participants and there is no agreement for any participants for category 4/5. It must be noted that this primarily a result of the low number of lengths reported in the benchmark as category 4/5 on A roads.
- For B/C roads broadly the same behaviour is seen as on A roads, but the agreement on the diagonal is slightly reduced. However, on B/C roads nearly all participants report a proportion of the sub-sections in category 4/5 that were reported by the benchmark (although this is less than 20% for the majority of participants). This reflects the increasing number of poor condition sub-sections on B/C roads.
- The spatial agreement (the same lengths reported as the same condition by the RCM device and the benchmark) on U roads is less consistent across nearly all categories for nearly all participants. Again, this will be affected by the lower number of U road lengths in general present on the study network.

These qualitative observations can be quantified, by device, using the precision values, as shown in Table 5.3. Here, cells are shown as red (value <50%) yellow (value>=50%) and green (value >=75%) to highlight the performance achieved. The table suggests that the vast majority of lengths are reported within one condition category of the benchmark on A and B/C roads, but there is poorer agreement on U roads. The number of lengths where there is exact agreement (precision 0) between the benchmark and the RCM devices is highest on A roads. For the participant fleet the data suggest that, when comparing RCM condition categories reported over 100m sub-section lengths benchmark condition categories, you are likely to obtain at least 50% of lengths in the same condition category as the benchmark, and over 90% within 1 category, on A roads - Figure 5.7. This reduces for lower classes of road (for this dataset).

Figure 5.6: Plot of individual condition categories reported, relative to section and chainage, by participant B (Upper, initial data. Lower, updated data), and examples of images collected by the HARRIS3 survey at three locations on the study network. The colour coding shows the class (A, B C, U as blue, orange, yellow and grey shading)

Table 5.3: Precision values for each participant by road classification (updated data)

Participant	A Roads	A Roads	B/C Roads	B/C Roads	U Roads	U Roads
	Precision 0	Precision 1	Precision 0	Precision 1	Precision 0	Precision 1
A	53.3%	90.4%	46.4%	86.7%	52.8%	75.5%
B	52.5%	94.8%	45.8%	81.4%	43.4%	79.2%
C	53.3%	92.9%	55.0%	94.0%	58.5%	86.8%
D	51.5%	92.6%	49.9%	92.2%	24.5%	75.5%
E	49.8%	91.3%	48.1%	87.1%	52.8%	81.1%
F	65.6%	97.2%	61.2%	96.3%	59.6%	90.4%
G	50.0%	90.1%	48.9%	90.5%	20.8%	71.7%
H	79.0%	94.8%	37.6%	82.5%	66.0%	86.8%
I	58.3%	93.5%	39.4%	80.3%	50.0%	88.5%
J	50.3%	88.0%	48.9%	90.0%	49.1%	66.0%
K	60.7%	96.9%	45.8%	88.8%	52.8%	66.0%
L	45.7%	90.4%	45.7%	90.3%	43.4%	71.7%
M	40.1%	87.3%	47.9%	91.0%	43.4%	77.4%
N	54.6%	94.1%	51.1%	92.3%	30.2%	79.2%
O	45.1%	92.3%	24.1%	67.0%	17.0%	62.3%
P	44.6%	74.7%	41.2%	77.2%	42.9%	57.1%

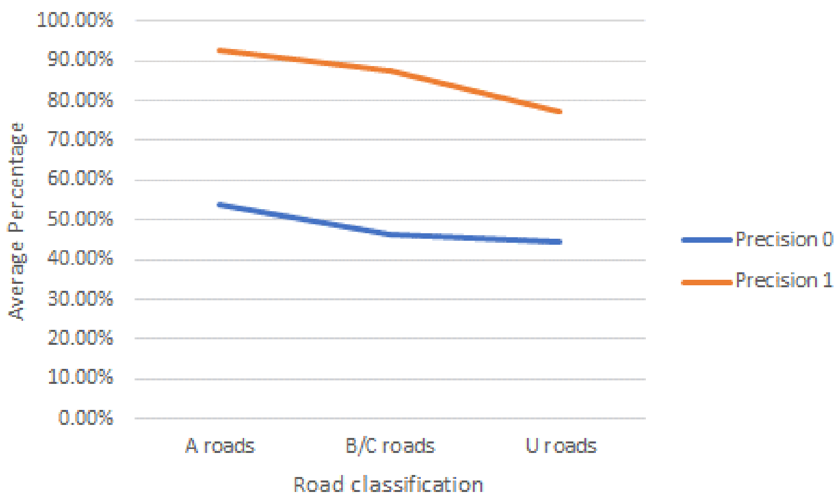


Figure 5.7: Average precision values obtained across the participants for each class of road

5.3.2 *Implications of using shorter (10m) sub-section lengths*

For the spatial comparisons the condition category values reported in relation to section and chainage along the length of the study network are compared individually. When reporting over 100m sub-section lengths (Section 5.3.1) the influence of locational errors of a few m (the level of error expected for the participants' data) is reduced. However, such locational misalignments are anticipated to have greater influence when 10m sub-section lengths are used. The influence of this was questioned by the network study stakeholder group when discussing the choice of reporting length to apply when undertaking performance assessments of RCM devices.

Table 5.4 compares the precision values obtained from confusion matrices produced for classified roads using 10m sub-section lengths and 100m sub-section lengths. The use of shorter reporting lengths does reduce the measured performance, but only (on average) by 2-3% (absolute). The relatively small overall difference in precision values when using shorter sub-section lengths is lower than anticipated. It is possible that the behaviour of the data on the network affects the extent of the influence. For example, a contiguous length of consistent condition categories (e.g. lengths of adjacent subsections all reported as 1s, of which a few examples can be seen in Figure 5.6), may be less affected by locational misalignment if the RCM device also reports the same category continuously.

Table 5.4: Comparison of precision values obtained when using 10m and 100m lengths to obtain the confusion matrices - on classified roads

Participant	10m Precision 0	10 m Precision 1	100m Precision 0	100m Precision 1	Difference (100m-10m) Precision 0	Difference (100m-10m) Precision 1
A	46.5%	87.1%	49.3%	88.2%	2.80%	1.10%
B	47.6%	88.8%	52.4%	93.1%	4.80%	4.30%
C	52.9%	93.0%	54.2%	93.5%	1.30%	0.50%
D	47.6%	90.0%	50.6%	92.4%	3.00%	2.40%
E	48.1%	89.0%	48.8%	88.8%	0.70%	-0.20%
F	58.7%	96.0%	63.2%	96.7%	4.50%	0.70%
G	45.6%	88.2%	49.4%	90.3%	3.80%	2.10%
H	77.4%	94.5%	81.9%	96.9%	4.50%	2.40%
I	56.0%	92.6%	60.4%	93.8%	4.40%	1.20%
J	49.0%	89.6%	49.5%	89.2%	0.50%	-0.40%
K	51.5%	93.7%	54.7%	95.8%	3.20%	2.10%
L	42.8%	88.0%	45.7%	90.4%	2.90%	2.40%
M	41.2%	86.6%	44.5%	89.4%	3.30%	2.80%
N	48.0%	89.9%	52.6%	93.0%	4.60%	3.10%
O	39.4%	85.7%	39.2%	88.8%	-0.20%	3.10%
P	40.3%	73.6%	42.5%	76.2%	2.20%	2.60%
Average	**49.54%**	**89.14%**	**52.43%**	**91.03%**	**2.89%**	**1.89%**

Note that, as a result of the averaging across all categories, the value reported for the overall precision may not reflect differences seen between individual cells (in either the 100m and 10m confusion matrices). For example, a participant might have a relatively high precision score if their data frequently agreed with the benchmark for categories 1 and 2 (usually the most commonly-seen categories), but this could obscure infrequent agreement achieved for categories 3/4/5.

The 100m and 10m matrices for participants F and L are shown in Table 5.5 (A roads). For participant F the difference between the 100m and 10m precision 0 value is 6.9% (65.6%-58.7%) and for precision 1 is 1.2%. However, a cell for cell comparison of the 100m and 10m values on the diagonal (precision 0) and off diagonal (precision 1) shows that there are individual differences greater than 1.2% on the diagonal and greater than 6.9% off the diagonal. However, in general the figures within the 10m and 100m matrices are relatively consistent with each other (i.e. 10m 'F' vs. 100m 'F') for these example participants.

Table 5.5: Example of effect of 100m (top) and 10m subsection lengths on the confusion matrices for participants F and L (updated data)

Data	Precision 0	Precision 1
F	65.6%	97.2%

		RCM			
		1	2	3	4/5
Bench.	1	83	16	1	0
	2	30	55	14	0
	3	14	42	38	6
	4/5	0	100	0	0

Data	Precision 0	Precision 1
L	45.7%	90.4%

		RCM			
		1	2	3	4/5
Bench.	1	56	38	6	0
	2	39	44	18	0
	3	41	39	20	0
	4/5	0	100	0	0

Data	Precision 0	Precision 1
F	58.7%	96.0%

		RCM			
		1	2	3	4/5
Bench.	1	77	21	2	0
	2	28	52	17	2
	3	10	46	34	10
	4/5	0	44	44	11

Data	Precision 0	Precision 1
L	42.8%	88.0%

		RCM			
		1	2	3	4/5
Bench.	1	54	36	9	1
	2	39	41	19	1
	3	38	42	21	1
	4/5	0	56	44	0

5.4 A note on the application of machine learning for calibration

Participant H generated their updated data using a machine learning algorithm to train their system. It can be seen from the above, and Appendix D, that the updated data from participant H was the most consistent with the benchmark. Participant H provided datasets in which they had trained their machine learning algorithm using 85, 50% and 8% of the Benchmark data. In the above we have used the 50% dataset. However, Table 5.6 compares the performance achieved using the 50% and 8% datasets. The performance achieved by the 8% dataset is comparable with the rest of the participants.

The performance achieved using the 50% dataset suggests that machine learning has potential to deliver high levels of performance. However, in this case this this was achieved using a substantial proportion of the benchmark dataset. This presents the risk that the performance is representative of what can be achieved on the specific study network, and not the local authority network as a whole. To ensure the trained system is able to perform to this level on the whole network would require the benchmark network to be fully representative of the whole network, with sufficient data withheld from the RCM provider, to ensure that a robust "blind" assessment can be carried out.

Table 5.6: Confusion matrices (percentages of reference) for participant H for datasets obtained using machine learning tools trained using 50% and 8% of the benchmark data.

Data	Precision 0	Precision 1
H: ML 50%	79.0%	94.8%

		RCM			
		1	2	3	4/5
Bench.	1	81	13	6	0
	2	12	79	8	1
	3	12	14	75	0
	4/5	0	100	0	0

Data	Precision 0	Precision 1
H: ML 8%	51.5%	89.2%

		RCM			
		1	2	3	4/5
Bench.	1	59	29	12	0
	2	34	49	17	0
	3	29	33	37	0
	4/5	0	100	0	0

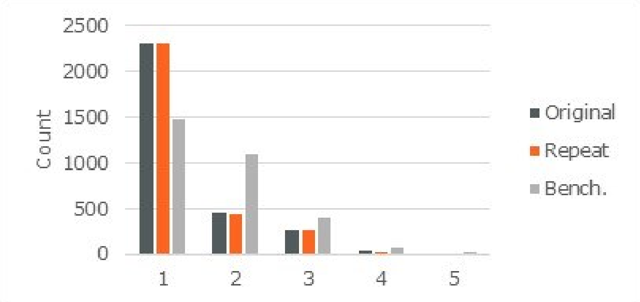
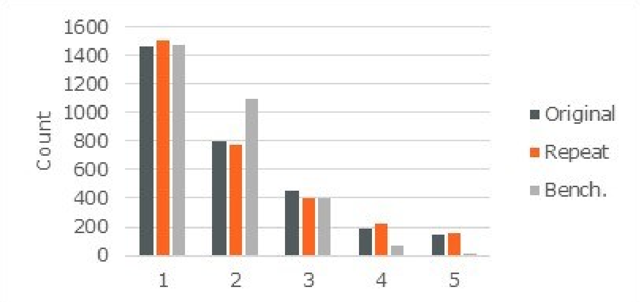
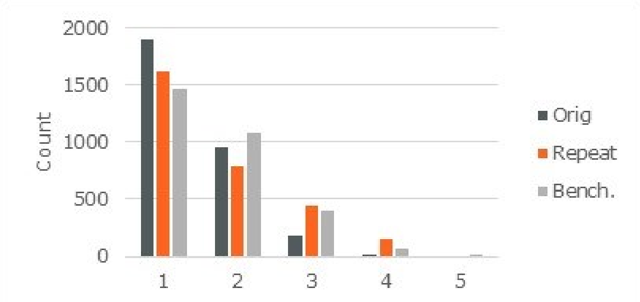
5.5 Repeatability

Repeat condition category data was provided by several of the participants from the repeat loops (Section 3.2). The RCM categories reported in the repeat data were compared with the categories reported in the original dataset provided by that participant, and with the benchmark condition categories. These are shown in Table 5.7. In this table repeatability refers to the comparison between RCM "runs", and reproducibility refers to the comparison of the RCM to the benchmark. Note: the reproducibility values differ from the comparisons with the benchmark shown above because this assessment has been carried out using only the subset of the surveyed network where repeat runs were carried out (repeat loops).

