ICES WKNEPS REPORT 2018

ECOSYSTEM OBSERVATION STEERING GROUP

ICES CM 2018/EOSG:25

REF ACOM AND SCICOM

Report of the Workshop on *Nephrops* Burrow Counting (WKNEPS)

2-5 October

Aberdeen, UK

ICES
International Council for the Exploration of the Sea

CIEM
Conseil International pour l'Exploration de la Mer

International Council for the Exploration of the Sea
Conseil International pour l'Exploration de la Mer

H. C. Andersens Boulevard 44–46
DK-1553 Copenhagen V
Denmark
Telephone (+45) 33 38 67 00
Telefax (+45) 33 93 42 15
www.ices.dk
info@ices.dk

Recommended format for purposes of citation:

ICES. 2018. Report of the Workshop on *Nephrops* Burrow Counting (WKNEPS). 2-5 October. Aberdeen, UK. ICES CM 2018/EOSG:25. 44 pp.
https://doi.org/10.17895/ices.pub.8180

Contents

Executive summary

The Workshop on ***Nephrops*** **burrow counting (WKNEPS)**, met in Aberdeen, Scotland on 2–5 October 2018. The workshop included 20 participants from eight countries (Iceland, Ireland, UK-Scotland, UK-England and Wales, UK-Northern Ireland, Sweden, Spain and France). Previous workshops WKNEPHBID (ICES, 2008) and WKNEPS (ICES, 2017a) developed standards for the compilation of survey specific reference sets, how to generate reference counts to test reviewer counting skills and also the content and use of *Nephrops* burrow training material. This workshop investigated the use of Lin's Concordance Correlation Coefficient (CCC), which measures the ability of counters to exactly reproduce each other's counts using reference image sets and to develop and test a standard approach to generate reference counts. Reference sets from six Functional Units (FU) were presented to test this process: Farn Deeps (FU 6), North Minch (FU 11), western Irish Sea (FU 15), Labadie, Jones and Cockburn Banks (FU 20-21), Bay of Biscay (FU 23-24) and Gulf of Cadiz (FU 30). Reference set from Iceland (FU 1) was available and reference counts were generated by consensus and also annotated. Burrow count training using survey footage from FU 3-4 and data quality review of the all reference sets were also carried out during the workshop in group sessions.

1 WKNEPS Terms of Reference

The **Workshop on *Nephrops* burrow counting (WKNEPS)**, co-chaired by Adrian Weetman and Jennifer Doyle will meet at Marine Science Scotland, Aberdeen, Scotland, UK during **2 – 5 October 2018** to:

a) To build capacity in burrowing counting skills and support counting procedures for new and developing surveys across Europe. (Science Plan Topic addressed, 31)

b) To analyse challenges and differences among *Nephrops* grounds. (Science Plan Topic addressed, 31)

c) To update the UWTV SISP based on WKNEPS conclusions. Redefine counting protocols if necessary. (Science Plan Topic addressed, 31)

d) To define periodicity of this type of training workshops (Science Plan Topics addressed, 27,31)

WKNEPS will report by 28 January 2019 for the attention of EOSG and SCICOM.

2 Introduction

The use of underwater television surveys (UWTV) to assess and advise on *Nephrops* stocks can be considered to be well established across many national institutes and is comprehensively documented (ICES 2018, Leocádio, *et al.*, 2018). Figure 2.1 shows the distribution of UWTV surveys and their status in providing data for stock assessment. These surveys fall under the planning and coordination of the working group on *Nephrops* surveys (WGNEPS) which meets annually (ICES, 2017b).

National institutes have been developing and building expertise in *Nephrops* burrow identification for their given survey areas for a number of years. The first burrow workshop (WKNEPHBID) in 2008 standardized the process of compiling reference sets and generating reference counts. The reference sets are used to test the reviewer's counting performance using Lin's CCC method (ICES, 2008, Lin, 1989) before being permitted to review the survey footage. At WKNEPHBID reference counts were generated from reference sets by consensus agreement for FU 6 (Farne Deeps) and FU 7 (Fladen) by three experienced reviewers with a weighting applied to the local expert. For FU 15 (western Irish Sea) an average of the three reviewers was used due to limited time and the high densities encountered. After this workshop national institutes then continued to compile reference sets and counts for their survey areas based on the recommendations from WKNEPHBID. The next burrow workshop (WKNEPS) generated annotated (time-stamped) consensus counts for FU 1 only, as this is only practical for very low density *Nephrops* grounds (ICES, 2017a). Other surveys presenting reference sets at this workshop could not complete the process due to high variability in the counts reflecting the challenging nature of the grounds (FU 30 for example).

Intersessional work carried out to generate reference counts for FU 14 showed high variability in the counts of the three international reviewers thus reflecting the difficulty in identifying *Nephrops* burrows in this area due to multiple burrowing species interactions (ICES, 2017b). WGNEPS 2017 then discussed the utility of applying Lin's CCC statistics as a method to develop reference counts. The working group then recommended to test this at the next burrow counting workshop in Aberdeen. This method would screen for intra-reviewer and inter-reviewer reproducibility using good quality image data. This would also be preferable to consensus counting where it would not be feasible to consensus agree each burrow in survey areas with moderate to high densities coupled with multiple other burrowing mega fauna interactions. This objective method could also reduce any potential effect of a "dominant" or "shy" counter in a consensus agreement scenario. Ultimately the aim was to test this method and to develop a common approach and methodology to calculate reference counts across the various institutes.

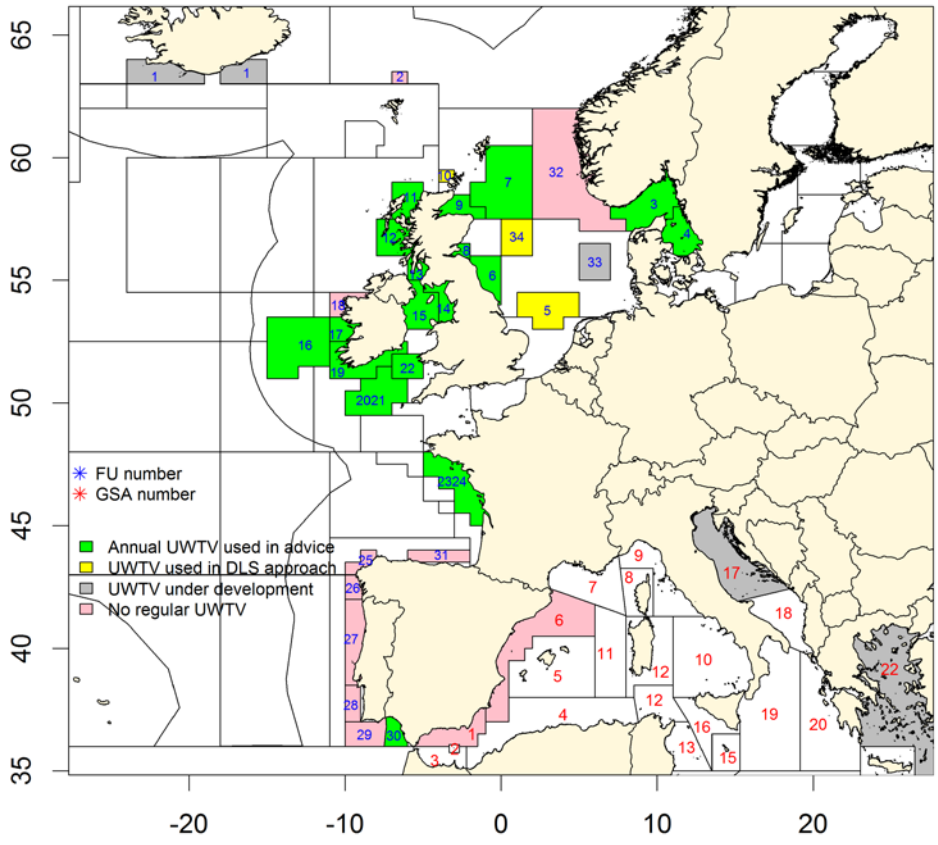


Figure. 2.1 *Nephrops* UWTV survey coverage in 2018 (FU: Functional Unit, GSA: Geographical Sub Area, DLS: Data Limited Stock).

3 Compilation of Reference Sets and work programme

The reference set for a given survey area will be comprised of nine stations covering the range of densities typically encountered in that area: that is, 3 x high density stations, 3 x moderate density stations and 3 x low density stations. The visual clarity must be "good" and clear, optimal speed (~0.7 - 0.8 knot) and good ground contact by the sledge. Hence the best quality footage selected for the reference set in order to test the reviewers effectively. The reference stations are then to be named 1 to 9.

In practice each underwater TV (UWTV) station is a 10 minute recording which means that minute one will be counted as a familiarisation or "warm-up" minute and where the count for this is also recorded. Count data from minutes 2 to 8 for each reference station will be used in the analysis. Each reference station will comprise of 7 minutes of count data to allow for robust analysis using Lin's CCC as this analysis requires a minimum of 5 observations from each of the two or more reviewers.

How to review the reference sets.

All reviewers are to count the footage in line with SISP (in prep) recommendations. These are:

- It is advised that a full review of one reference station to be undertaken first to obtain overall familiarity with footage.
- All reviews to be done in isolation.
- The first minute of the reference station is to be counted as a warm –up minute and the count recorded.
- Then count 7 minutes and to disregard any minute where 30 seconds or more not countable due to poor visual clarity or sledge stopped or gliding.
- If returning from a break of more than 1 hour then also to review one reference station as a warm-up.

Previously it has been recommended that reference counts for survey areas such as FU 1 and FU 16 should be annotated (time-stamped) and then consensus agreed where there is a difference of 1 or more burrows (ICES, 2017a).

Standard count data sheet template.

A standard worksheet to record each reviewers' counts was provided prior to the so that the data could be read into r-scripts compiled at the workshop workshop (Annex 3).

"r" scripts and code.

The "r" scripts for assessing the count data and generating the final reference counts were developed during the workshop and can be found on the ICES WGNEPS github site (https://github.com/ices-eg/wg_WGNEPS).

Work programme.

The mix of image data formats, file size and reviewing equipment being used on UWTV surveys across the various institutes poses many logistical challenges when holding a workshop. Standard definition format requires specialised cathode ray tube monitors (CRT) for review, while the high definition format file size may be too large to send electronically (approximately 4 GB per station). To ensure maximum efficiency for this workshop national and international reviews were completed prior to the workshop with some exceptions (FU 1, FU 6 and FU 30), due to specialised screens

requirements and file formats and size. The international reviews for these were completed during the workshop. For FU 23-24 (Bay of Biscay) five stations were reviewed during the workshop due to data quality issues with the reference set.

It was planned that a consensus review of a random selection of reference stations would be undertaken at the workshop and the results compared to the Lin's directed reference counts. However, this was not possible due to the time spent in group review sessions of footage from many surveys areas. This was deemed important so that participants would gain experience in identifying Nephrops burrows from the range of survey areas as the workshop was a mix of experienced and novice reviewers.

The next section describes in detail how each institute compiled the reference set.

High definition format.