Table 5.7 shows that the repeatability of the RCM technologies is typically high, and better than the reproducibility. This can be seen visually in the histograms, and also in the values in the confusion matrices and the Chi-square values. However, there is one exception to this (dataset C), which shows a higher variation for the repeatability. As with the comparisons with the benchmark there are lower levels of performance in the repeatability in the higher categories (due to the low number of sub-sections having high condition category values).

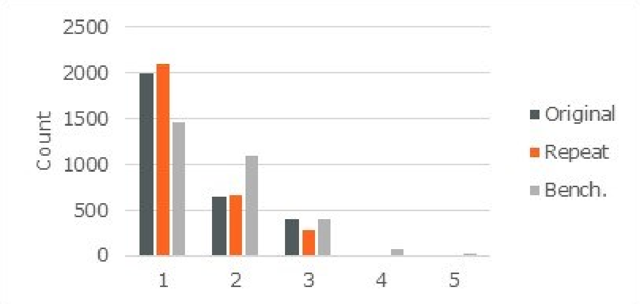
It can also be seen from Table 5.7 that the participants that show high repeatability in the distributions also show repeatability confusion matrices with higher levels of performance than the corresponding reproducibility confusion matrices. However, whilst the repeatability performance is higher, it can be seen that the values on the diagonal are only around 40-50% for categories other than 1. This demonstrates the challenge of achieving high performance values in the confusion matrix assessment. The locational issues affecting the comparison with the benchmark will also be present in the repeatability test (and hence affect the performance). In addition, the reducing percentage of lengths that fall into the higher (poorer) condition categories make it increasingly challenging to match condition categories on a sub-section for sub-section basis (i.e. spatially).

Table 5.7: Repeatability, for the devices providing repeat data

<table>
<tr><th>Data</th><th>Histogram</th><th>Repeatability confusion matrix</th><th>Reproducibility confusion matrix</th><th>Repeat-ability Chi Sq</th><th>Reprod-ucibility Chi Sq</th></tr>
<tr><td>A</td><td>
</td><td>
<table>
<tr><td></td><td></td><td colspan="4">Repeat</td></tr>
<tr><td></td><td></td><td>1</td><td>2</td><td>3</td><td>4/5</td></tr>
<tr><td rowspan="4">Original</td><td>1</td><td>83</td><td>12</td><td>5</td><td>1</td></tr>
<tr><td>2</td><td>60</td><td>23</td><td>16</td><td>1</td></tr>
<tr><td>3</td><td>45</td><td>24</td><td>28</td><td>3</td></tr>
<tr><td>4/5</td><td>48</td><td>16</td><td>26</td><td>10</td></tr>
</table></td><td>
<table>
<tr><td></td><td></td><td colspan="4">Original</td></tr>
<tr><td></td><td></td><td>1</td><td>2</td><td>3</td><td>4/5</td></tr>
<tr><td rowspan="4">Bench.</td><td>1</td><td>84</td><td>11</td><td>5</td><td>0</td></tr>
<tr><td>2</td><td>72</td><td>16</td><td>11</td><td>1</td></tr>
<tr><td>3</td><td>56</td><td>27</td><td>16</td><td>1</td></tr>
<tr><td>4/5</td><td>69</td><td>13</td><td>11</td><td>7</td></tr>
</table></td><td>0.88</td><td>913.84</td></tr>
<tr><td>B</td><td>
</td><td>
<table>
<tr><td></td><td></td><td colspan="4">Repeat</td></tr>
<tr><td></td><td></td><td>1</td><td>2</td><td>3</td><td>4/5</td></tr>
<tr><td rowspan="4">Original</td><td>1</td><td>81</td><td>15</td><td>3</td><td>1</td></tr>
<tr><td>2</td><td>33</td><td>49</td><td>13</td><td>4</td></tr>
<tr><td>3</td><td>10</td><td>30</td><td>41</td><td>16</td></tr>
<tr><td>4/5</td><td>2</td><td>9</td><td>18</td><td>32</td></tr>
</table></td><td>
<table>
<tr><td></td><td></td><td colspan="4">Original</td></tr>
<tr><td></td><td></td><td>1</td><td>2</td><td>3</td><td>4/5</td></tr>
<tr><td rowspan="4">Bench.</td><td>1</td><td>63</td><td>25</td><td>8</td><td>3</td></tr>
<tr><td>2</td><td>39</td><td>27</td><td>19</td><td>15</td></tr>
<tr><td>3</td><td>23</td><td>27</td><td>28</td><td>22</td></tr>
<tr><td>4/5</td><td>14</td><td>14</td><td>18</td><td>54</td></tr>
</table></td><td>14.43</td><td>861.81</td></tr>
<tr><td>C</td><td>
</td><td>
<table>
<tr><td></td><td></td><td colspan="4">Repeat</td></tr>
<tr><td></td><td></td><td>1</td><td>2</td><td>3</td><td>4/5</td></tr>
<tr><td rowspan="4">Original</td><td>1</td><td>67</td><td>24</td><td>7</td><td>2</td></tr>
<tr><td>2</td><td>37</td><td>31</td><td>25</td><td>7</td></tr>
<tr><td>3</td><td>6</td><td>19</td><td>40</td><td>35</td></tr>
<tr><td>4/5</td><td>0</td><td>50</td><td>0</td><td>50</td></tr>
</table></td><td>
<table>
<tr><td></td><td></td><td colspan="4">Original</td></tr>
<tr><td></td><td></td><td>1</td><td>2</td><td>3</td><td>4/5</td></tr>
<tr><td rowspan="4">Bench.</td><td>1</td><td>85</td><td>15</td><td>0</td><td>0</td></tr>
<tr><td>2</td><td>56</td><td>34</td><td>10</td><td>0</td></tr>
<tr><td>3</td><td>11</td><td>79</td><td>10</td><td>0</td></tr>
<tr><td>4/5</td><td>4</td><td>62</td><td>29</td><td>5</td></tr>
</table></td><td>6423.94</td><td>337.64</td></tr>
</table>

Data	Histogram	Repeatability confusion matrix	Reproducibility confusion matrix	Repeat-ability Chi Sq	Reprod-ucibility Chi Sq

E



Repeatability confusion matrix (Original × Repeat):

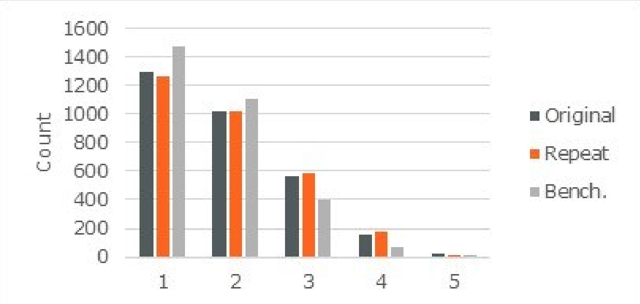
Original	Repeat 1	Repeat 2	Repeat 3	Repeat 4/5
1	89	10	1	0
2	41	50	9	0
3	13	38	50	0
4/5	-	-	-	-

Reproducibility confusion matrix (Bench. × Original):

Bench.	Original 1	Original 2	Original 3	Original 4/5
1	76	15	9	0
2	58	25	17	0
3	49	32	19	0
4/5	58	27	14	0

Repeatability Chi Sq: 39.06
Reproducibility Chi Sq: 453.82

G



Repeatability confusion matrix (Original × Repeat):

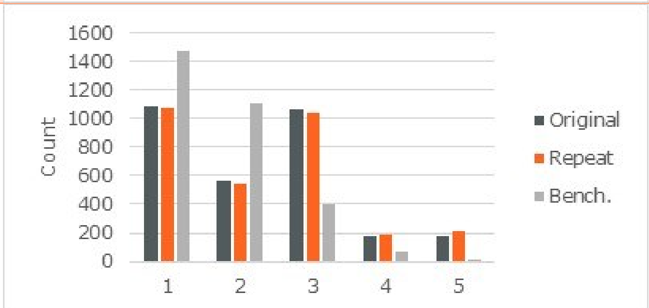
Original	Repeat 1	Repeat 2	Repeat 3	Repeat 4/5
1	80	18	2	0
2	21	60	17	1
3	2	29	57	12
4/5	0	3	32	55

Reproducibility confusion matrix (Bench. × Original):

Bench.	Original 1	Original 2	Original 3	Original 4/5
1	57	29	11	2
2	31	37	26	6
3	29	43	20	9
4/5	4	9	45	42

Repeatability Chi Sq: 4.10
Reproducibility Chi Sq: 197.38

O



Repeatability confusion matrix (Original × Repeat):

Original	Repeat 1	Repeat 2	Repeat 3	Repeat 4/5
1	74	16	7	2
2	30	35	32	1
3	9	16	62	8
4/5	4	4	33	24

Reproducibility confusion matrix (Bench. × Original):

Bench.	Original 1	Original 2	Original 3	Original 4/5
1	34	20	36	9
2	39	18	32	11
3	35	16	37	12
4/5	5	1	26	68

Repeatability Chi Sq: 5.85
Reproducibility Chi Sq: 2360.65

6 Discussion

The introduction of new Road Condition Monitoring technologies over recent years has provided an opportunity to reconsider the approach taken to the collection and reporting of condition data on the local road network. This could lead to a transition from the current network wide SCANNER survey to an approach that is less restrictive in terms of the technologies deployed. The intention is to introduce this new regime via a new Publicly Available Standard (PAS), which will define future national reporting requirements. However, if national reporting of condition is to be achieved in this new regime there will be a need to address a number of the technical challenges associated with obtaining common data, achieving consistency and ensuring the quality of the data provided. This network study has sought to understand these challenges and identify potential options to address them. The following sections discuss some of the observations of the study and the implications for the PAS.

6.1 RCM devices and data

The study engaged with the industry to identify the types of technology currently available for collection and reporting of condition data, and how these may be applied for national reporting. This has identified five groups of RCM technology, which have been categorised and labelled using a lettering system based on the name of the technology group, e.g. E for Engineer's inspection. This deliberately removes the implication of "hierarchy" (as would be suggested through the use of a contiguous scale (A-E)) and makes it easier for end users to identify which letter corresponds to which group. To account for potential future technologies a sixth group (Other Technology) has been added. It is recommended that this be used in the PAS. However, should further specific groups of technologies be identified the list should be updated.

6.2 Condition categories for national reporting

In the current regime for the collection of data for national reporting (SCANNER) a specific set of defects are defined, which must be measured in a defined way to obtain a dataset of parameters. Centrally defined rules are applied to these parameters to determine the Road Condition Indicator (RCI) for national reporting. Current standardisation processes focus on consistency in the measurement of these parameters, with the resulting consistency in the RCI implied from this. For the new technologies, providers of RCM data have undertaken their own development programmes to deliver vendor-specific defects that should be of practical value to Highway authorities seeking to understand the condition of their networks (which we have referred to as additional data). The RCM technologies provide a wide range of additional data covering different aspects of condition. These data are collected and reported in different ways.

The network study has developed an end-result approach to categorising condition. In this approach the expected outcome is defined, not the method. RCM data providers themselves determine the methodology used to collect data and determine the condition category. The PAS steering group set the network study group the objective to define five condition categories. These have been established based on the actions required to restore the length of pavement to a condition that does not require maintenance. The study has avoided the need to develop an approach based on defects, rules and parameters. Hence the approach taken to calculating the condition category is determined by the RCM provider. The objective is then set for RCM providers to demonstrate their capability in the reporting of these condition categories.

The logic/practicality of this approach is somewhat supported by the observations of Section 3.5 (additional data), in which it was clear that the condition parameters provided by current RCM devices are collected and reported in different ways, leading to different types, rating and severities of parameters. Achieving a common comparable rating of condition using rules applied to these different parameters could present a significant development challenge, and could lead to a need to standardise some parts of the collection process, which is a requirement that is not desirable for the new PAS.