3.1 Iceland (FU 1)

The third UWTV survey was carried out in FU1 between 11 and 19 of June 2018 following the same procedures from other areas conducting similar surveys. A reference footage set was generated from this survey as there was 100% coverage of the grounds. The visual clarity ranged from medium to good and the speed was usually around 1 nm/hour. 9 stations each 8 min. distributed on most of the discrete grounds where chosen as the reference set (Figure 3.1). There are few known burrowing megafauna of concern in this region other than the crab Geryon Trispinosus.

The footage was shot with HD camera (Kongsberg OE14-502F) and video files are saved as MPEG and are around 1.6 GB in size. Each video was timestamped, with information on geographical coordinates, speed, depth, distance covered by wheel and how much cable is out.

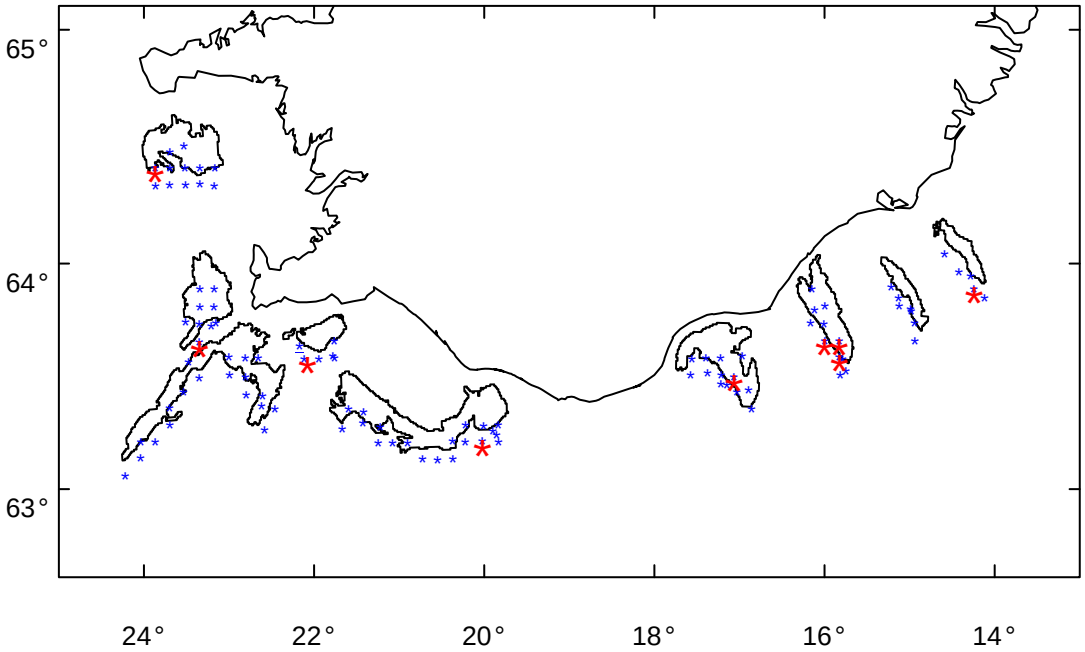


Figure 3.1. Overview of the UWTV stations (black asterisk) from the 2018 survey and the reference footage stations (red asterisk) in FU 1. The polygons depict the *Nephrops* ground and are based on cumulative VMS data.

3.2 Skaggerak and Kattegat (FU 3-4)

Reference footage for FU 3-4 was not available at this workshop. Survey footage was available and this was used to train and develop counting skills for the national reviewers present (See section 5).

3.3 Farne Deeps (FU 6)

Cefas has performed annual UWTV surveys in Farn Deeps since 1997. The survey design consists of a randomized fixed grid of 110 stations where at each station a sledge mounted TV camera is deployed, and a clear 10-minute tow is recorded. Before the survey, all counters have to pass the Lin's CCC test using a reference footage created in 2010. However, the quality of the videos increased significantly in 2016 due to an upgrade in the UWTV gear (Cefas bought new monitors, camera and cable, and onboard control system), and a new reference footage is now needed.

For the workshop, a set of 9 videos x 7-minute footage segments recorded in MP4 were selected from the 2018 survey. The segments were representative of the ground, and they included areas with different density and level of complexity for burrow identification (i.e. presence of small Nephrops burrows, or interactions with other burrowing species). The quality of the footage was good in terms of speed, ground contact, and visibility.

3.4 Gulf of Cadiz (FU 30)

The reference video set has been generated from footage obtained during the 2018 TV survey. This year, a new 4K Ultra HD camera (SONY Handycam FDRAX33) was used and video files were saved as MP4 which are around 4.0 GB in size. Reference footage was chosen taking into account the quality in terms of ground contact, sledge speed and visual clarity. The reference footage covered a range of *Nephrops* densities (low, moderate and high) in relation to the mean density of this FU. In FU 30, density is considered low (mean density 0.12 burrows/m^2 in 2018) compared with other Atlantic FUs, so the maximum number of *Nephrops* burrows systems per minute was lower than 15. In this area, *Nephrops* inhabit with other burrowing megafauna, mainly the squad lobster *Munida spp*. and the crabs *Goneplax rhomboides* or *Monadaeus couchii* among others. This fact means in some areas the bioturbation is very high and, as consequence, the complexity in the identification of *Nephrops* burrows increases. Based on expert knowledge, the reference footage chosen also covered a range of species interaction.

Standard definition format.

3.5 North Minch (FU 11)

In 2018 Marine Scotland Science (MSS) updated the *Nephrops* burrow identification training reference set for the North Minch (FU11) in accordance with the recommendations made at WGNEPS 2017. The footage was first recorded on to the hard drive of a Panasonic DMR-EX86EB in MPEG2 format and then burned onto Verbatim 4.7GB -R DVDs, with no more than four clips per disk. This new collection of nine video clips used footage from the 2018 survey, which was carried out on board MRV Scotia, provided footage using only the most recent electronic equipment in operation by MSS (Konesberg OE-14366 video camera and SeaLED lights). Each of the analogue video clips presented a true representation of the wide ranging benthic conditions encountered in this area; from soft muddy grounds to noticeably coarse regions, where stones and occasionally rocks could be observed. This variety of sediments affects the fauna present in the North Minch, and this too was clearly visible in the excerpts that were

selected, whereby in addition to examples of wide ranging densities of *Nephrops* burrow complexes, some clips illustrated areas containing high densities of other burrowing megafauna, with other areas where sea pens were more frequently observed, and on the coarser grounds where squat lobsters (*Munida* sp.) were present. This selection of footage resulted in presenting the full range of benthic features and biological inhabitants typically found during recent *Nephrops* UWTV surveys in the North Minch.

3.6 Western Irish Sea (FU15)

The FU 15 UWTV survey is an established survey series that commenced in 2003 with a consistent protocol. It is an annual survey occurring in late July to early August of each year whereby 100 stations are surveyed on a fixed isometric grid with 4.5 nm spacing with a randomized starting point are surveyed. Survey footage (10 – 15 mins) from the selected nine stations was directly burned onto a DVD using a Pioneer HDD/DVD recorder (DVR-550HX), and then burned onto a standard definition DVD.

In accordance with the recommendations made at WGNEPS 2017, nine stations (Figure 3.6) were selected as reference footage by the scientist in charge from the 2016 and 2017 surveys. The selection of stations was based on expert opinion to ensure the footage was representative of the range in burrow density, visual clarity, environmental conditions and presence non-target burrowing species experienced during the survey series. The *Nephrops* habitat in FU 15 has comparatively higher burrow densities compared with other Functional Units and a significant presence of non-target burrowing species (e.g. *Calocaris macandreae, Goneplax rhomboides*) which provides challenges to reviewers when identifying *Nephrop's* burrow complexes.

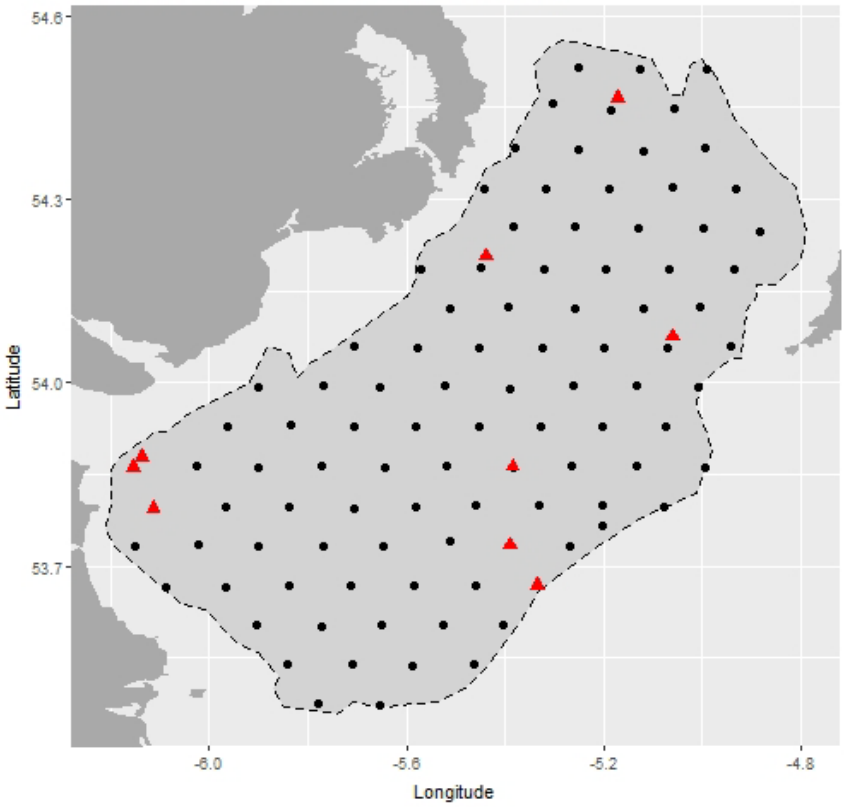


Figure 3.6 Overview of the UWTV stations (black circles) from the 2016 survey and the reference footage stations (red triangles) selected from 2016 and 2017 stations in FU 15 (note each year's stations are randomly offset). Dashed line polygon depicts the *Nephrops* FU 15 survey ground extent.

3.7 Labadie, Jones and Cockburn Banks (FU 20-21)

It was decided to use only stations from the last three years (2016, 2017 and 2018 surveys) in order to have high quality data in terms of ground contact, sledge speed and visual clarity. From each of the years, the stations with the highest counts were preselected, together with a random sample of 20% of the stations with more than 15 burrow systems. This first process resulted in a total of 16 stations from each year as possible candidate stations for the reference set. Based on expert knowledge of the ground, stations were then chosen to cover the range of species interactions and small scale topography relief, where on this ground there are a lot of undulations on the seabed. It was also deemed necessary to include one station with recently trawled ground to test the reviewers. After a precursory review in the laboratory the final nine stations were selected with two stations from the 2016 survey, three from 2017 and four from 2018.

The selected nine stations were saved into a Pioneer HDD/DVD recorder (DVR-550HX), and then burned onto a DVD.