6.3 The study network

When identifying the study network the aim was to establish a network that would be practical to survey whilst containing a range of road condition categories. However, the percentage of the network falling into categories 4 and 5 was lower than expected, and it was not practical within the timescale of the work to expand the network. This has affected the analysis. For a robust assessment of both distributions and spatial agreement, there is a need for a significant percentage of the network to fall into each category, and in particular that sufficient lengths fall within categories 4 and 5 (ideally several km in each). However, the majority of road networks lengths tend to fall into the better condition categories. In future assessments of RCM devices it may be necessary to undertake initial surveys of a much larger initial network to identify suitable lengths for inclusion in the benchmarking network.

During the study the steering group observed that the characteristics of different types of road can also vary across local authorities, particularly for U roads. For example, differences between the west country and eastern counties. It may therefore also be necessary to include a range of roads from across the country to ensure that it is representative of the overall network.

6.4 Benchmarking

As discussed in Section 3.3, a range of options were considered for obtaining benchmark data for the network study. These included (a) using the average of the RCM providers themselves, (b) using engineers, or (c) a hybrid of these. It was decided that engineers' surveys would be undertaken, and to consider the other options during the analysis phase. One of the key questions raised for options a and c was how to decide which RCM datasets to include within the benchmark, because the participants may not show consensus. Examination of the initial dataset showed that this is a significant risk. Although there was some consensus, the initial datasets showed large differences between the distributions of condition categories reported by each device, such that no acceptable objective approach could be determined to justify the inclusion of any specific RCM dataset within the benchmark. Clearly, consensus improved with the updated datasets, but this built on the engineers' benchmark. Whilst we are in the early stages of transition to the new RCM regime it seems appropriate to restrict benchmarking datasets to engineers' ratings.

Engineers' inspections were undertaken as driven surveys, as walked surveys would have been impractical within the timetable, and traffic management expensive. However, the level of detail achievable in driven surveys is low. This has two main impacts, the first is that the engineers' surveys will tend to change rating less frequently. This is one of the reasons we have focussed more closely on 100m subsection data in this report. The second impact is that there is a risk that the severity of defects might have been harder to establish in a consistent manner. However, this is a reason for using the consensus of several engineers to establish the rating. Given the above recommendation for longer, widely distributed test networks for future testing of RCMs (e.g. to demonstrate capability for national reporting) It is likely that full walked engineer's surveys would be impractical for future exercises. The risks to quality of the benchmark data may be alleviated through the use of a wider range of engineers. Also, as noted in Section 3.3.3, the study found that the guidance given to engineers, and the attention paid to "training", had an effect on consensus. Therefore, the training component also forms an important part of the benchmarking regime.

6.5 Understanding the performance of RCM devices

6.5.1 Distributions

Acknowledgement should be given to the RCM providers, who undertook their own development to determine condition category values from the data they collected. Each provider developed their own methodology. It was not the purpose of the study to investigate how these had been determined, but to investigate how they compared with the benchmark. The initial observations show that RCM devices are able to report condition categories that broadly relate to the consensus of the engineers who have assessed the network. There is visual agreement in the condition vs distance plots, and this shows high potential for the use of the data in road asset management.

However, the objective is to establish a process such that RCM data will be reported to a level of consistency that will enable robust reporting of condition at the national level. Our first level of comparison was the distributions. This is because condition national reporting is currently at the distribution level (e.g. percent red, amber, green). The results are promising, especially following the calibration / updated data phase. On both A and B/C roads the devices reported broadly comparable proportions of the network to be in poorer condition categories. There is currently no benchmark for the level of consistency we might expect in national reporting for these types of device. However, the SCANNER survey has been applying the RCI to report the "percentage red" for several years. We take as a rough benchmark the consistency that SCANNER has achieved in year-on-year reporting of condition on the road network, as reported in The Road Condition in England (RCE) annual report[16]. Examination of the RCE report suggests that, between 2009 and 2021 the year-on-year change in the RCI percentage red was within ±2% and ±3% (absolute) for almost 90% of authorities on the A and B/C road networks respectively. This suggests that, for the new RCM regime to achieve a comparable level of consistency the cross-fleet agreement should be at at this level for the percentage reported in category 4/5. Whilst this was not achieved by several systems in the initial data, in the updated data this level of agreement was observed for nearly all systems with only a couple of outliers (who did not provide updated data). It is emphasised that this comparison provides a broad context only. The consistency estimated from the year-on-year SCANNER data will be affected by year-on-year changes in condition. Also, because some Local Authority networks are surveyed by the same SCANNER device in subsequent years, the consistency we experienced in national reporting may include a component of repeatability in addition to reproducibility. However, the study shows some promise for the establishment of a consistent set of RCM devices on classified roads, provided that the questions over calibration, network size, and the proportion of the network that falls in the higher condition categories can be addressed.

In terms of the statistical comparisons between the device and benchmark data, these suggest that the Pearson test is not a sensitive/useful method for quantifying performance, but the Chi-square does seem to highlight the devices showing visibly higher levels of agreement. In addition, it is reassuring that the Chi-square shows good performance for the repeat datasets. However, the observation of one device with poor repeatability suggests that high levels of repeatability cannot be assumed for RCM devices.

On U roads the performance on the study network was not as high, with significant differences in the proportions of the network to be in poorer condition categories. This may reflect the more challenging nature of this network, but also the lower proportion of U roads present in the study network. The outcomes suggest that there would lower confidence in the ability to achieve consistent national reporting using RCM data collected on U roads. This aligns with current national reporting for U roads, where data shows greater variability than for classified roads. This suggests further work may be needed to improve the comparability of RCM data for U roads.

6.5.2 Spatial comparison

The spatial comparison is more challenging. To achieve high performance values the device needs to report the same locations on the network having both poor and good condition as the benchmark. In the initial datasets there were examples of where this was not achieved, with bias present resulting in some devices reporting generally lower or higher categories than the benchmark. Much of this was reduced in the updated datasets. The statistical performance figures (delta and precision values) appear quite low, at ~50% and 75% on A and B/C roads for precision 0 and precision 1 respectively. However, these are reasonable levels of performance for a first network trial. As for the network distributions above, we

[16] See: https://www.gov.uk/government/statistical-data-sets/road-condition-statistics-data-tables-rdc#condition-of-local-authority-managed-roads-rdc01. Note that this comparison assumes that no change should take place between years, which is not likely to be the case, and hence provides a broad guideline only.

have no specific reference against which to test this level of performance. However, the requirements for the measurement of cracking in the SCANNER specification provide a broad context. Volume 5 of the specification[17] defines the spatial performance for the reporting of cracking in terms of the percentage of lengths reported by the device as high (more cracking), medium or low (less cracking) that are also reported by the reference as high, medium or low (50m sub-sections). For low the target is 75%, for high it is 65%. If we consider the confusion matrices in Appendix D (taking 1 vs 1 to be the ‘low’ category and 4/5 vs 4/5 to be the ‘high’ category) we can see that the participants came very close to the performance for low. There may be scope for improvement for high. However, it is emphasised that SCANNER is used here as an example only, to provide context. The combined condition categories developed in this work are very different to the individual parameters reported by SCANNER, with specific challenges associated with the establishment of benchmark data, location referencing and technological differences between different types of RCM technology. Also, as noted in this study, SCANNER testing has shown that achieving levels of performance close to the specification requires that the network test sites contain significantly more lengths in poor condition than achieved in this study.

As for the comparison of distributions, the devices demonstrated higher levels of repeatability than reproducibility in the spatial comparisons. However, a further observation regarding repeatability is that, whilst assessing the repeatability of dedicated survey vehicles and dedicated mobile technologies seems straightforward (i.e. undertake a repeat survey), for other technology types the approach is less clear. For technology types such as Application-based mobile technologies and vehicle telematics it may require a second dataset to be collected covering a different date range (taking care to account for maintenance/deterioration between the datasets) or selecting random subsets of the data from within the original dataset.

6.5.3 Other observations

The performance achieved in the study drew on the use of calibration data for several (but not all) of the devices. For the further calibration the RCM providers had access to all of the benchmark data, but only a subset of providers used this to calibrate their devices. The question of calibration was one that was discussed throughout the study. RCM providers cannot be expected to “guess” the basis for condition category values, and hence the data may be an essential component of achieving consistency. However, as noted in 5.4 there is risk that its application leads to a dataset that is too fine tuned to the network on which the assessment is being carried out. This will require careful attention in any future studies.

The work has also examined the difference in performance when using 10m and 100m reporting lengths. This has not been conclusive. For the network distributions the differences were quite small, but not insignificant. For the spatial comparisons we expected the measurement of location (fitting to the network) to have a greater effect on the performance when reporting over shorter lengths (than longer lengths). However, although

[17] See: https://www.ciht.org.uk/media/11987/scanner_specification_volume_5_october_2009.pdf

this may be a contributing factor in the reduction in performance we have seen when reporting over 10m lengths, the reduction was not as large as expected. As 10m lengths are likely to be applied for national reporting it may be desirable to use 10m lengths in future testing of RCM devices. However, there may be benefit in applying processing tools during the analysis to ensure that the benchmark and RCM data are optimally aligned prior to comparison. In the meantime, consideration could be given to using both reporting lengths until greater experience is gained with the systems.

7 Conclusions

The introduction of new RCM technologies will enable a transition from the current network wide SCANNER survey to an approach that is less restrictive in terms of the technologies deployed and condition data reported. The new regime will introduce a Publicly Available Standard (PAS) to define future national reporting requirements using RCM technologies. The study engaged with 11 potential providers who freely provided their time to undertake a practical demonstration of the collection of RCM data on the network. The study has provided insight into the challenges that may lie ahead when introducing the new standard and implementing it across local authorities. The key conclusions from the study are:

- An approach has been proposed to classify the wide range of RCM technologies by type. This may be useful for local authorities considering data collection regimes.
- RCM devices collect and report a wide range of condition data. Achieving a common comparable rating of condition using "rules" applied to such data could be challenging, and stifle innovation. Therefore, five end-result condition categories have been established, based on the extent of action required to restore the pavement to a condition that does not require maintenance. The approach taken to calculating the condition category is determined by the RCM provider.
- In addition to categories, RCM devices can provide additional data on the defects. However, this is reported in different ways by each device. Care would need to be taken interpreting this data, and in particular when switching between providers.
- The use of engineers' assessments as benchmark data for both calibration and performance testing appears to be robust. However, any future process to approve RCM devices will need to ensure that the benchmark network is representative of the network as a whole and that engineers' benchmark data is robust. The process should ensure any calibration process does not "over-tune" the data to the network deployed for the demonstration. It should also be ensured that comparisons with the benchmark are not adversely affected by locational errors. This may be achieved through longer reporting lengths (e.g.100m), or through the application of processing tools to better align the data.
- The outcomes of the study suggest there is potential for new RCM technologies to provide consistent condition categories on the local road network. However, there are differences between systems, which could affect the achievement of comparable, consistent condition data across local authorities. The calibration stage, in which providers were able to benchmark their data to a common level, reduced these differences. However, the consistency (across all RCM devices) may not yet be at a level that is comparable with the consistency of current devices applied for national reporting of pavement condition. Until this is achieved, greater variability may be expected in cross-year and cross-authority reporting.
- Ultimately, there may be a requirement to establish performance thresholds for approval of RCM devices. The study has proposed methods that could be used to quantify the performance. However, further work would be required to establish comparability thresholds, and to undertake approval tests.

Appendix A Guidance provided for the benchmark surveys

A.1 Overview

Thank you for participating in this study, which is being carried out to support the development of a Publicly Available Standard (PAS) for national reporting of road condition.

The objective of this survey is to obtain Engineers' ratings of the condition category of lengths of a test route. It is not a defect survey.

The study will compare condition category ratings with ratings provided by Road Condition Monitoring systems, to help understand the performance of current technologies and hence define the requirements of the PAS.

The network for the study is located neat Guildford in Surrey. The total length of the study network is approximately 90km. To meet practical constraints the engineers' surveys will be carried out on a subset (but the majority) of the overall network, broken down into a set of shorter drivable routes. Therefore, on the day of the study you will be undertaking several "survey runs".

Due to the length of the surveys, we anticipate that at least 4 hours will be required at the location of the route. In addition, for those arriving/leaving from TRL in Crowthorne, there will be a further 2 hours travel time required (in total) for travelling to/from the survey site.

Two engineers will sit in the front seat of a survey vehicle equipped with suitable hi-vis markings. You will be driven around the route by a TRL driver. During the survey the speed of the vehicle will normally be ~20-30mph (and no greater than 40mph).

You will use a tablet to record condition categories of the pavement from 1 (Good) to 5 (Poor) along the length of the route, as discussed in Section A.2. The tablets are linked to GPS receivers and will record locations continuously, to link your category ratings to their locations on the route.

Note: there are hazards associated with carrying out the survey, the risk assessment should be read and signed before starting the survey.

A.2 Assessing Condition Categories in this Survey

As noted above, the objective of this survey is to obtain engineer's condition category ratings for each length of the route. This Section provides guidance on the definitions of the categories and how to apply them.

A.2.1 The Pavement Condition Categories

The pavement condition categories are based on the engineer's (i.e your) judgement of the actions required to restore that length of pavement to an 'as new' condition. This is based on your visual assessment of the deterioration and your experience of the ride quality.

Five categories are defined for the study, which are explained within the following table in terms of the category name, the general level of deterioration associated with that category and the level of action you would anticipate is required to restore the condition to 'as new'. The list of deterioration and actions is not exhaustive, and specific defects or extents are not specified. It is expected that you would apply your own judgement to rate each length within the broad definitions provided in the table.

Table 7.1: Pavement Condition Categories

	Category*	Example deterioration	Action level to restore to as new condition
1	**No deterioration**.	Pavement appears in good condition and provides a smooth running surface.	Pavement not considered for maintenance.
2	**Minor (and/or Aesthetic) deterioration**.	Aesthetic and/or minor deterioration, associated with very early signs of ageing. Pavement presents superficial defects. Occasional/minor ride quality issues.	'Light touch' maintenance. (e.g. minor patching).
3	**Moderate deterioration.**	Pavement is showing notable signs of ageing. Non-structural defects visible Evidence of minor potholes or development of potholes. Ride quality is deteriorating.	Localised intervention or mid-life preventative maintenance. (e.g. surface dressing, patching, crack sealing, patching of anti-skid surface).
4	**Moderate to severe deterioration.**	Pavement is substantially aged. Defects across a significant extent of the surface. Evidence of failed maintenance/ patches. Indications of structural failure/onset of structural failure. Major potholes, Ride quality is compromised.	Preventative, perhaps full carriageway maintenance. (e.g. resurfacing, with thin overlay/surface dressing, multiple patching, edge haunching, renewal of anti-skid surface).
5	**Severe deterioration.**	Extensive deterioration/defects. Structural failure. Serviceability may soon be compromised, or is no longer serviceable. Very poor ride quality.	Significant maintenance is required, which is likely to include full carriageway resurfacing or reconstruction.

**The category is based on a judgement of the action level typically required to restore that length of pavement to an "as new" condition, based on visual assessment of the deterioration and experience of the ride quality.*

A.2.2 Further Guidance on Making an Assessment of the Condition Category

Condition category assessments will be made using the App[18]. To make an assessment you press the button that reflects your opinion of the condition category of the pavement of the carriageway visible to you. As you are driven along the route you will record a new category on the App when (in your opinion) the condition category of the pavement changes. Therefore, you are rating the pavement condition on a continuous basis, changing the allocated category as the condition changes along the length of the network. Note:

- To avoid the risk of attempting to record the categories at an overly high level of detail, we recommend that you approach the survey from the mindset "*of planned*

[18] Further guidance / instructions were provided on the App, which are not included in this report.

maintenance and scheme identification", but do *not* attempt to include prioritisation into your rating. Hence such a survey would not expect report "rapid" changes in condition categories (i.e. over very short lengths).

- When assessing the condition category, you should rate the condition of the whole carriageway pavement. On a two lane dual carriageway this would mean the condition of both the running lanes in the direction of travel. For a two-way road this would be the condition of the lanes in the direction of travel and the reverse direction of travel.
- The condition of Roundabouts SHOULD NOT be assessed. Therefore, to make sure that the condition category is "reset" after each roundabout, you should record your opinion of the condition category of the pavement of the carriageway visible to you after leaving each roundabout.
- The condition of junctions SHOULD be assessed e.g. the centre area between traffic lights.

Your ratings should be undertaken "*out of context*". Hence you would not place a length in a poorer condition category (i.e., give it more significant maintenance) because it is in "this location" than if it was located in a "different" location. In summary, assessments **should not be influenced by**:

- *The strategic priority of the road* e.g. A main artery A road of high strategic importance showing the same deterioration as an A road of low strategic importance would have the same requirements to restore it to as new condition.
- *The location or environment* of the road, e.g. Urban or rural are treated the same – they have the same requirements to restore to 'as new' condition.
- *The traffic that the road is taking.*
- *Maintenance prioritisation and other asset management considerations,* e.g., whether that length may have a higher priority (due to any of the above items or for other reasons). This includes the concept of "asset sweating" in which maintenance may be delayed in order to get the maximum life from the pavement.

However, when determining the condition category you **should take into account** the type of pavement being assessed, when thinking about the level of work needed to restore it to its 'as new' condition, e.g. an A road may have had better "as new" ride quality than a U road. You would not expect to restore the U road to a level that is better than when it was new, or restore a U road it to the level of ride quality that would be expected of an A road – hence this may be taken into account when rating the condition category (extent of work required to restore to as new condition).

It will not be practical during the study to provide road class definitions of all the lengths being surveyed (i.e. engineers may not know if this is a A, B, C or U road), but this guide includes the route maps and shows the road classifications. TRL staff will provide advice on the road classification where practical. However, your best judgement may have to be made for some of the lengths.

Appendix B Example images from the study network

Table 7.2: Example images from the study network, according to the locations shown in Figure 7.1

 1	 2	 3
 4	 5	 6
 7	 8	 9

 10	 11	 12
 13	 14	 15
 16		

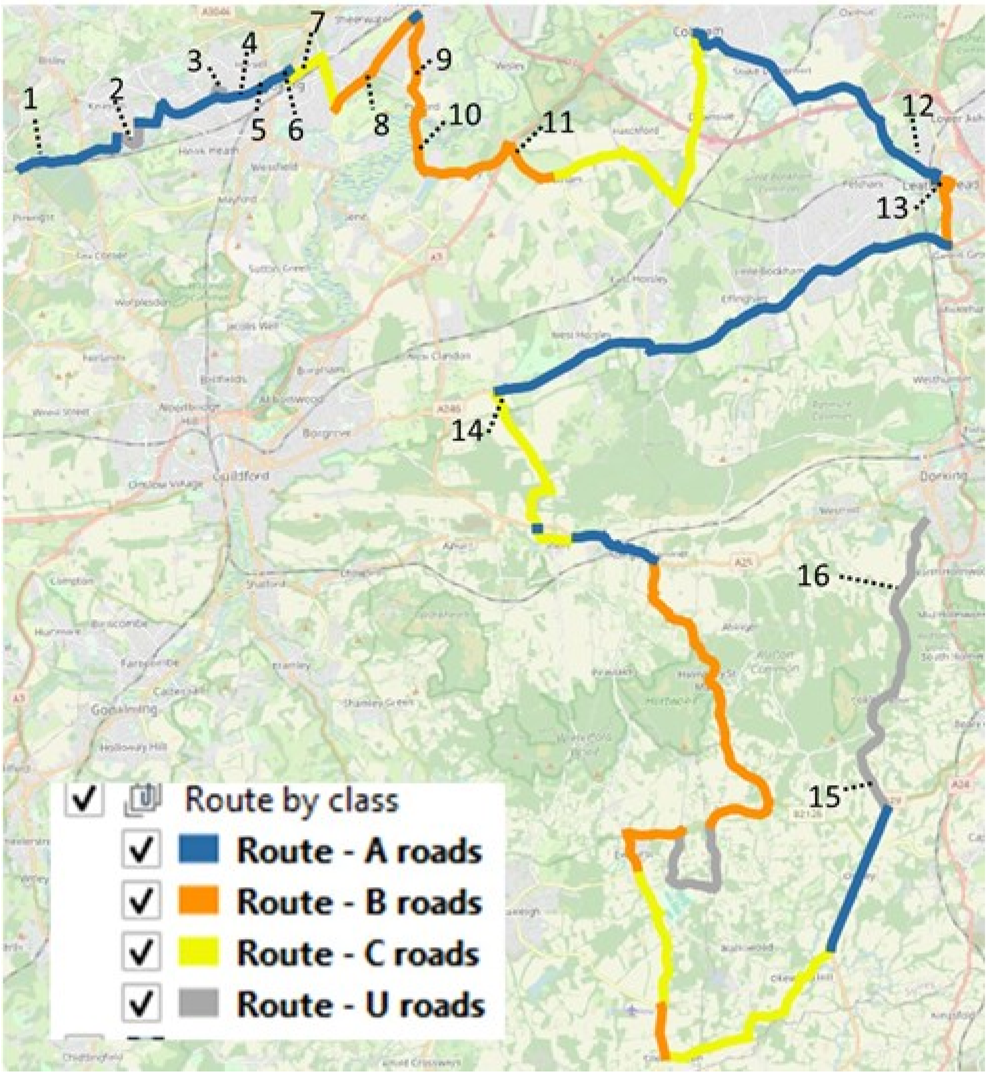


Figure 7.1: Study Network[19]

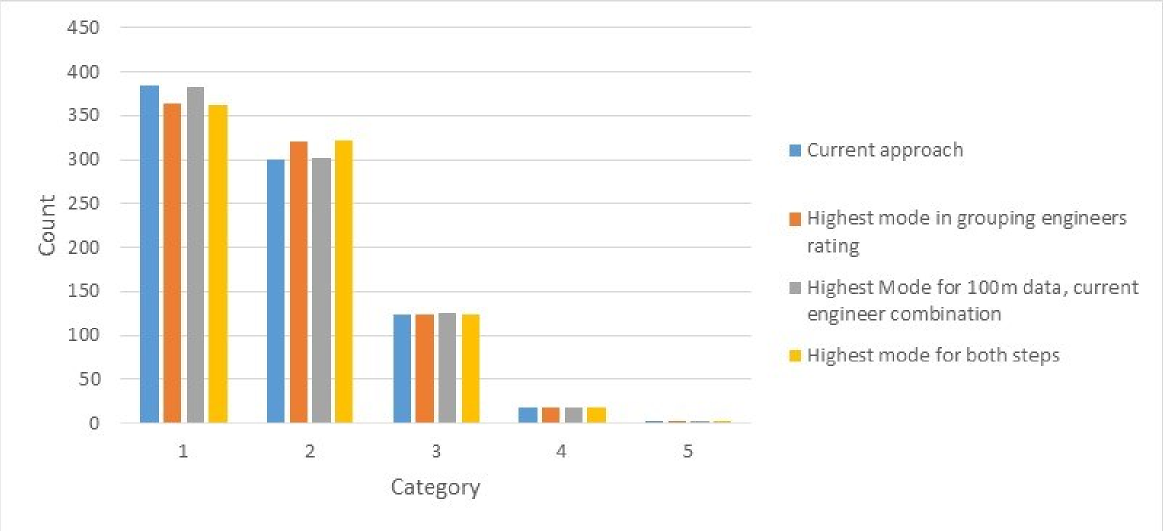
[19] OpenStreetMap contributors. (2015). Map data copyrighted OpenStreetMap contributors and available from https://www.openstreetmap.org

Appendix C Note on the use of the mode

C.1 Calculation of the mode

For this work the benchmark and RCM data were aggregated via calculation of the mode. Some of the data processing was undertaken in Excel using the formula “MODE”. In Excel, where more than one mode exists (multimodal lengths) the first encountered value is reported (e.g. the mode for 1,2,1,2 would be reported as 1, whereas 2,1,2,1 would be reported as 2). Data processing was also undertaken using MS Access, which does not have a mode function, and therefore this was obtained using SQL. The SQL code delivered the lowest value in the case of multimodal lengths (e.g. both 1,2,1,2 and 2,1,2,1 would be reported as 1).

The approach taken in the study therefore presents the risk of inconsistency. These should be minimised in future implementation (e.g in the PAS) or studies (e.g, those approving systems for national reporting). It is suggested that the highest mode value should be used in the case of multimodal value lengths. This will be consistent and will reduce the likelihood that lengths containing a high proportion of high sub-section ratings, will be reported with lower condition category values. Reporting the highest mode value for multi-modal lengths will also result in an increase in the proportion of the overall dataset reported in the higher categories. This may improve the ability to achieve the required number of lengths containing high condition categories when establishing test networks for the approval of RCM technologies. However, it should be noted that the different approaches to the mode only have a small effect on the overall distributions – as shown in the comparisons for the benchmark datasets presented in Figure 7.2. Therefore, the general conclusions from this study should be unaffected.