3.8 Bay of Biscay (FU 23-24)

The fifth UWTV survey in Bay of Biscay (FU23-24) was carried out between 19th April and 2nd May 2018 onboard RV Celtic Voyager. Until now, FU22 (Smalls area) reference footage has been used for training and to validate the French team to recognize *Nephrops* burrows, following survey procedures from other areas.

In 2018, following the workshop recommendations, footage was selected from those acquired during the 2018 survey, trying to select stations with the best quality in terms of ground contact and visual clarity. Three range of densities were chosen : high, medium and low density. To compile the reference set onto physical DVD the Imovie software was used.

4 Generating Reference Counts

This section reports on the process involved in using Lins' CCC method to develop the reference counts for the FUs presented at the workshop with the exception of FU 1 which is annotated (time-stamped).

Lin's Concordance Correlation Coefficient (CCC) (Lin, 1989) measures the ability of counters to exactly reproduce each others counts on a scale of 1 to –1 where 1 is perfect concordance (i.e. a pairwise plot will have all points lying along the 1:1 line). The first step is to test the CCC values when comparing counts within an individual termed the intra performance check , that is, can the reviewer reproduce the same counts. A threshold of 0.5 was set for this step where a reviewer's count data are dismissed if they do not pass this step. Step two is to test the CCC values when comparing the mean counts of each reviewer also termed the inter performance check. A threshold of 0.5 was used along with a rule where if a reviewer fails more than 50% of its pairings then they are dismissed from the process. The reference counts were calculated as the mean of all the remaining reviewers with no weighting applied. It was agreed at the workshop that it was acceptable to calculate reference counts in cases when only two reviewers passed step 1 and 2. This is similar to the process for verification of survey count data. Step three is finally to test the robustness of the reference set using all the reviewer's counts.

The decision tree shown in Figure 4 documents the steps followed at the workshop.

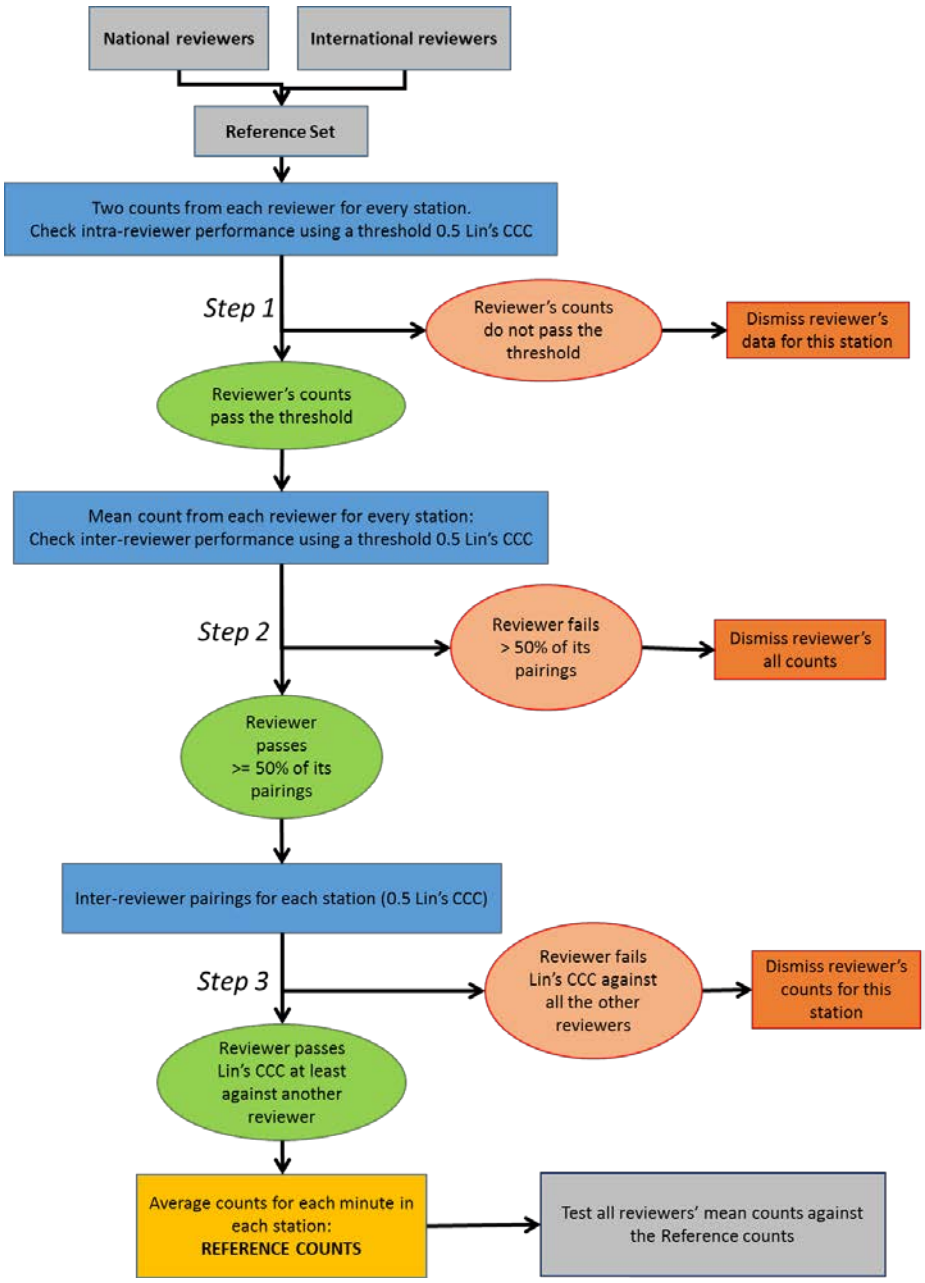


Figure 4. Decision tree in using Lin's CCC statistics to screen the reference count data for intra-reviewer (step1) and inter-review consistency (step 2) and robustness of reference counts (step 3).

High definition format.

4.1 Iceland (FU 1)

Snapshot examples of the reference footage are shown in Figures 4.1.1. The counts for the area were less than 11 systems per minute, and the reference set had on average 4 systems per minute (range: 1.2–7.0). To generate the reference counts, two national counters and one international counter reviewed the footage. The reference set was timestamped by individual counters independently. Once the counts were completed the results compared and footage was reviewed in a group and counts agreed on if there were any discrepancies.

Table 4.1.1 shows the time-stamps for a reference station from FU 1. The reference counts were then finalized for this area at the workshop.

Table 4.1.1. Example from one reference stations in FU 1 showing timestamped counts.

	A	B	C	D	E	F	G
98	11						
99	Station	602	Time:1-8				
100							
101							
102	ID:						
103	Min	Burrow Count	Time Stamp				
104	1	5	14L, 44L, 48L, 49R, 53L				
105	2	2	19R, 50L				
106	3	2	0L, 41L				
107	4	5	10C, 16c, 22l, 23c, 38r				
108	5	1	37R				10C FALLIN?
109	6	5	2C, 18L, 20L, 42C, 56R				
110	7	7	19L, 31R, 38L, 45C, 47L, 51L, 54C				
111	8	4	3R, 16L, 33R, 37L				
112	9	31					
113	10						
114	11						
115							
116	Station	603	Time:1-8				
117							
118	ID:						
119	Min	Burrow Count	Time Stamp				
120	1	5	6C-R, 15C, 28L-R, 26C, 41R				
121	2	7	10L, 19R, 28L, 36C-37 STÓRT, 40L-R, 43R-L, 58R				55L SKRÍTIÐ, 56 FALLIÐ
122	3	9	16C, 18C, 38L, 43C, 44R, 48R, 52R, 54L, 57L				
123	4	6	14L, 24L, 37C-L, 40C, 52R, 58R-L				
124	5	6	3C, 15R, 28C, 35L-R, 46C, 55L				
125	6	8	15L, 18C, 24R, 27C, 29L, 38C, 51C, 58R-L				
126	7	7	10R, 17L, 32L, 42L, 46C, 56R-L, 58L				
127	8	7	0L, 2C, 9L, 39R, 42L, 53R, 56R				
128	9	55					

Recount_FU1

Figure 4.1.1. Snapshots from FU 1 ground reference footage with *Nephrops* burrows visible. The distance between the lasers (green line) is 0.95 meter.

4.2 Farne Deeps (FU 6)

The new reference footage was independently counted twice by four national and one international reviewer. In all cases the counts were made using the OLED monitors (Sony 25-inch professional PVM-A250) used during the surveys, in order to replicate the quality of the videos used for the real counts.

The procedure to create the new reference counts was as detailed in the flow chart above (Figure 4). First, an intra-counter Lin's CCC test was conducted to identify inconsistencies between the two counts made per person in each station. Those cases where the result was lower than 0.5 were eliminated of the analysis. For the other cases, an average of the two counts per minute was estimated. Then, the Lin's CCC test was conducted again to identify inconsistencies between counters in each station. The reference counts were finally estimated as an average of the counts that passed the threshold of 0.5 in the test.

Figure 4.2.1 shows the average count of those counters that passed the intra-counter Lin's CCC test in each station (91% of the total counts). Overall, the variation among counts was small, being the maximum difference 8 burrows·min^{-1} in stations 7 and 8. The time-trends were also very similar between counters in most of the stations, and there was not a different pattern between international and national counters.

At least two pairs of counts passed the inter-counter Lin's CCC test in each station (Table 4.2.1) and they were used to estimate the reference counts. The lowest level of agreement was found in station 6, where the burrow density is low and it is easier to fail the test.

The reference counts are shown in Figure 4.2.2 together with the counts used for their estimation.

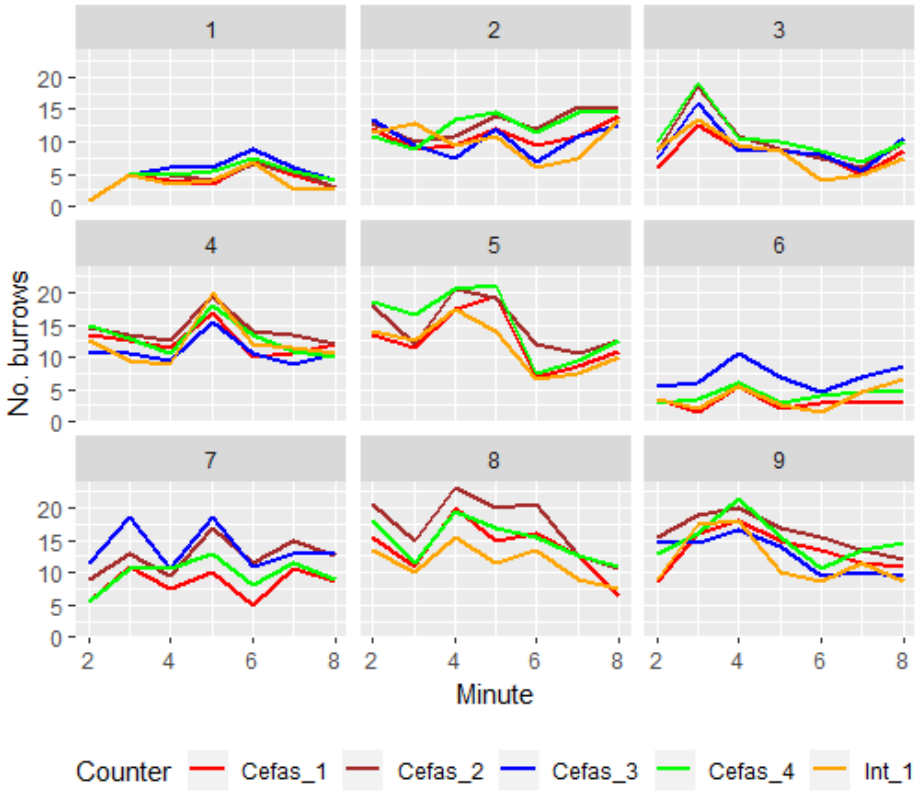


Figure 4.2.1. Burrow counts in each station after passing the intra-count Lin's CCC test.

Table 4.2.1. Inter-counter Lin's CCC test results. In bold are the pairs that passed the test (CCC > 0.5) and were used to estimate the reference counts. The empty pairs not passed the Intra-counter Lin's CCC and were not used in this analysis.

Station 1					
	Cefas_1	Cefas_2	Cefas_3	Cefas_4	Int_1
Cefas_1		**0.96**	**0.77**	**0.87**	**0.85**
Cefas_2			**0.81**	**0.91**	**0.75**
Cefas_3				**0.94**	**0.65**
Cefas_4					**0.74**

Station 2					
	Cefas_1	Cefas_2	Cefas_3	Cefas_4	Int_1
Cefas_1		**0.52**	**0.75**	0.42	0.40
Cefas_2			0.38	**0.76**	-0.04
Cefas_3				0.16	**0.58**
Cefas_4					-0.10

Station 3					
	Cefas_1	Cefas_2	Cefas_3	Cefas_4	Int_1
Cefas_1		**0.70**	**0.83**	**0.61**	**0.71**
Cefas_2			**0.92**	**0.97**	**0.76**
Cefas_3				**0.87**	**0.75**
Cefas_4					**0.69**

Station 4					
	Cefas_1	Cefas_2	Cefas_3	Cefas_4	Int_1
Cefas_1		**0.65**	**0.74**	**0.75**	**0.72**
Cefas_2			0.46	**0.83**	**0.73**
Cefas_3				**0.62**	**0.74**
Cefas_4					**0.80**

Station 5					
	Cefas_1	Cefas_2	Cefas_3	Cefas_4	Int_1
Cefas_1		**0.78**		**0.83**	**0.85**
Cefas_2				**0.83**	**0.66**
Cefas_3					
Cefas_4					**0.72**

Station 6					
	Cefas_1	Cefas_2	Cefas_3	Cefas_4	Int_1
Cefas_1			0.15	**0.55**	0.49
Cefas_2					
Cefas_3				0.24	0.33
Cefas_4					**0.54**

Station 7					
	Cefas_1	Cefas_2	Cefas_3	Cefas_4	Int_1
Cefas_1		0.32	0.26	**0.68**	
Cefas_2			**0.66**	**0.51**	
Cefas_3				0.30	
Cefas_4					

Station 8					
	Cefas_1	Cefas_2	Cefas_3	Cefas_4	Int_1
Cefas_1		**0.70**		**0.84**	**0.73**
Cefas_2				**0.75**	0.40
Cefas_3					
Cefas_4					**0.55**

Station 9					
	Cefas_1	Cefas_2	Cefas_3	Cefas_4	Int_1
Cefas_1		**0.57**	0.43	**0.63**	**0.69**
Cefas_2			0.50	**0.61**	0.44
Cefas_3				**0.56**	**0.59**
Cefas_4					**0.59**

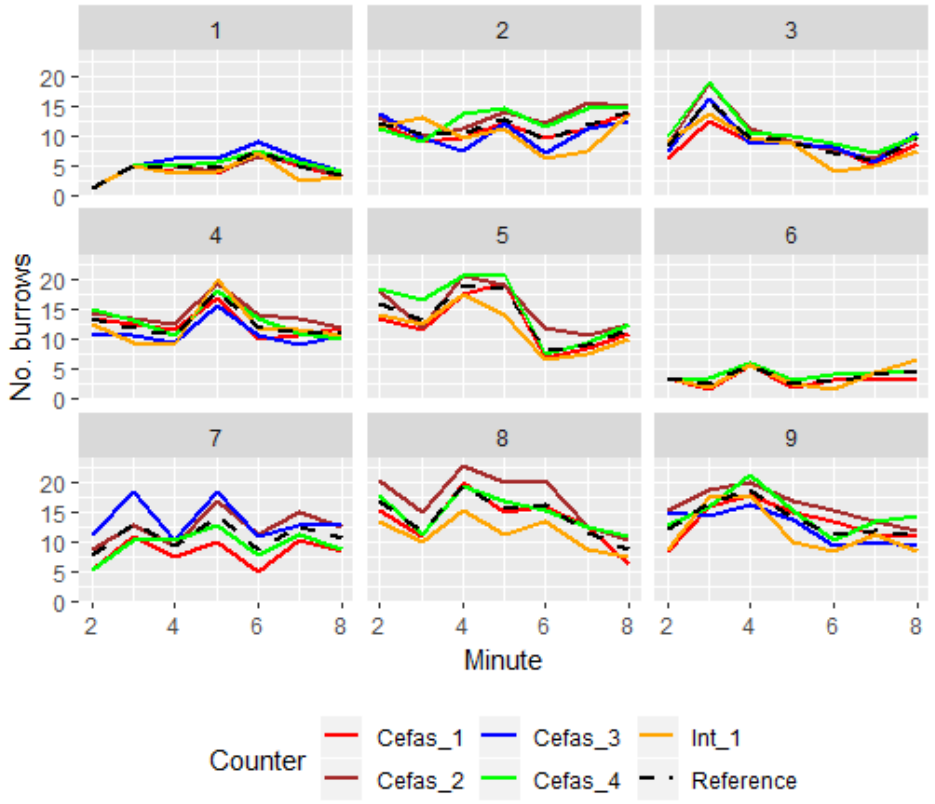


Figure 4.2.2. Burrow counts used to estimate the reference counts. The reference counts are shown in dashed line.

4.3 Gulf of Cadiz (FU 30)

To generate the reference counts, three national reviewers (C1, C2 and C3) and two international reviewers (IC1 and IC2), counted twice and independently the nine stations chosen according to workshop recommendations. Footage was time-stamped only by the national reviewers. Due to the large file size (4GB) which hindered sending the videos, the international reviews were done during the Workshop and there was not the time to make time stamping annotations. On the other hand, the use of Lin's CCC to assess the reviewer performance and to generate reference counts independently wanted to be tested.

Decision tree according to Figure 4 was used to generate the reference footage process for this survey area. Lin's CCC test was used to check intra-reviewer consistency and inter-reviewer differences. R-script was developed and supplied during the Workshop for this purpose. A threshold of 0.5 was agreed to check both intra and inter reviewer consistency. When a reviewer did not have consistency with the own counts (Lin's CCC lower than 0.5) for any station, their counts were dismissed for the station (Figure 4, step 1). Thus, the IC2 reviewer was dismissed in 4 stations (RF01, RF03, RF04 and RF05), the IC1 reviewer in 2 (RF05, RF06) and the C1 reviewer in 1 station (RF06) of the total nine stations reviewed (Figure 4.3.1).

Mean counts per minute from each reviewer were calculated, and inter-reviewer correlations were also checked using Lin's CCC and a threshold of 0.5. When the inter-reviewer pairings failed more than 50%, all counts of this reviewer were dismissed and were not taken into account in the process (Figure 4, step 2). So, counts for IC2 were dismissed for the process to generate reference counts for this FU (see Figure 4.3.2).

Figure 4.3.3 shows the Lin's CCC values for each of the reviewer's pairings without IC2.

According to the step 3 of the decision tree (Figure 4), only counts that passed at least 50% of the time against another reviewer were used in order to calculate the reference counts set. Green cells in Figure 4.3.2 correspond to the counts used. Only the national counts of C2 and C3 were used for generate the reference count for the stations 5 and 6. This is acceptable by the WKNEPS, as it is the current quality control method of survey counts.

The average of the counts per minute of all the reviewers who passed the previous steps were used to generate the reference counts for FU 30.

Testing the Reference counts.

Lin's CCC values for each of the reviewer's mean counts were also tested against the reference counts. All reviewers showed Lin's CCC values higher than the threshold of 0.5 (Figure 4.3.4). Reviewer IC2 was dismissed when the reference counts were generated such is explained previously. However, this reviewer passed the Lin's CCC test when their mean counts were tested against the reference counts although the IC2 spread data were very high. The reason could be that IC2 counts are within the range of the other reviewers' and for the step 1 and 2 of the decision tree they are not in the extreme cutoff for exclusion. Probably, a major training of this reviewer would improve their identification skills.

Recommendations

National counts to include time-stamped annotation for each station selected for the Reference set but only from minutes 2 to 8. It's recommended to annotate the whole station (10 minutes) for training purposes.

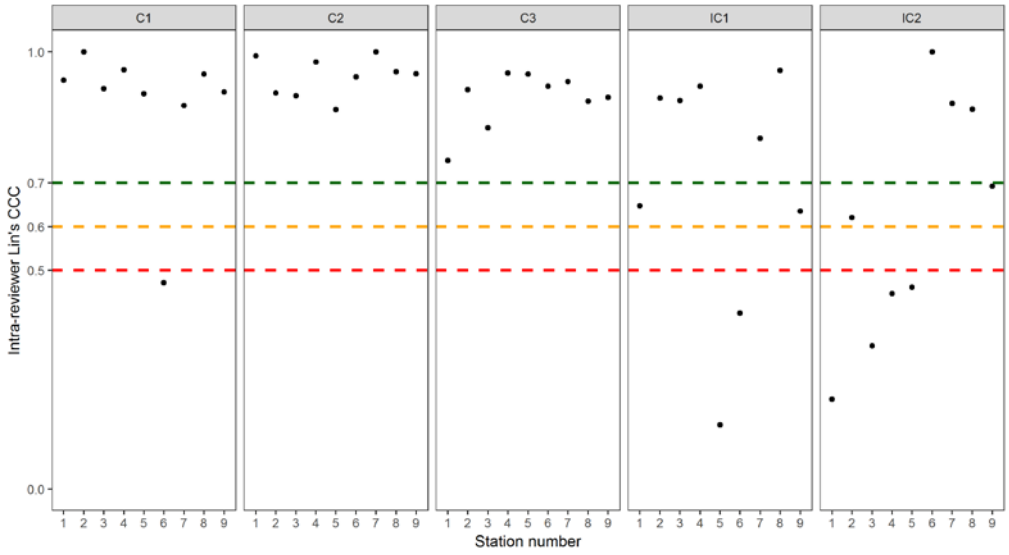


Figure 4.3.1. Intra-reviewer Lin's CCC value for each reviewer. Red value correspond to Lin's CCC lower than the threshold of 0.5.