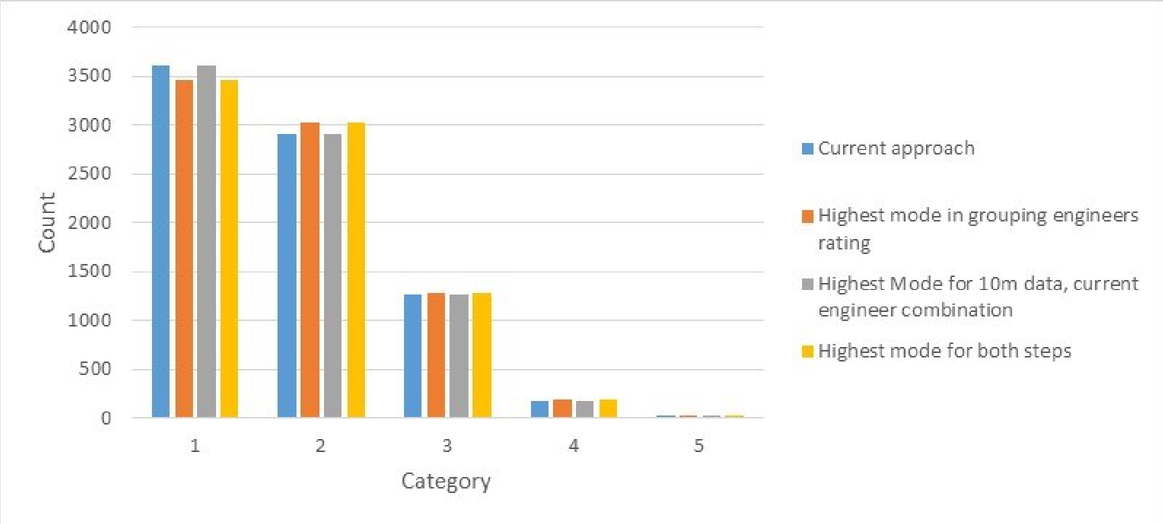


Figure 7.2: Distribution of benchmark dataset by category for different mode calculations (10m dataset bottom, 100m top)

Appendix D Results – Updated RCM data

D.1 Contents of this appendix

D.1.1 100m category values reported by section and chainage

The 100m category values are reported along the length of the study network for each device. These plots also show the benchmark dataset and have the following formatting:

- All classes of road are shown. A roads lengths are identified with blue shading, B roads with orange shading, C roads with yellow shading lines and U roads with grey shading
- The blue line is the benchmark dataset, and the orange line is the RCM dataset.

D.1.2 Performance measures

The summary tables of statistics are provided for each road class. These contain:

- Chi Sq. The Chi-square test statistic is shown for each RCM distribution. The critical value (confidence level 0.05, 3 degrees of freedom (because categories 4 and 5 have been combined) for this data is 7.815. Values lower than this are highlighted green, with values less than 2xcritical value highlighted yellow and 3xcritcal value orange.
- P-Value. The P-value is shown for the comparison between the benchmark and RCM distributions. To highlight differences, P-values>0.05 have been considered to show no statistically significant difference between the distributions. Values higher than this are highlighted green.
- Pearson. The Pearson correlation coefficient between the distributions. Values higher than 0.8 (strong correlation) are highlighted green.
- Diff in % 4/5: The difference between the percentage of the study network reported by the benchmark and by the RCM for condition categories 4 and 5. Differences <5% (absolute) are shown in green.
- Fractional delta Δ1, Δ2, Δ3, Δ4/5. Values lower than 0.75 are shown in green.
- Precision 0, 1. The precision values summarising the overall spatial agreement with the benchmark. Values> 75%, are highlighted green. Values >50% are highlighted yellow.

D.1.3 Bar charts

The bar charts show

- The histograms (reported as %) of the proportion of the network reported within each category by each device. Category 4 and 5 are reported as a combined value.
- The difference between the percent of the network reported as category 4 and 5 by the RCM device and by the benchmark. This is a graphical representation of the data presented in the table in column "Diff in % 4/5".

D.1.4 Confusion matrices

The confusion matrices follow the bar charts. A grid of 16 confusion matrices is shown (one for each RCM dataset). For example, Table 7.3 for device L on U roads

Table 7.3: Example confusion matrix for participant L on U roads

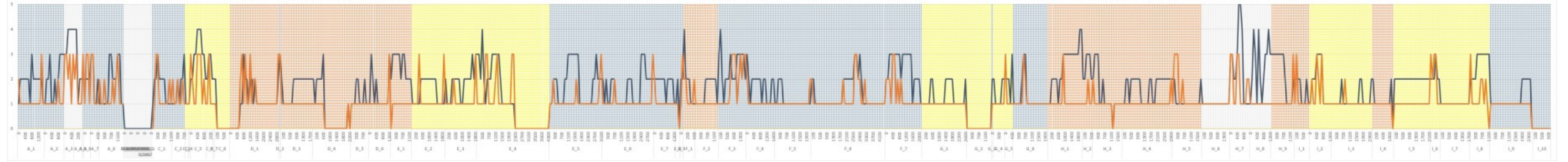
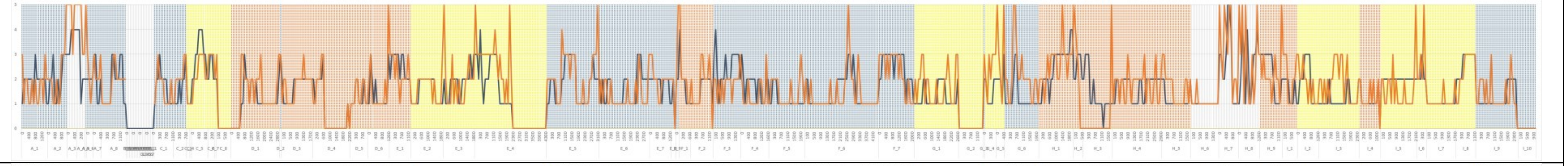
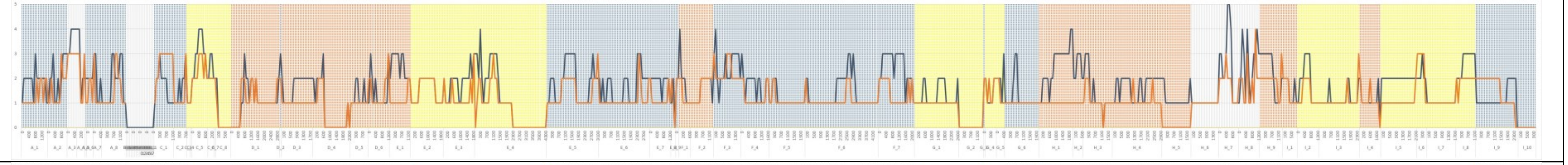
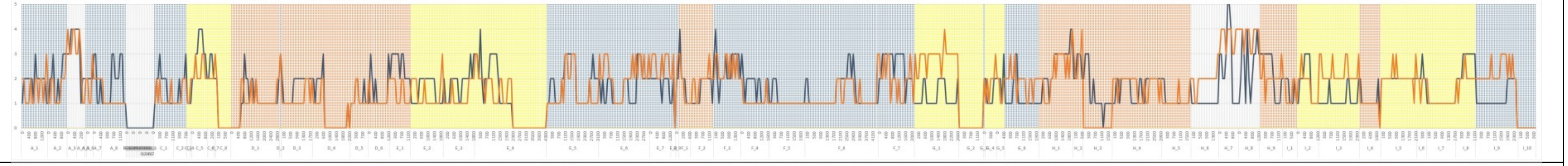
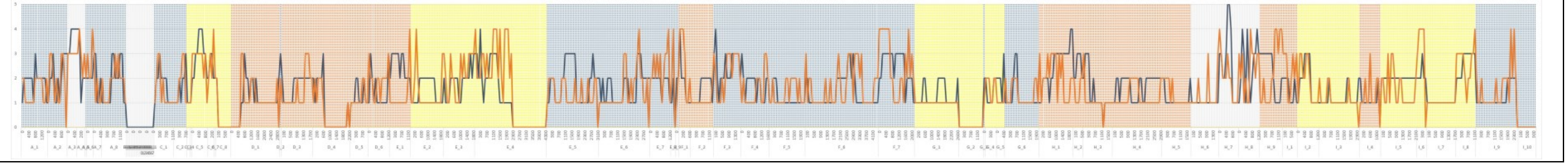
Data	Precision 0	Precision 1
L	43.4%	71.7%

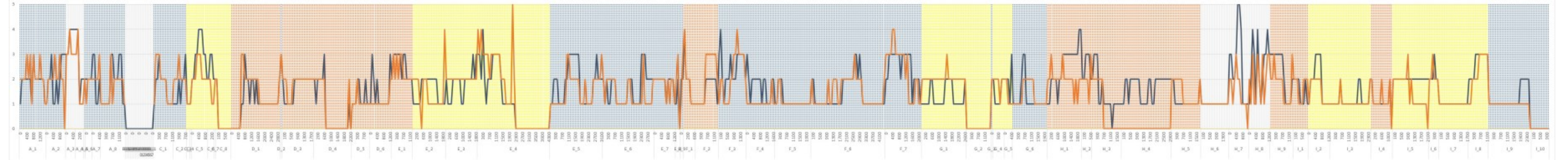
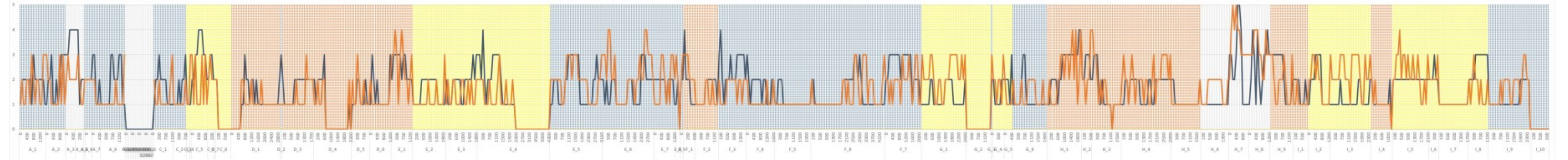
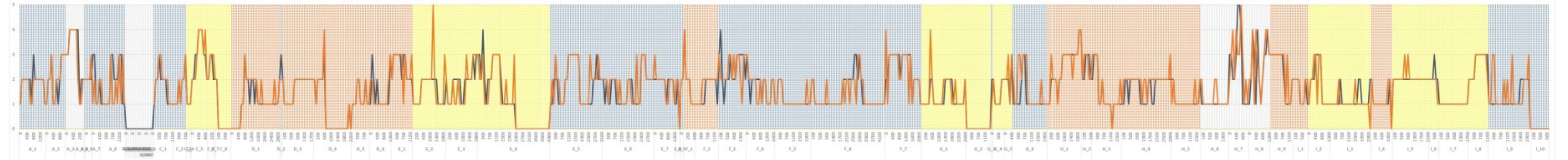
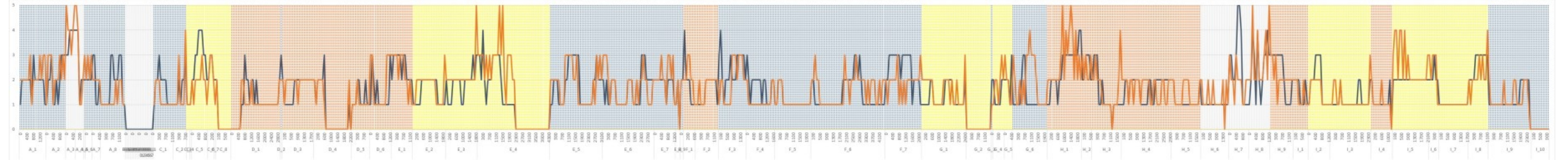
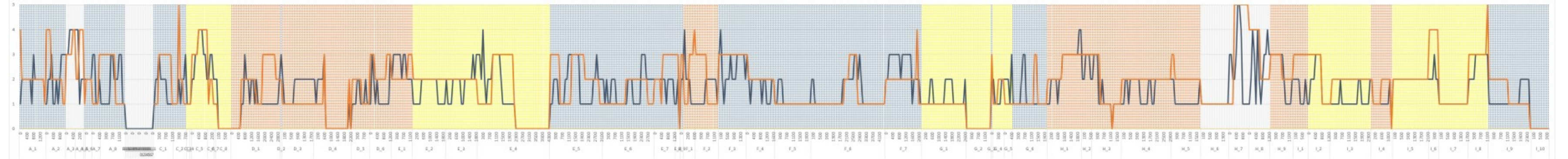
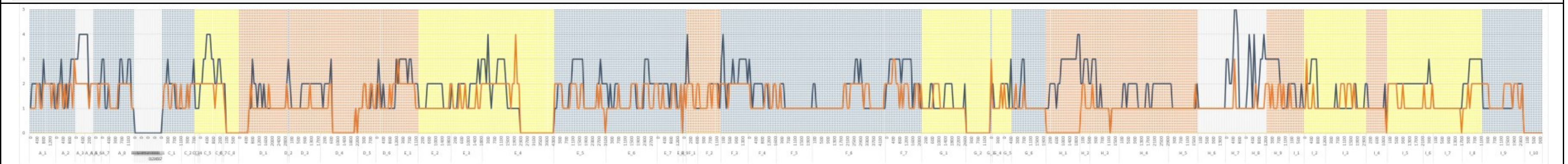
		RCM			
		1	2	3	4/5
Bench.	1	60	16	12	12
	2	0	29	43	29
	3	11	33	22	33
	4/5	0	50	17	33