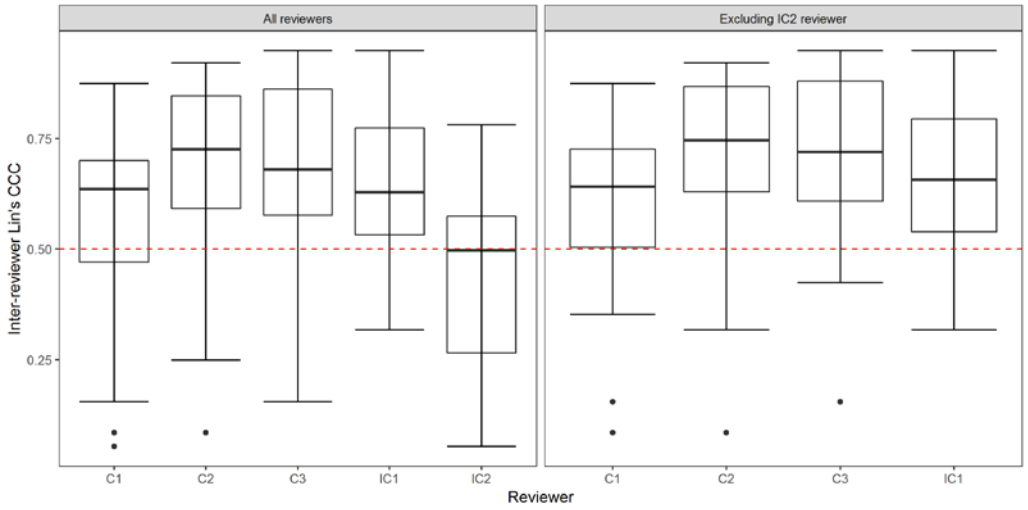


Figure 4.3.2. Step 2. Inter-reviewer Lin's CCC performance check. Boxplots show the distribution of Lin's CCC values for each of the reviewer's pairings. On the left panel, first check using mean counts of all the reviewers: note that the median (black band inside the box) of the IC2 reviewer is just below the 0.5 threshold (red-dashed line). On the right panel, second check excluding the counts of the IC2 reviewer.

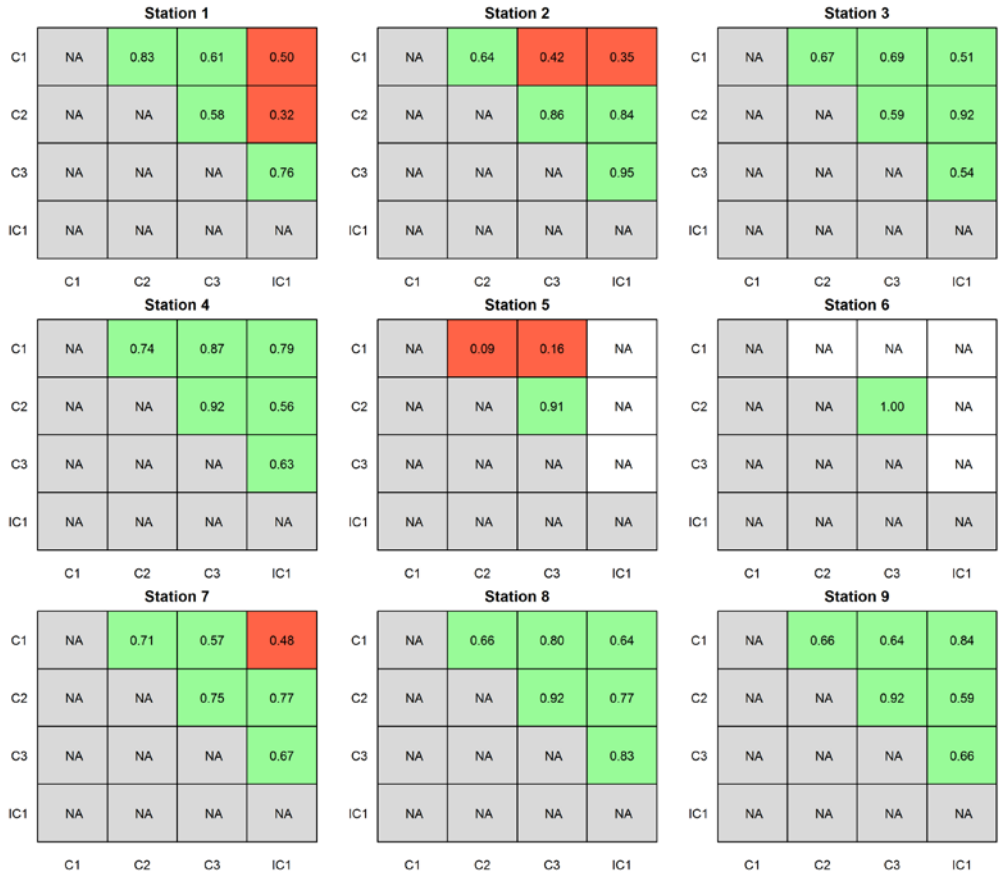


Figure 4.3.3: Step 3. Inter-reviewer Lin's CCC performance for each station. Pairings which passed the 0.5 threshold (in green) where later used for generating the reference counts. Pairings which did not pass the 0.5 threshold are shown in red. Counts which were dismissed in Step 1 are shown in white cells. International reviewer's pairings (IC2) are not shown, as all its counts were dismissed in Step 2.

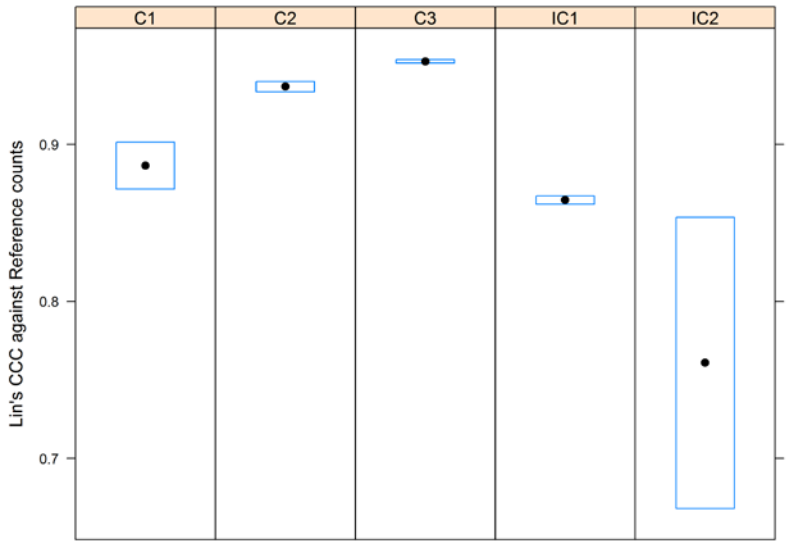


Figure 4.3.4. Lin's CCC values distribution for each of the reviewer's mean counts tested against the reference counts.

Standard definition format.

4.4 North Minch (FU 11)

To generate burrow abundance values from the 2018 North Minch reference set, four members of staff from Marine Scotland Science (MSS1-4) and one international counter (INT1) were invited to independently review each video clip twice, as described in the *Nephrops* UWTV survey SISP. All reviewers had previous experience of reviewing similar footage. Results were recorded onto a standardized Microsoft Excel template (Annex 3). Initially the results from all five reviewers were analysed by Lin's CCC (using a script written in R and created by C. Mesquita from MSS and is available on the GitHub website at https://github.com/ices-eg/wg_WGNEPS), where each reviewer's data were compared against themselves. Those sets that failed to pass the 0.5 threshold were then dismissed from any further analysis, resulting in eight of the 45 datasets being rejected (intra-reviewer analysis, Figure 4.4.1). As described in the Decision Tree (Figure 4) the mean counts from each of the remaining burrow counts for each video clip were then calculated, and applying Lin's CCC (using an amended version of the earlier R script modified by E. Bell from CEFAS), the mean counts from each reviewer were compared against each other (inter-reviewer analysis), and those reviewers where their datasets fell below the threshold of 0.5 had all their data dismissed from future analysis. This step in the script had a negative affect on the international reviewer's data, due to the differences in the way the equipment on the sledge was arranged between the two institutes, and the unfamiliar heterogenic nature of the footage used in the reference set.

A fundamental criterion for the analysis to progress was that at least two reviewers passed the 0.5 threshold for each reference set, which was achieved for all the complete FU11 reference set –although only the minimum requirement was met for station five. As expected the most frequent number of disagreements was observed in stations with low densities such as station 1, (Table 4.4.1).

Further analysis using the R script incorporating Lin's CCC provided a plot of the mean of the reviewers' values for each reference set (Figure 4.4.2). These results were achieved using a variety of reviewer combinations having been generated using the methodologies agreed in the SISP and processed according to the logic flow plan as described in the Decision Tree (Figure 4). These values were then tabulated using a modified version of the R script to produce final and statistically agreed values for each station.

These agreed reference values were then applied to the original, first set of values recorded by each of the reviewers, to observe how the results compared. In producing the agreed reference values using the earlier R scripts, some trends were observed where there was a suggestion that some reviewer's data may display outliers and find meeting the 0.5 threshold difficult. This final analysis provided a clear indication where further training and support was required, and there by fully meeting the purpose of this exercise.

Recommendations.

It was agreed at WKNEPS the first two minutes from each video clip for each reference set should be annotated. This can be achieved by digitally annotating footage when high definition footage is available, or in the case of the North Minch (FU11) where analogue footage is recorded, by 'time stamping' the selected video.

International reviewers with similar survey grounds and sledge set-up should be involved in generating reference set count data.

MSS annually surveys seven functional units for assessment purposes, and new reference footage using video collected on the 2018 UWTV surveys should be created for six of the seven areas (Devils Hole (FU34) and Fladen (FU7) have similar features so one reference set would be acceptable for both areas). Reference count data should then be generated for each reference set using the approach discussed and approved by WKNEPS 2018.

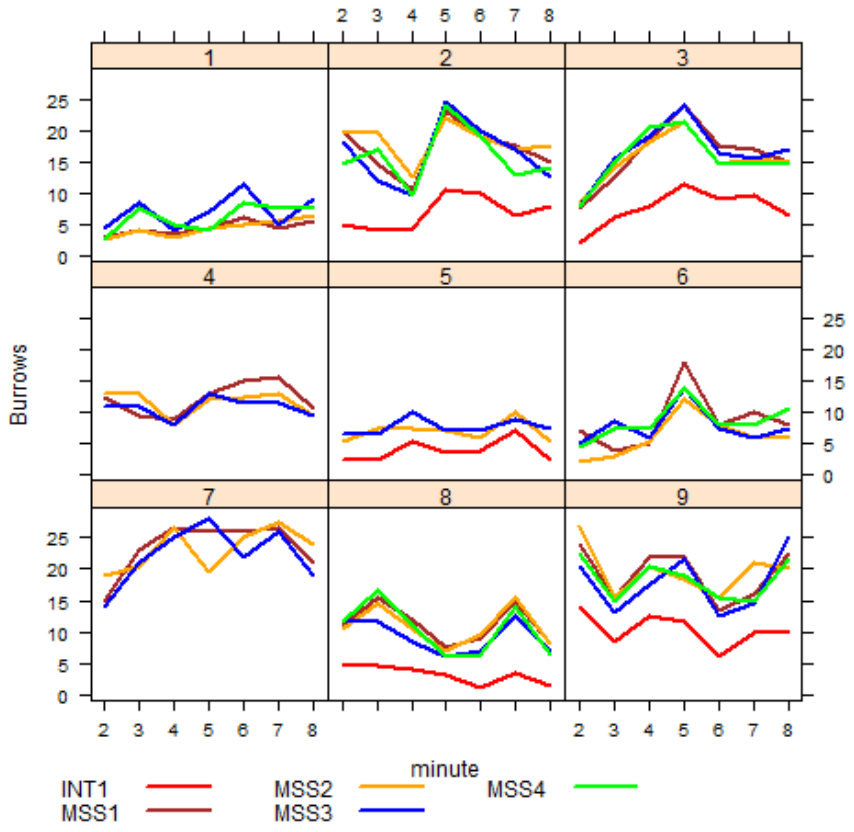


Figure 4.4.1. Intra-reviewer plots generated by the Lin's CCC analysis in the WKNEPS R script, showing burrow density by minute for each station by each reviewer that passed the 0.5 threshold when applied to their own data.

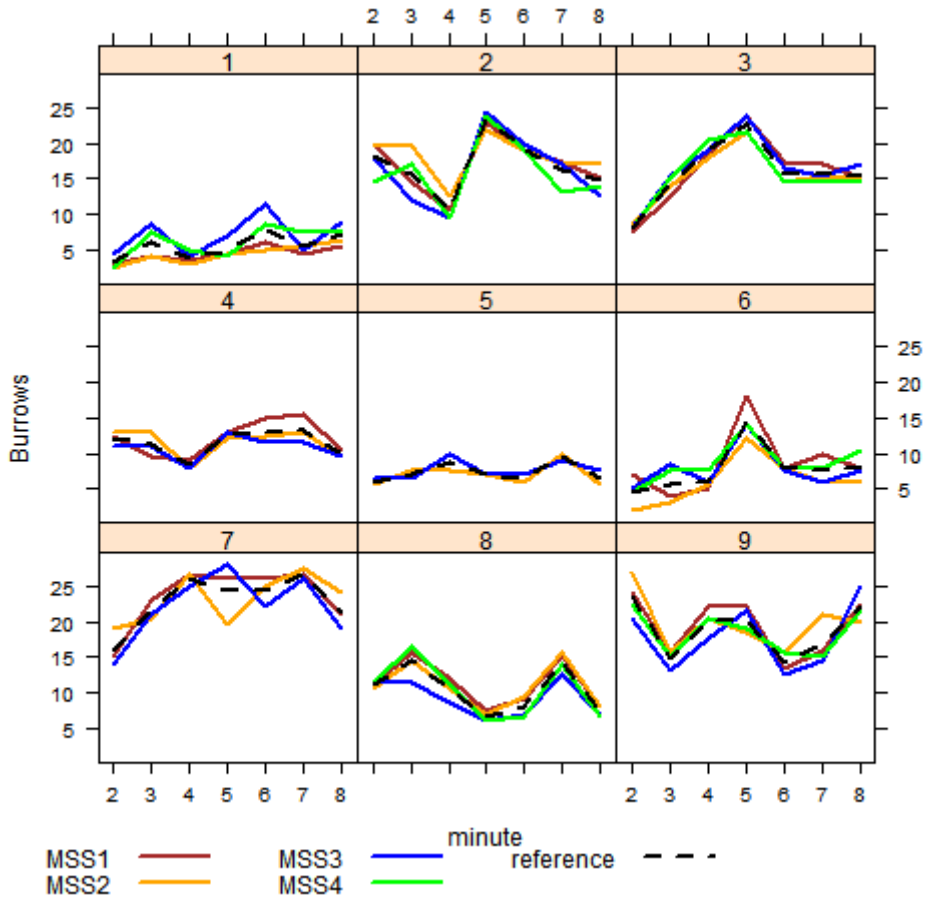


Figure 4.4.2. Inter-reviewer plots generated by the Lin's CCC analysis in the WKNEPS R script, showing burrow density and the agreed reference values, by minute for each station by each reviewer that passed the 0.5 threshold when their data were applied to data from all the other remaining reviewers.

Table 4.4.1. Inter-reviewer correlation results using Lin's CCC by each station for the North Minch (FU11), with acceptable results highlighted in green and those that failed to reach the 0.5 threshold highlighted in red. The international reviewer's data (INT1) is excluded from the output as this failed to pass the 0.5 threshold at an earlier stage in the process.

Reference Set 1

	MSS1	MSS2	MSS3	MSS4
MSS1	NA	0.810078	0.313467	0.40607
MSS2	NA	NA	0.262455	0.465765
MSS3	NA	NA	NA	0.615519
MSS4	NA	NA	NA	NA

Reference Set 2

	MSS1	MSS2	MSS3	MSS4
MSS1	NA	0.785377	0.924818	0.759111
MSS2	NA	NA	0.59687	0.670532
MSS3	NA	NA	NA	0.808417
MSS4	NA	NA	NA	NA

Reference Set 3

	MSS1	MSS2	MSS3	MSS4
MSS1	NA	0.920115	0.944538	0.886121
MSS2	NA	NA	0.932263	0.959554
MSS3	NA	NA	NA	0.927894
MSS4	NA	NA	NA	NA

Reference Set 4

	MSS1	MSS2	MSS3	MSS4
MSS1	NA	0.580899	0.503923	NA
MSS2	NA	NA	0.731638	NA
MSS3	NA	NA	NA	NA
MSS4	NA	NA	NA	NA

Reference Set 5

	MSS1	MSS2	MSS3	MSS4
MSS1	NA	NA	NA	NA
MSS2	NA	NA	0.511703	NA
MSS3	NA	NA	NA	NA
MSS4	NA	NA	NA	NA

Reference Set 6

	MSS1	MSS2	MSS3	MSS4
MSS1	NA	0.67018	0.690919	0.71734
MSS2	NA	NA	0.684014	0.662224
MSS3	NA	NA	NA	0.853135
MSS4	NA	NA	NA	NA

Reference Set 7

	MSS1	MSS2	MSS3	MSS4
MSS1	NA	0.584612	0.882488	NA
MSS2	NA	NA	0.365346	NA
MSS3	NA	NA	NA	NA
MSS4	NA	NA	NA	NA

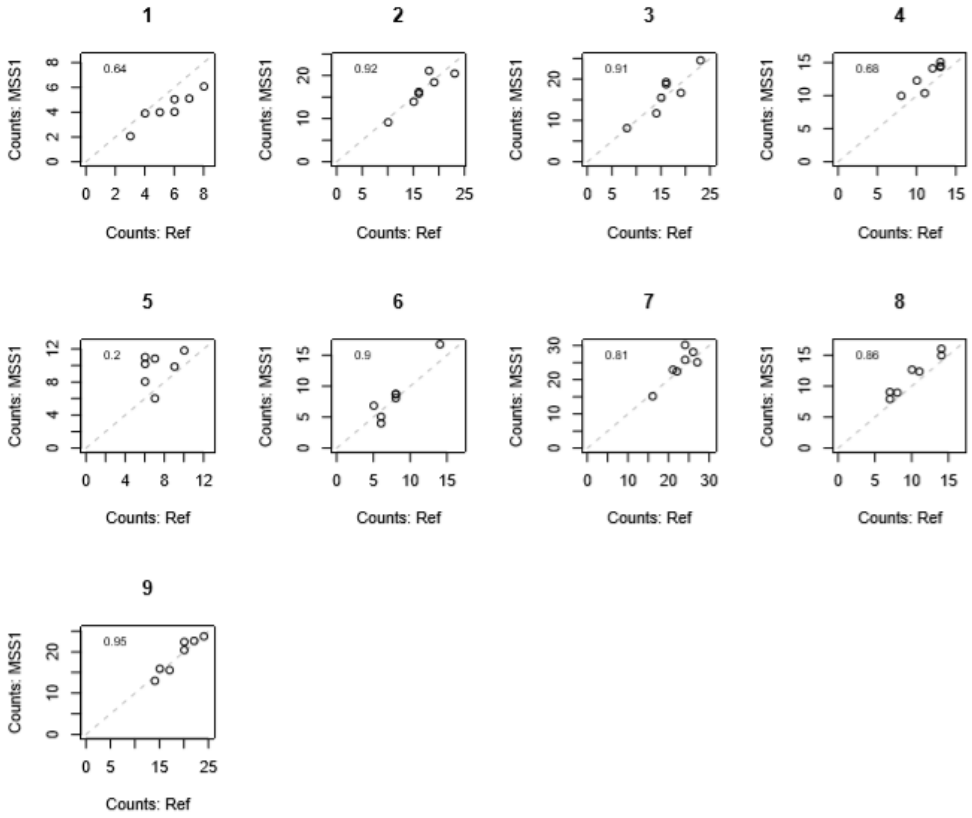
Reference Set 8

	MSS1	MSS2	MSS3	MSS4
MSS1	NA	0.966498	0.703621	0.921131
MSS2	NA	NA	0.758	0.880263
MSS3	NA	NA	NA	0.78366
MSS4	NA	NA	NA	NA

Reference Set 9

	MSS1	MSS2	MSS3	MSS4
MSS1	NA	0.70549	0.819913	0.89129
MSS2	NA	NA	0.44315	0.652572
MSS3	NA	NA	NA	0.778452
MSS4	NA	NA	NA	NA

Table 4.4.2. Sample ouput where the agreed reference count values for the North Minch (FU11) were applied to reviewer MSS1's original, first run of dat,a and then assessed using Lin's CCC to illustrate the correlation between the two datasets.



4.5 Western Irish Sea (FU 15)

The footage from each of these nine survey stations was reviewed by 5 trained AFBI staff and an international counter (labelled as A-F to anonymise identity). Each surveyor reviewed 8 minutes of footage three times independently whereby the 1st minute and 1st count of a station was discounted from analysis.

A summary of the reviewers counts of the reference footage is provided below (Figure 4.5.1), whereby the sum of burrows per minute was calculated from each surveyor's count of reference footage and the mean of the second and third counts plotted for each surveyor (A-F) and station.

At each station although the absolute number of burrows identified were similar trends in burrow densities whereby relative peaks and troughs of burrow densities were observed by all counters. This highlights the variability that occurs in the identification of *Nephrops* burrow complexes whereby the current set of surveyors had a wide range in experience of counting *Nephrops* burrows.