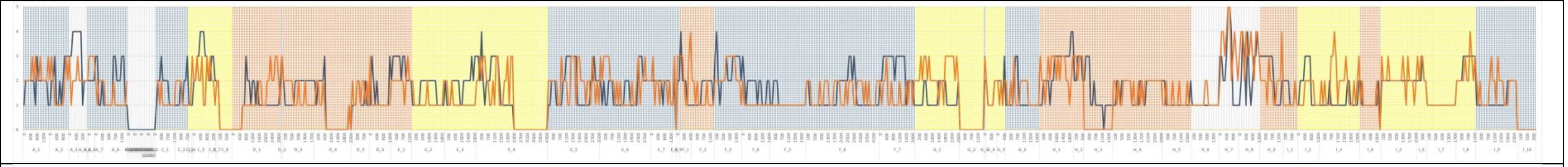
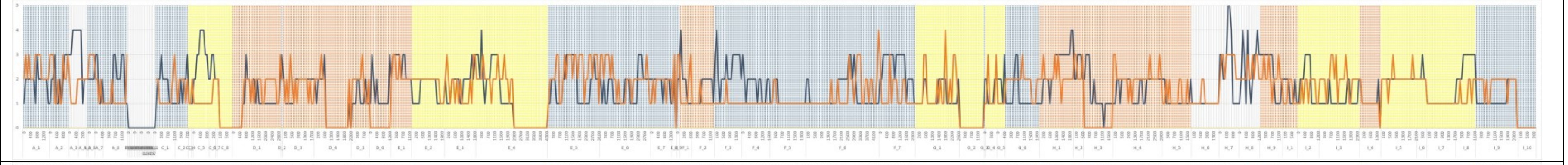
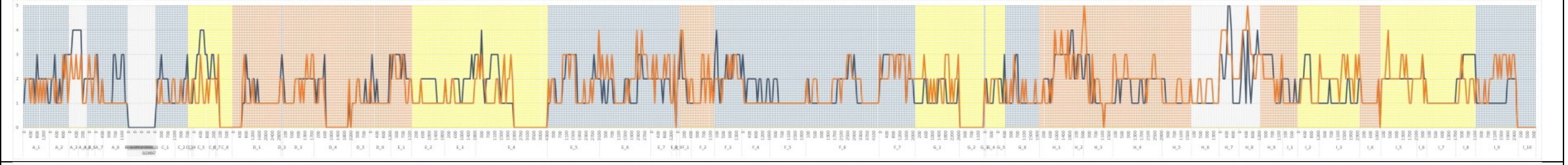
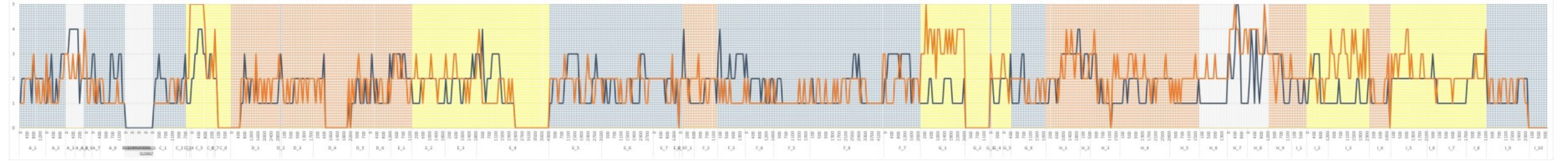
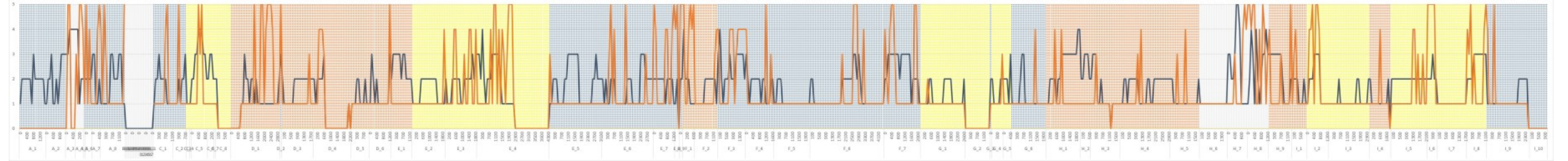
- E.g. 22% of the lengths reported as category 3 by the benchmark were also reported as category 3 by the RCM, 17% of the lengths reported as 4/5 by the benchmark were reported as 3 by the RCM.
- The leading diagonal is colour coded using red where the value is less than 25%, amber (25 to 50%) and yellow (50 to 75%). Values >75% are shown as green.
- The precision values (which are also shown in the summary table) are also shown at the top of the matrix.

D.2 Categories reported by section and chainage

Table 7.4: Plots of condition category values for each participant (orange line is RCM data, blue line is benchmark)

A	
B	
C	
D	
E	

F	
G	
H	
I	
J	
K	

L	
M	
N	
O	
P	

D.3 100m Performance measures, A roads

Table 7.5: Summary: 100m data, A roads

Data	Chi Sq	P-Value	Pearson	Diff in % 4/5	Delta Δ1	Delta Δ2	Delta Δ3	Delta Δ4/5	Precision 0	Precision 1
A	168.90	0.000000	0.77	-0.3%	0.76	0.74	0.53	1.00	53.3%	90.4%
B	49.19	0.000000	1.00	2.2%	0.03	0.01	0.02	7.00	52.5%	94.8%
C	59.56	0.000000	0.91	-0.3%	0.44	0.30	0.61	1.00	53.3%	92.9%
D	1.05	0.788377	1.00	-0.3%	0.00	0.02	0.02	1.00	51.5%	92.6%
E	172.27	0.000000	0.98	4.0%	0.05	0.10	0.18	13.00	49.8%	91.3%
F	9.57	0.022584	0.99	0.6%	0.11	0.05	0.26	2.00	65.6%	97.2%
G	13.69	0.003359	0.99	0.9%	0.11	0.08	0.20	3.00	50.0%	90.1%
H	1.31	0.726541	1.00	0.0%	0.06	0.03	0.12	0.00	79.0%	94.8%
I	1.62	0.653814	1.00	0.3%	0.00	0.03	0.10	1.00	58.3%	93.5%
J	34.50	0.000000	0.98	1.5%	0.06	0.13	0.37	5.00	50.3%	88.0%
K	52.83	0.000000	0.96	-0.3%	0.18	0.17	0.92	1.00	60.7%	96.9%
L	4.58	0.205037	0.99	-0.3%	0.00	0.10	0.22	1.00	45.7%	90.4%
M	0.12	0.989153	1.00	0.0%	0.01	0.03	0.02	0.00	40.1%	87.3%
N	4.01	0.259855	1.00	0.6%	0.01	0.01	0.00	2.00	54.6%	94.1%
O	38.75	0.000000	0.95	0.0%	0.05	0.27	0.76	0.00	45.1%	92.3%
P	2400.64	0.000000	0.63	15.9%	0.64	0.99	0.83	47.00	44.6%	74.7%

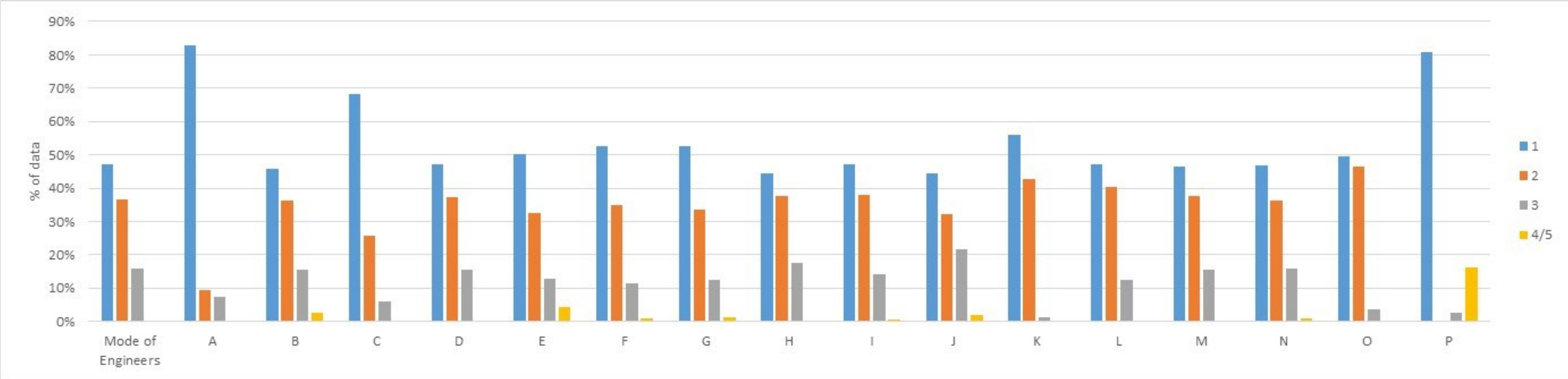


Figure 7.3: Distribution of ratings for participants A to P, A road 100m data

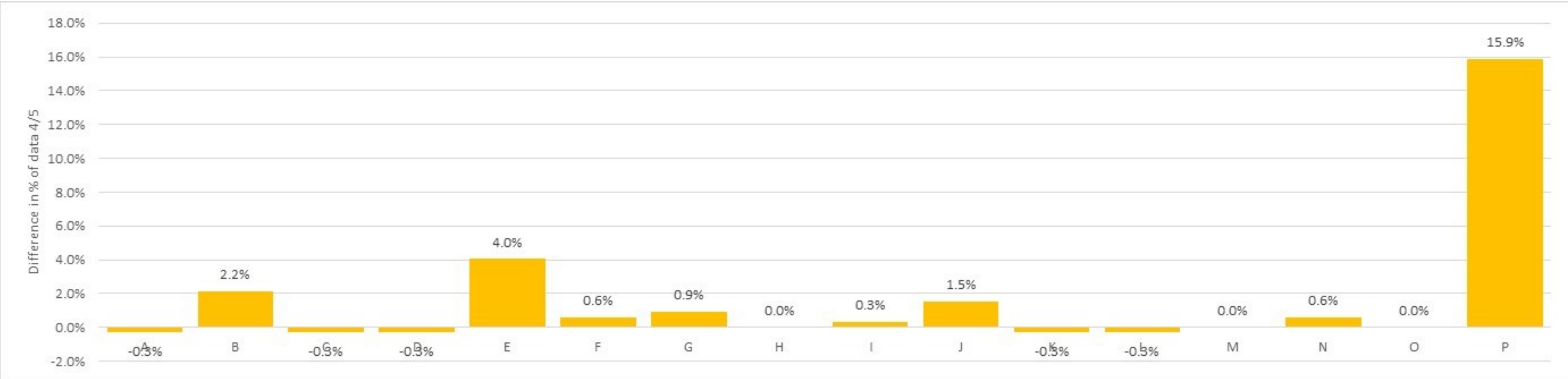


Figure 7.4: Difference between % recorded by benchmark and RCM dataset for categories 4 and 5 for participants A to P, A road 100m data

Table 7.6: Confusion matrices (percentages of benchmark), 100m data A roads

Data	Precision 0	Precision 1
A	53.3%	90.4%

		RCM			
		1	2	3	4/5
Bench.	1	95	5	0	0
	2	77	13	10	0
	3	59	18	24	0
	4/5	100	0	0	0

Data	Precision 0	Precision 1
B	52.5%	94.8%

		RCM			
		1	2	3	4/5
Bench.	1	65	32	3	0
	2	39	39	18	4
	3	6	41	47	6
	4/5	0	100	0	0

Data	Precision 0	Precision 1
C	53.3%	92.9%

		RCM			
		1	2	3	4/5
Bench.	1	86	12	3	0
	2	60	31	9	0
	3	35	55	10	0
	4/5	0	100	0	0

Data	Precision 0	Precision 1
D	51.5%	92.6%

		RCM			
		1	2	3	4/5
Bench.	1	63	28	8	0
	2	38	45	17	0
	3	22	47	31	0
	4/5	0	0	100	0

Data	Precision 0	Precision 1
E	49.8%	91.3%

		RCM			
		1	2	3	4/5
Bench.	1	69	27	3	0
	2	35	37	21	7
	3	28	38	22	12
	4/5	0	100	0	0