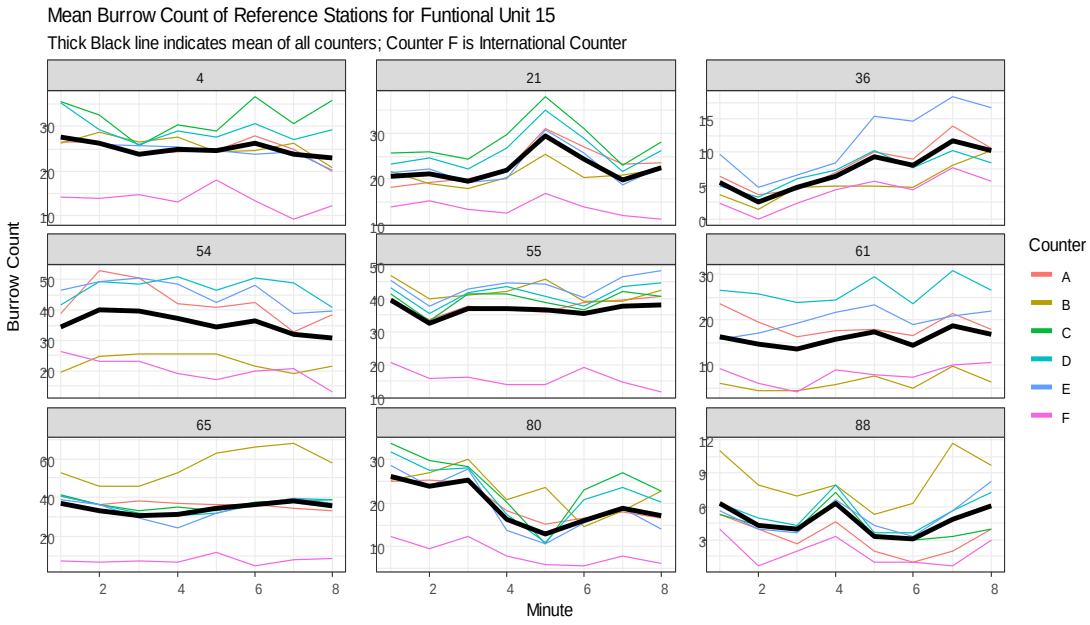


Figure 4.5.1. Mean count by reviewer and station for FU15 reference set.

Intra-Counter Consistency (Step 1)

This initial step (Step 1) for developing reference counts is to ensure surveyors have a minimum intra-counter consistent when evaluated the same footage. Lin's Concordance Correlation Coefficient (CCC) (Lin 1989) technique was applied to pairwise comparisons of multiple counts of the same footage by an individual surveyor.

A comparison of burrow counts was carried out for each counter at each station between the second and the third counts and the Lin CCC statistic calculated. A threshold CCC value of 0.5 was agreed at the workshop as the minimum value which observations are suitable to develop reference counts. The results are reported below (Figure 4.5.2).

Figure 4.5.2. Step 1. Intra-reviewer Lin's CCC performance check for each of the reviewers (in panels) and each of the stations (x-axis). Value within bar displays CCC for pairwise intra-counter comparison. Green bar represents CCC value at >0.7 threshold; orange bar represents CCC value at >=0.5; red bar represents CCC value at <0.5.

Those paired counts of an individual counter for a station which are below the 0.5 threshold (marked in red) were discounted from further analysis. Only counters B and D obtained the CCC threshold value of 0.5 for all of the stations reflecting the high burrow density, poor visibility and burrows from other species that occur in FU15. Stations 36, 54, 80 and 88 had all counters pass the intra-counter CCC threshold, indicating that there is variability in the ability to identify burrows between stations.

The arithmetic mean burrow count per minute for all stations using only counters who passed the CCC threshold in stage 1 was calculated and utilized as the burrow count for that counter and station.

A pairwise comparison between these mean counts from each counter for the respective station was applied. For a surveyor's counts to be suitable for reference counts they must have passed at least 50 % of their comparisons at the CCC 0.5 threshold. No surveyor passed a minimum of 50 % of their comparisons and as such no reference counts could be generated with the current dataset.

Testing the Reference counts.

The application of the decision tree shown in Figure 4.5.3 to the current dataset results in reference footage not being obtained for FU15 due to a lack of inter-counter consistency (Step 2). Therefore subsequent testing of the reference counts could not be carried out.

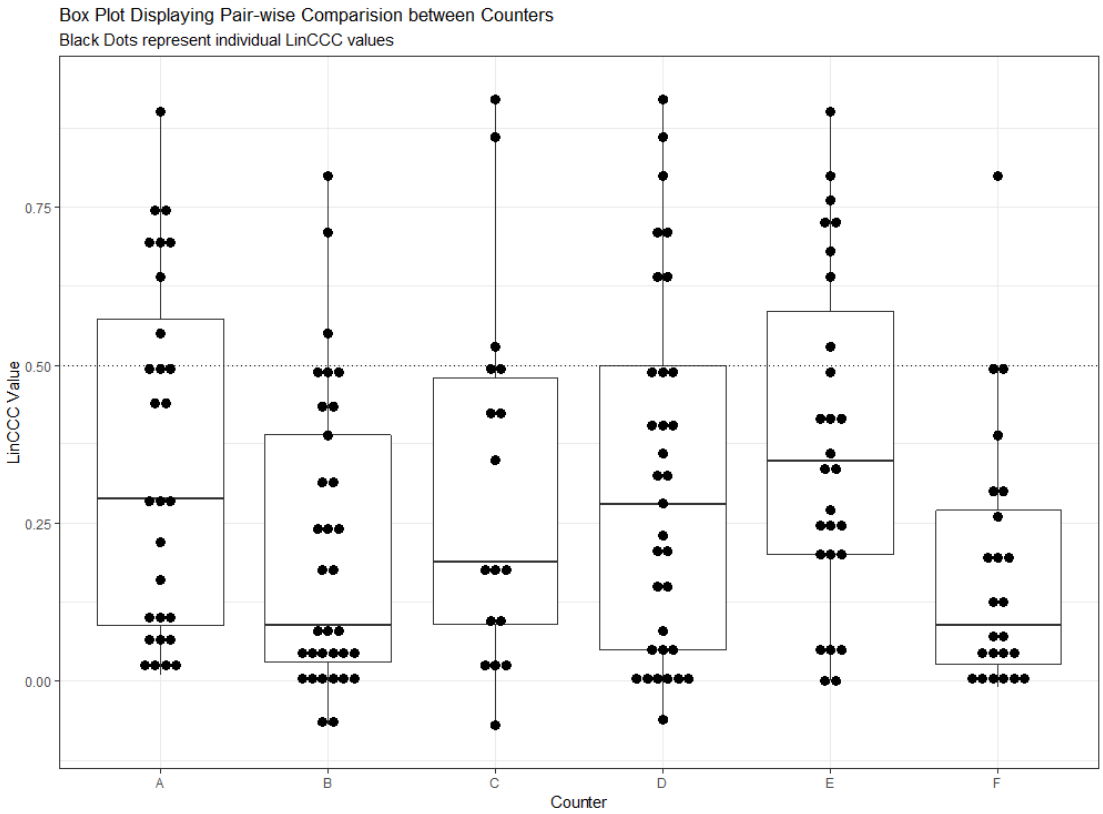


Figure 4.5.3: Step 2. Inter-reviewer Lin's CCC performance check. Boxplot displaying CCC value pairwise comparison between counters. Black Dots represent individual CCC values.

Conclusions.

Participants of the workshop concluded that the footage used for generating the reference counts was suitable for this purpose and acknowledged the difficulties due to high burrow densities and occurrence of non-target burrowing species in FU15.

Applying the methodology agreed at the current WKNEPS resulted in reference counts not being generated for FU15. It is well reported that identifying and counting *Nephrops* burrows from UWTV survey footage is a subjective methodology with variation occurring due to misclassification of burrows from/to another species or wrongly attributing *Nephrops* burrows to burrow complexes. Variability can occur not only between counters but also by the same counter evaluating the same on footage multiple occasions.

The working group suggests the following recommendations for FU15 reference footage which AFBI intend to action prior to the 2019 survey.

- As a training exercise, the first two minutes of each station should be analysed by randomized pair of counters and each individual burrow complex timestamped and annotated.
- Prior to assessing reference footage all surveyors will review this training footage to ensure consistency in burrow complex identification and any uncertainty reviewed with the original pair of counters.
- The reference footage should be re-assessed by 5 trained members of AFBI staff and an independent international reviewer. The current methodology

propose by the workshop should be reapplied to generate reference counts to test this method.

4.6 Labadie, Jones and Cockburn Banks (FU 20-21)

Figure 4. shows the decision tree in how to use Lin's Concordance Correlation Coefficient (CCC) to independently assess and generate the reference counts (Lin, 1989). Each station was counted twice and independently by five national and one international reviewers. Intra-reviewer performance was checked using Lin's CCC with a threshold of 0.5 to screen for internal consistency. When a reviewer did not pass the threshold for any of the stations, this resulted in their counts being dismissed for those stations (Figure 4, step 1). Only three stations out of 54 (six reviewers by nine stations) were dismissed by this process for FU20-21 (Figure 4.6.1)

Mean counts per minute from each reviewer were calculated, and inter-reviewer correlations were checked using Lin's CCC with a threshold of 0.5. When a reviewer failed more than 50% of its inter-reviewer pairings, all the counts of this reviewer were dismissed (Figure 4, step 2). Following this, the international reviewer's counts were dismissed from the process (Figure 4.6.2).

In order to calculate the reference counts for station, only counts that passed at least against another reviewer were used (Figure 4, step 3). The matrix in Figure 4.6.3 shows a different scenario for each of the stations reviewed. For example, in station number 4, counts from national reviewer numbers 1, 2 and 3 were used to generate the reference counts for this station. For station number 2 only two reviewers were used to generate the reference count; this is deemed acceptable by the WKNEPS, as it is the current quality control method of survey counts. The average of the counts per minute of all the reviewers who passed the previous steps were used to generate the reference counts for FU20-21.

Testing the Reference counts.

This workshop decided that it would be useful to also test the Lin's directed reference counts using the reviewers' data. This is important in scenarios where some reviewers' counts were dismissed in step 1, as they failed the intra performance check. Only the international reviewer failed against the reference counts for FU20-21 (Figure 4.6.4). This was as expected, as the international reviewer failed more than 50% of its pairings in step 2, and therefore its counts were not used to calculate the reference counts.

Conclusions.

The Lin's process is a useful method to objectively assess and generate reference counts for FU20-21. The previous reference set was based on the average of five national reviewers' data and this set was developed in year 2 of the survey series when local expertise was being developed. The decision tree chart shows clearly the steps involved in this process. The fact that the international reviewer did not perform well against the national reviewers for this reference set can be partly explained by the difference in camera set up, which varies across the national labs. Also, this survey area FU20-21 would be viewed as a challenging ground to count given that the burrow systems are not often of the classic morphology due to local area conditions, such as topography and also due to unknown benthic fauna species interactions.

The WKNEPS also recommended that the first two minutes of each reference station is to be annotated for training in identifying burrows for each survey area.