Data	Precision 0	Precision 1
F	65.6%	97.2%

		RCM			
		1	2	3	4/5
Bench.	1	83	16	1	0
	2	30	55	14	0
	3	14	42	38	6
	4/5	0	100	0	0

Data	Precision 0	Precision 1
G	50.0%	90.1%

		RCM			
		1	2	3	4/5
Bench.	1	67	24	8	1
	2	42	41	16	1
	3	31	47	20	2
	4/5	100	0	0	0

Data	Precision 0	Precision 1
H	79.0%	94.8%

		RCM			
		1	2	3	4/5
Bench.	1	81	13	6	0
	2	12	79	8	1
	3	12	14	75	0
	4/5	0	100	0	0

Data	Precision 0	Precision 1
I	58.3%	93.5%

		RCM			
		1	2	3	4/5
Bench.	1	73	20	7	1
	2	27	54	19	0
	3	18	55	25	2
	4/5	0	100	0	0

Data	Precision 0	Precision 1
J	50.3%	88.0%

		RCM			
		1	2	3	4/5
Bench.	1	63	22	14	1
	2	30	41	25	3
	3	24	41	35	0
	4/5	0	0	100	0

Data	Precision 0	Precision 1
K	60.7%	96.9%

		RCM			
		1	2	3	4/5
Bench.	1	81	19	0	0
	2	41	58	1	0
	3	18	76	6	0
	4/5	0	100	0	0

Data	Precision 0	Precision 1
L	45.7%	90.4%

		RCM			
		1	2	3	4/5
Bench.	1	56	38	6	0
	2	39	44	18	0
	3	41	39	20	0
	4/5	0	100	0	0

Data	Precision 0	Precision 1
M	40.1%	87.3%

		RCM			
		1	2	3	4/5
Bench.	1	50	39	11	1
	2	44	37	19	0
	3	43	37	20	0
	4/5	100	0	0	0

Data	Precision 0	Precision 1
N	54.6%	94.1%

		RCM			
		1	2	3	4/5
Bench.	1	68	26	5	1
	2	34	45	20	1
	3	16	45	37	2
	4/5	0	100	0	0

Data	Precision 0	Precision 1
O	45.1%	92.3%

		RCM			
		1	2	3	4/5
Bench.	1	56	41	3	0
	2	45	49	5	1
	3	37	59	4	0
	4/5	100	0	0	0

Data	Precision 0	Precision 1
P	44.6%	74.7%

		RCM			
		1	2	3	4/5
Bench.	1	88	0	2	10
	2	76	1	3	20
	3	67	0	4	28
	4/5	100	0	0	0

D.4 100m Performance measures, B and C roads

Table 7.7: Summary of updated results, 100m data B and C roads

Data	Chi Sq	P-Value	Pearson	Diff in % 4/5	Delta Δ1	Delta Δ2	Delta Δ3	Delta Δ4/5	Precision 0	Precision 1
A	336.74	0.000000	0.69	-1.6%	0.93	0.88	0.51	1.00	46.4%	86.7%
B	32.38	0.000000	0.97	1.8%	0.18	0.03	0.53	1.14	52.4%	91.9%
C	152.50	0.000000	0.84	-1.6%	0.65	0.40	0.78	1.00	55.0%	94.0%
D	4.99	0.172832	0.99	-0.9%	0.05	0.10	0.08	0.57	49.9%	92.2%
E	85.17	0.000000	0.86	3.6%	0.29	0.42	0.05	2.29	48.1%	87.1%
F	30.43	0.000001	0.93	-0.5%	0.15	0.34	0.36	0.29	61.2%	96.3%
G	7.97	0.046707	0.98	0.2%	0.09	0.17	0.14	0.14	48.9%	90.5%
H	2.54	0.468681	1.00	0.7%	0.05	0.00	0.11	0.43	84.0%	98.4%
I	23.66	0.000029	0.96	2.0%	0.16	0.18	0.14	1.29	61.9%	94.0%
J	58.90	0.000000	0.86	1.5%	0.33	0.12	0.64	1.00	48.9%	90.0%
K	115.76	0.000000	0.90	-0.9%	0.50	0.27	0.96	0.80	50.2%	94.9%
L	1.36	0.715776	1.00	-0.7%	0.01	0.00	0.03	0.43	45.7%	90.3%
M	31.39	0.000001	0.94	-1.4%	0.14	0.30	0.34	0.86	47.9%	91.0%
N	2.68	0.443818	1.00	0.4%	0.06	0.03	0.14	0.29	51.1%	92.3%
O	289.05	0.000000	0.53	7.3%	0.65	0.50	0.22	4.71	35.0%	86.3%
P	732.45	0.000000	0.56	11.5%	0.82	0.99	0.73	7.43	41.2%	77.2%

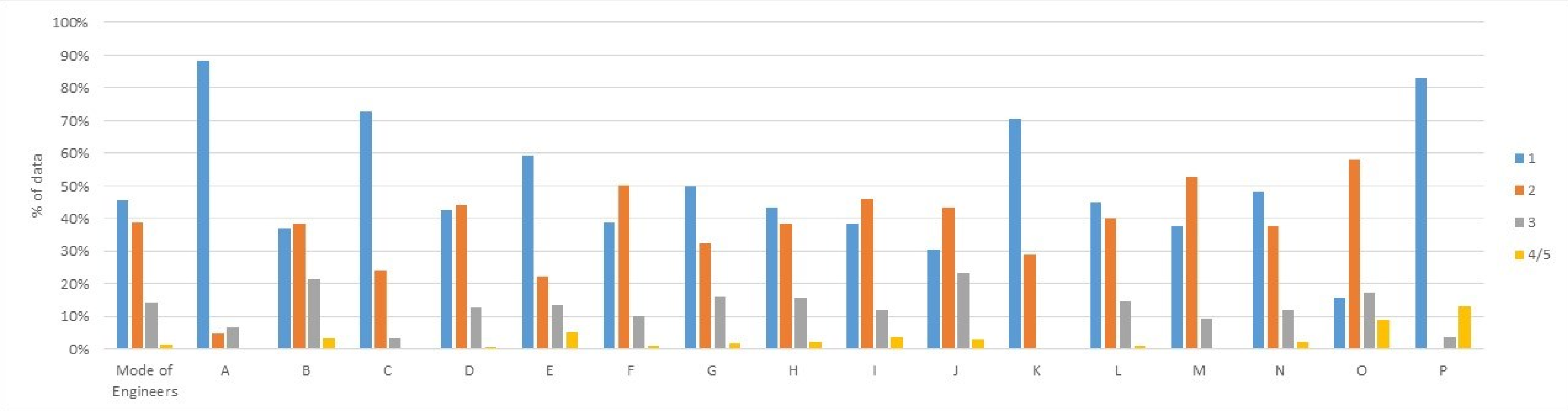


Figure 7.5: Distribution of ratings for participants A to P, B and C roads 100m data

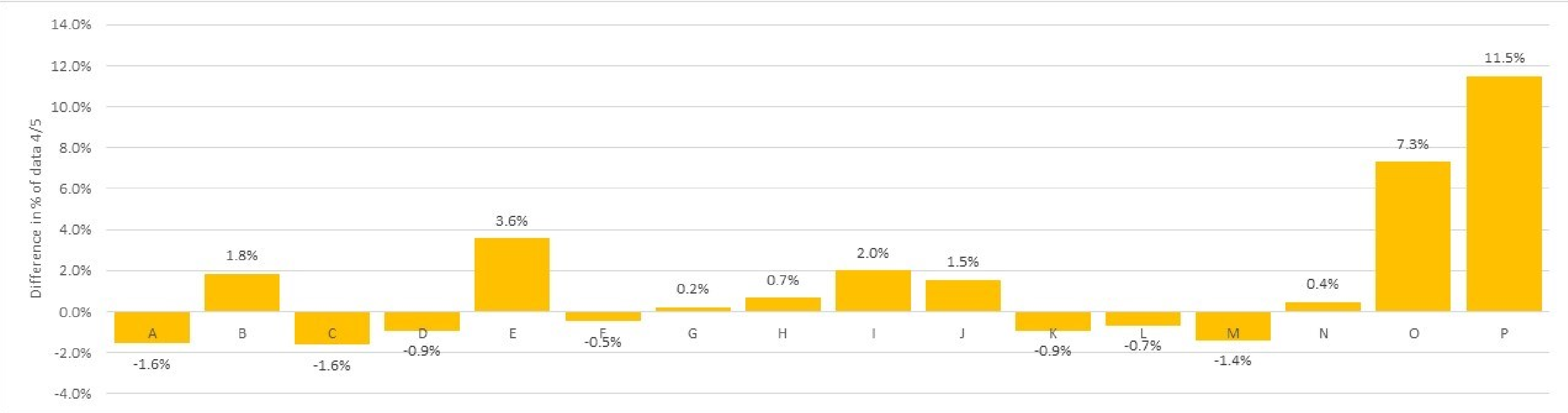


Figure 7.6: Difference between % recorded by benchmark and RCM dataset for categories 4 and 5 for participants A to P, 100m data B and C roads

Table 7.8: Confusion matrices (percentages of benchmark), 100m data B and C roads

Data	Precision 0	Precision 1
A	46.4%	86.7%

		RCM			
		1	2	3	4/5
Bench.	1	94	2	4	0
	2	89	5	7	0
	3	76	11	13	0
	4/5	43	14	43	0

Data	Precision 0	Precision 1
B	52.4%	91.9%

		RCM			
		1	2	3	4/5
Bench.	1	59	30	10	1
	2	26	49	22	4
	3	3	40	48	8
	4/5	0	0	86	14

Data	Precision 0	Precision 1
C	55.0%	94.0%

		RCM			
		1	2	3	4/5
Bench.	1	94	6	0	0
	2	65	31	3	0
	3	35	57	8	0
	4/5	29	29	43	0

Data	Precision 0	Precision 1
D	49.9%	92.2%

		RCM			
		1	2	3	4/5
Bench.	1	58	31	11	0
	2	38	51	10	1
	3	13	62	25	0
	4/5	0	43	43	14

Data	Precision 0	Precision 1
E	48.1%	87.1%

		RCM			
		1	2	3	4/5
Bench.	1	74	17	6	4
	2	51	27	17	5
	3	42	23	25	9
	4/5	0	43	43	14

Data	Precision 0	Precision 1
F	61.2%	96.3%

		RCM			
		1	2	3	4/5
Bench.	1	67	30	2	1
	2	18	70	11	1
	3	9	61	28	2
	4/5	0	43	43	14

Data	Precision 0	Precision 1
G	48.9%	90.5%

		RCM			
		1	2	3	4/5
Bench.	1	67	25	8	0
	2	39	38	21	2
	3	28	41	27	5
	4/5	29	29	29	14

Data	Precision 0	Precision 1
H	84.0%	98.4%

		RCM			
		1	2	3	4/5
Bench.	1	86	12	1	0
	2	9	82	8	1
	3	3	11	83	3
	4/5	0	0	14	86

Data	Precision 0	Precision 1
I	61.9%	94.0%

		RCM			
		1	2	3	4/5
Bench.	1	67	28	4	1
	2	18	67	11	3
	3	6	42	38	14
	4/5	0	57	43	0

Data	Precision 0	Precision 1
J	48.9%	90.0%

		RCM			
		1	2	3	4/5
Bench.	1	40	44	16	0
	2	29	53	15	3
	3	5	20	69	6
	4/5	29	0	29	43

Data	Precision 0	Precision 1
K	50.2%	94.9%

		RCM			
		1	2	3	4/5
Bench.	1	83	16	0	0
	2	70	28	1	0
	3	30	70	0	0
	4/5	0	100	0	0

Data	Precision 0	Precision 1
L	45.7%	90.3%

		RCM			
		1	2	3	4/5
Bench.	1	55	32	12	1
	2	43	44	12	1
	3	20	52	26	2
	4/5	14	43	43	0

Data	Precision 0	Precision 1
M	47.9%	91.0%

		RCM			
		1	2	3	4/5
Bench.	1	48	44	8	0
	2	30	61	9	1
	3	25	59	15	0
	4/5	57	43	0	0

Data	Precision 0	Precision 1
N	51.1%	92.3%

		RCM			
		1	2	3	4/5
Bench.	1	66	28	6	0
	2	37	46	15	2
	3	25	47	22	6
	4/5	14	29	43	14

Data	Precision 0	Precision 1
O	35.0%	86.3%

		RCM			
		1	2	3	4/5
Bench.	1	19	62	11	8
	2	14	58	22	6
	3	11	50	25	14
	4/5	14	14	29	43

Data	Precision 0	Precision 1
P	41.2%	77.2%

		RCM			
		1	2	3	4/5
Bench.	1	88	0	1	10
	2	82	0	5	13
	3	73	0	6	20
	4/5	43	0	14	43

D.5 100m Performance measures Results, U roads

Table 7.9: Summary of updated results, 100m data U roads

Data	Chi Sq	P-Value	Pearson	Diff in % 4/5	Delta Δ1	Delta Δ2	Delta Δ3	Delta Δ4/5	Precision 0	Precision 1
A	22.29	0.000	0.92	-22.6%	0.60	0.43	0.00	1.00	52.8%	75.5%
B	2.70	0.441	0.90	9.4%	0.12	0.14	0.11	0.42	43.4%	79.2%
C	16.03	0.001	0.74	-20.8%	0.12	0.86	0.22	0.92	58.5%	86.8%
D	34.75	0.000	-0.69	18.9%	0.84	1.00	0.44	0.83	24.5%	75.5%
E	19.79	0.000	0.49	-17.0%	0.16	0.43	1.11	0.75	52.8%	81.1%
F	10.98	0.012	0.84	-19.2%	0.25	0.14	0.33	0.83	59.6%	90.4%
G	37.49	0.000	-0.75	-3.8%	0.64	1.86	0.56	0.17	20.8%	71.7%
H	4.57	0.206	0.93	0.0%	0.20	0.71	0.00	0.00	66.0%	86.8%
I	33.53	0.000	0.14	-1.9%	0.33	2.00	0.56	0.08	50.0%	88.5%
J	10.67	0.014	0.68	17.0%	0.20	0.14	0.56	0.75	49.1%	66.0%
K	31.68	0.000	0.86	-22.6%	0.60	0.71	0.89	1.00	52.8%	66.0%
L	12.49	0.006	0.55	0.0%	0.36	1.14	0.11	0.00	43.4%	71.7%
M	46.58	0.000	0.15	-22.6%	0.24	2.14	0.33	1.00	43.4%	77.4%
N	23.88	0.000	-0.11	-5.7%	0.44	1.57	0.33	0.25	30.2%	79.2%
O	62.48	0.000	-0.97	7.5%	1.00	2.14	0.67	0.33	17.0%	62.3%
P	16.57	0.001	0.90	14.3%	0.20	1.00	0.63	0.78	42.9%	57.1%

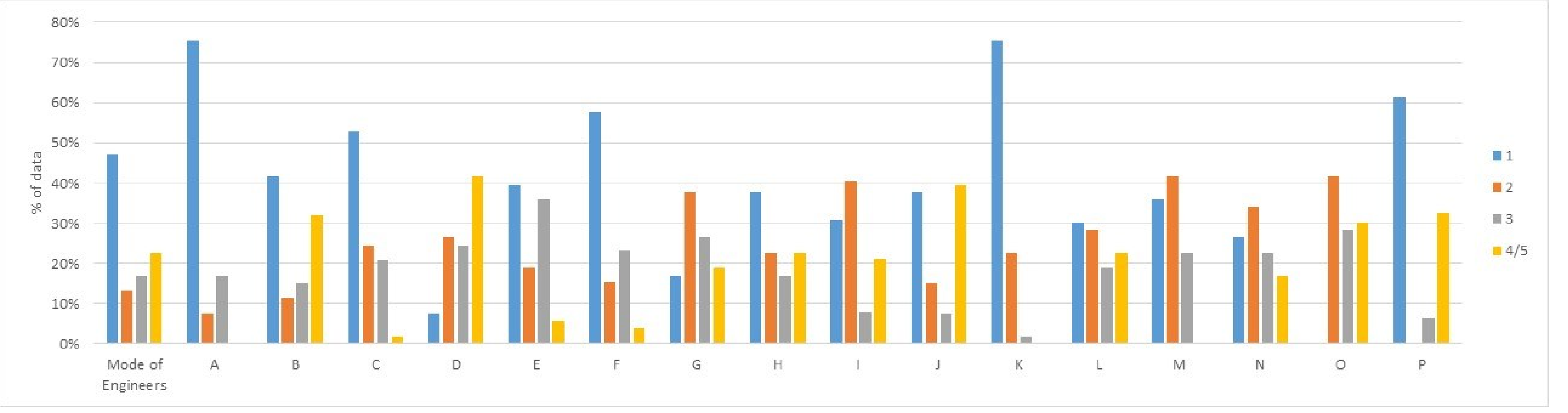


Figure 7.7: Distribution of ratings for participants A to P, 100m data U roads

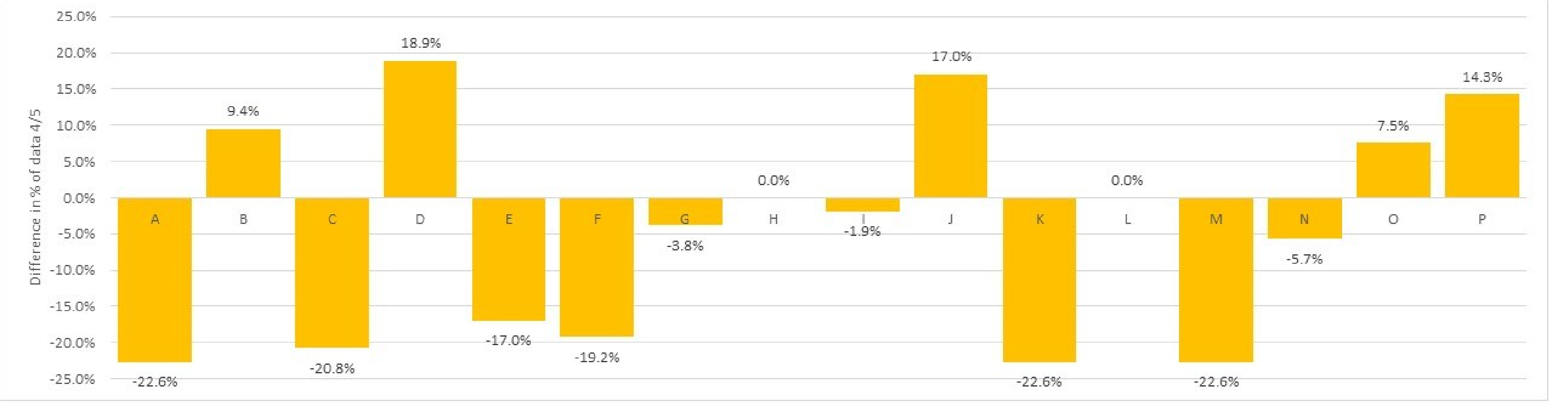


Figure 7.8: Difference between % recorded by benchmark and RCM dataset for categories 4 and 5 for participants A to P, 100m data U roads

Table 7.10: Confusion matrices (percentages of benchmark), 100m data U roads

Data	Precision 0	Precision 1
A	52.8%	75.5%

		RCM 1	RCM 2	RCM 3	RCM 4/5
Bench.	1	92	8	0	0
	2	100	0	0	0
	3	44	0	56	0
	4/5	50	17	33	0

Data	Precision 0	Precision 1
B	43.4%	79.2%

		RCM 1	RCM 2	RCM 3	RCM 4/5
Bench.	1	76	12	0	12
	2	14	14	57	14
	3	22	11	22	44
	4/5	0	8	17	75

Data	Precision 0	Precision 1
C	58.5%	86.8%

		RCM 1	RCM 2	RCM 3	RCM 4/5
Bench.	1	96	4	0	0
	2	29	57	14	0
	3	22	33	33	11
	4/5	0	42	58	0

Data	Precision 0	Precision 1
D	24.5%	75.5%

		RCM 1	RCM 2	RCM 3	RCM 4/5
Bench.	1	8	56	20	16
	2	29	0	14	57
	3	0	0	44	56
	4/5	0	0	25	75

Data	Precision 0	Precision 1
E	52.8%	81.1%

		RCM 1	RCM 2	RCM 3	RCM 4/5
Bench.	1	72	8	16	4
	2	14	57	29	0
	3	11	22	56	11
	4/5	8	17	67	8

Data	Precision 0	Precision 1
F	59.6%	90.4%

		RCM 1	RCM 2	RCM 3	RCM 4/5
Bench.	1	96	4	0	0
	2	57	29	14	0
	3	22	33	44	0
	4/5	8	17	58	17

Data	Precision 0	Precision 1
G	20.8%	71.7%

		RCM 1	RCM 2	RCM 3	RCM 4/5
Bench.	1	28	44	24	4
	2	29	14	29	29
	3	0	33	22	44
	4/5	0	42	33	25

Data	Precision 0	Precision 1
H	66.0%	86.8%

		RCM 1	RCM 2	RCM 3	RCM 4/5
Bench.	1	72	20	4	4
	2	14	71	0	14
	3	11	11	56	22
	4/5	0	8	25	67

Data	Precision 0	Precision 1
I	50.0%	88.5%

		RCM 1	RCM 2	RCM 3	RCM 4/5
Bench.	1	58	42	0	0
	2	29	71	0	0
	3	0	22	33	44
	4/5	0	33	8	58

Data	Precision 0	Precision 1
J	49.1%	66.0%

		RCM 1	RCM 2	RCM 3	RCM 4/5
Bench.	1	68	4	0	28
	2	14	14	0	71
	3	22	33	22	22
	4/5	0	25	17	58

Data	Precision 0	Precision 1
K	52.8%	66.0%

		RCM 1	RCM 2	RCM 3	RCM 4/5
Bench.	1	100	0	0	0
	2	57	43	0	0
	3	67	33	0	0
	4/5	42	50	8	0

Data	Precision 0	Precision 1
L	43.4%	71.7%

		RCM 1	RCM 2	RCM 3	RCM 4/5
Bench.	1	60	16	12	12
	2	0	29	43	29
	3	11	33	22	33
	4/5	0	50	17	33

Data	Precision 0	Precision 1
M	43.4%	77.4%

		RCM 1	RCM 2	RCM 3	RCM 4/5
Bench.	1	60	32	8	0
	2	0	71	29	0
	3	11	56	33	0
	4/5	25	33	42	0

Data	Precision 0	Precision 1
N	30.2%	79.2%

		RCM 1	RCM 2	RCM 3	RCM 4/5
Bench.	1	48	44	4	4
	2	29	0	29	43
	3	0	33	33	33
	4/5	0	33	50	17

Data	Precision 0	Precision 1
O	17.0%	62.3%

		RCM 1	RCM 2	RCM 3	RCM 4/5
Bench.	1	0	60	28	12
	2	0	29	14	57
	3	0	11	44	44
	4/5	0	33	25	42

Data	Precision 0	Precision 1
P	42.9%	57.1%

		RCM 1	RCM 2	RCM 3	RCM 4/5
Bench.	1	72	0	0	28
	2	57	0	0	43
	3	25	0	25	50
	4/5	67	0	11	22

Road Condition Monitoring Data

Network Study

In England, Local Authorities provide network-level Road Condition Monitoring (RCM) data to the Department for Transport each year, so that the DfT can publish annual official statistics for road condition in England (national reporting). Whilst a regime has been in place to collect this data since 2009, during this period significant progress has been made in the development of new and alternative technologies for the collection of data. This provides an opportunity to reconsider the approach taken to the collection and reporting of condition on the local road network, for both local and national reporting. The transition to a new regime is to be achieved via the introduction of a Publicly Available Standard (PAS 2161). This report describes a network study carried out to provide insight on the collection of RCM data, to support the development of the PAS. The study has compared the data provided by RCM technologies on a study network with benchmark data provided by engineers' surveys. The work has assisted in the development of definitions for a new set of condition categories that will be deployed for national reporting (in the PAS), established an understanding the data provided by current systems and the comparability of the condition categories reported, and insight into the implications for achieving consistency in a future PAS compliant data collection regime.

Other titles from this subject area

PPR1021	Highways England 2021 national accreditation for sideway-force skid resistance devices. S. Brittain. 2022
PPR863	Applications of machine learning in transport. K Nesnas, R Khatry, A Smirnov, D Peeling, S Mistry, M Crabtree, S Reeve. 2018
PPR991	Developing a Road Condition Indicator for Fretting. Emma Benbow, Mohammed Rweished, Daniel Spearman. 2018
PR974	Development of SCANNER and UKPMS: Task 1 - Consistency of SCANNER data and Task 2 - SCANNER Condition Parameters. E Benbow, A Wright, N Dhillon, M Harrington and K Nesnas. 2017
PPR290	Smaller, quicker, cheaper automated carriageway surveys. K Gallagher, A Wright, P King and A Pickett. 2009

TRL
Crowthorne House, Nine Mile Ride,
Wokingham, Berkshire, RG40 3GA,
United Kingdom
T: +44 (0) 1344 773131
F: +44 (0) 1344 770356
E: enquiries@trl.co.uk
W: www.trl.co.uk

ISSN: 2514-96522514-9652

DOI: 10.58446/mdft1813

PPR2042