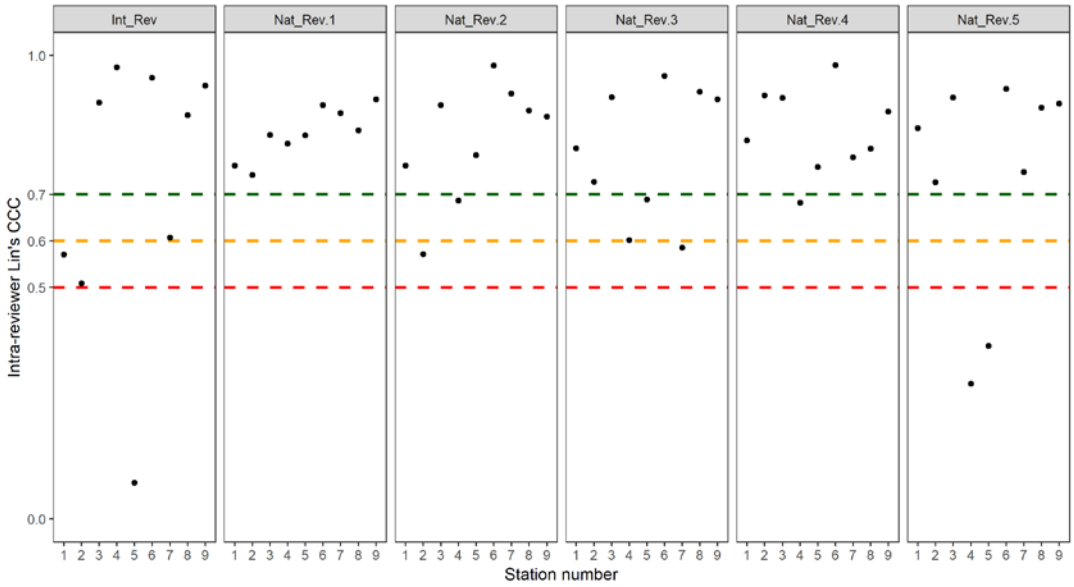


Figure 4.6.1: Step 1. Intra-reviewer Lin's CCC performance check for each of the reviewers (in panels) and each of the stations (x-axis). Dashed lines show different possible thresholds for the Lin's CCC.

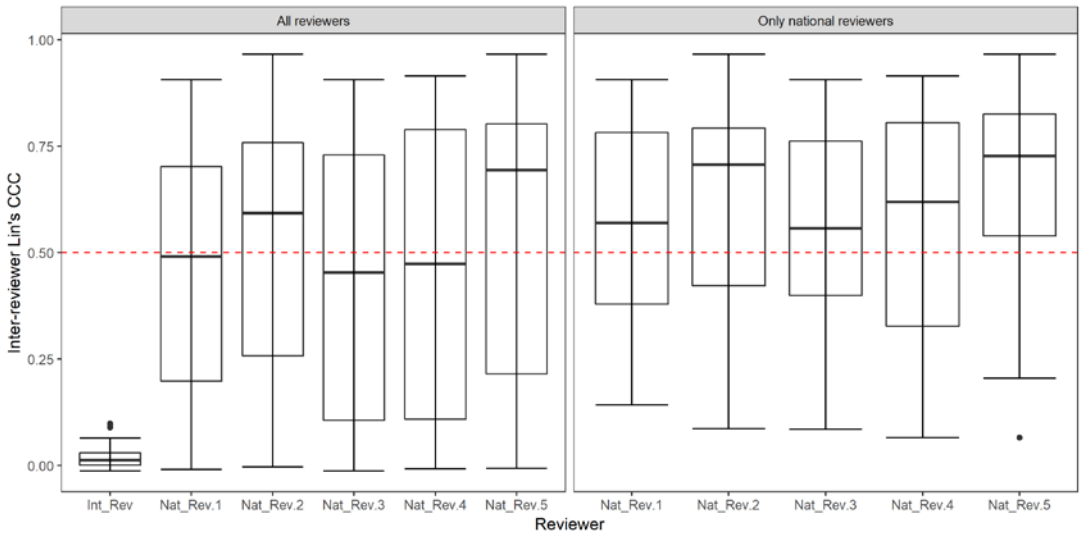


Figure 4.6.2: Step 2. Inter-reviewer Lin's CCC performance check. Boxplots show the distribution of Lin's CCC values for each of the reviewer's pairings. On the left panel, first check using mean counts of all the reviewers: note that the median (black band inside the box) of the international reviewer is well below the 0.5 threshold (red-dashed line). On the right panel, second check excluding the counts of the international reviewer.

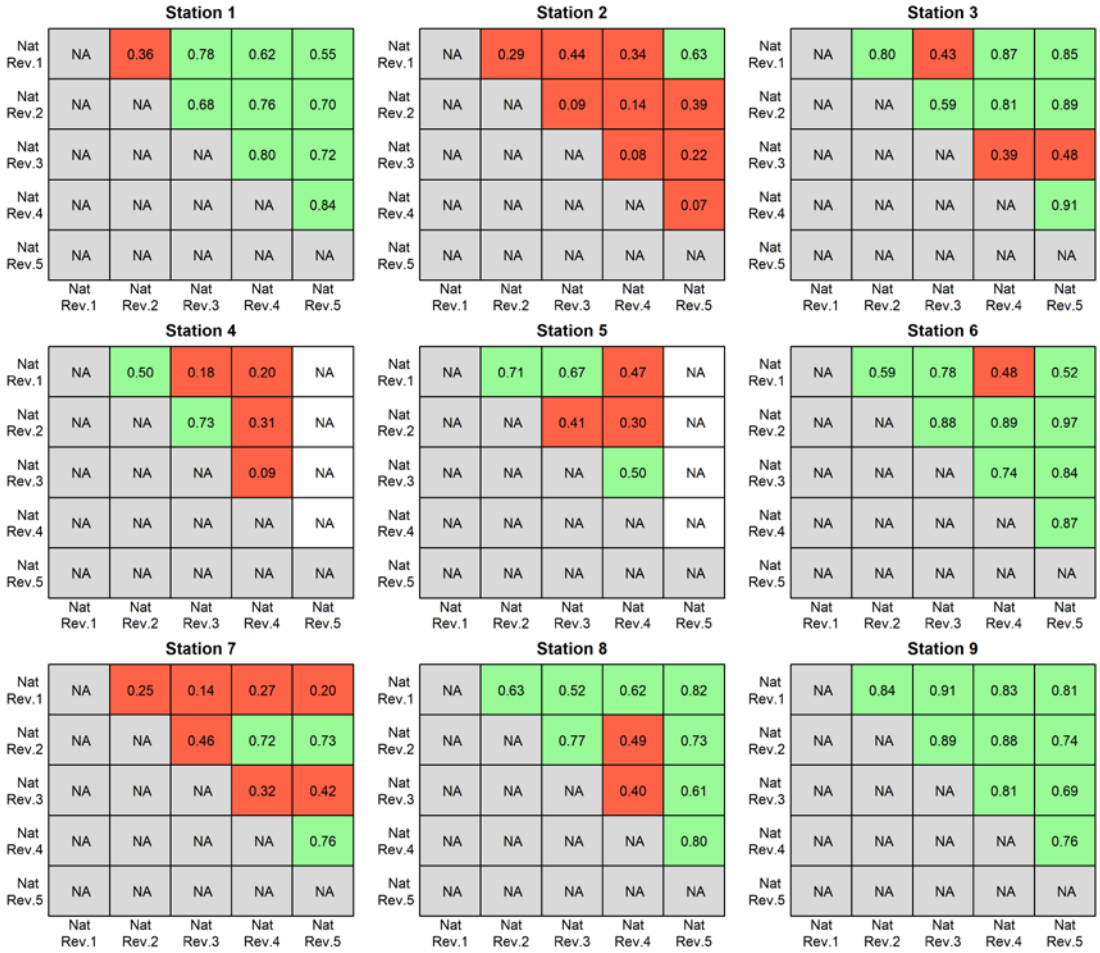


Figure 4.6.3: Step 3. Inter-reviewer Lin's CCC performance for each station. Pairings which passed the 0.5 threshold (in green) where later used for generating the reference counts. Pairings which did not pass the 0.5 threshold are shown in red. Counts which were dismissed in Step 1 for the national reviewer 5 in stations 4 and 5 are shown in white cells. International reviewer's pairings are not shown, as all its counts were dismissed in Step 2.

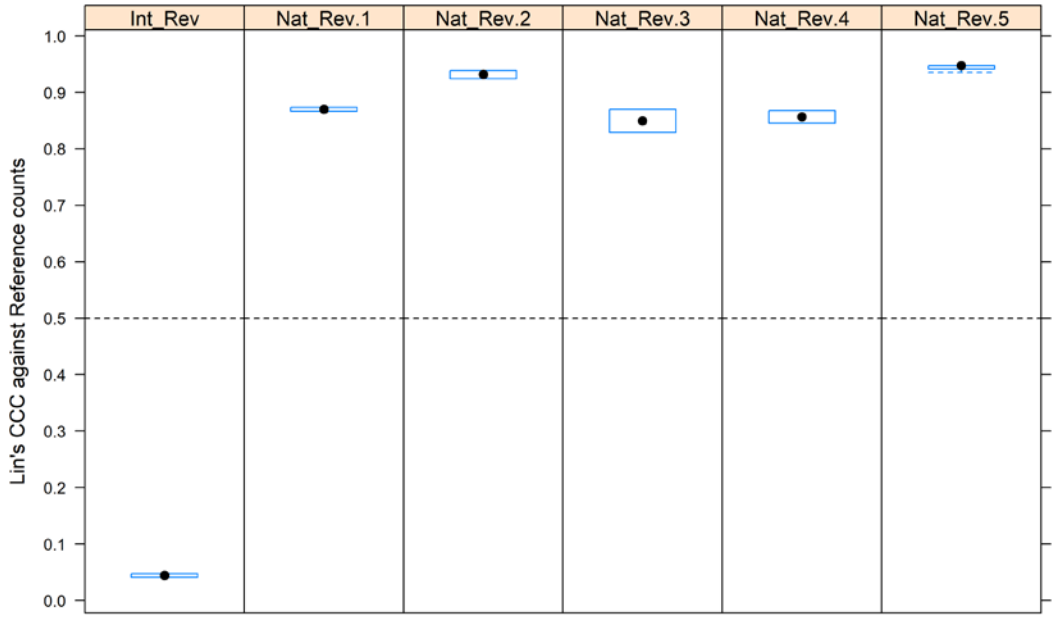


Figure 4.6.4: Lin's CCC values distribution for each of the reviewer's mean counts tested against the reference counts.

4.7 Bay of Biscay (FU 23-24)

Three national counters reviewed the reference set twice. The quality of the original DVDs was not "good", but it reflected the conditions during the 2018 survey. The reference set was sent to the international counter to be reviewed prior to the workshop. The international counter found the reference set to be very difficult to review due to poor visual clarity and highly pixelated footage and general poor quality.

The original DVDs of the selected reference stations were brought to the workshop and these compared to the compiled reference set in a group review session. It was then obvious that there was software processing format problem which had compressed the reference stations to a lower quality video file which explained the pixilation.

The workshop then decided that 5 reference stations would be reviewed during the workshop from the "original" DVDs. The three national reviewers and two international reviewers went through this process on CRT monitors where they reviewed the set in isolation twice and all stationed in the one review room.

Figure 4.7.1 shows the counts from the five stations reviewed of the original footage and there are trends of high variability between the counters for stations 1 to 3 and to somewhat lesser degree in stations 4 to 5. The mean count for each reviewer is shown in Figure 4.7.2. This shows that there are two counting trends in stations 1 to 3 whereas in stations 4 to 5 the count data are within a closer range.

The dataset was not further tested using the decision tree in Figure 4 due to the problems encountered with the quality of the reference set and time constraints.

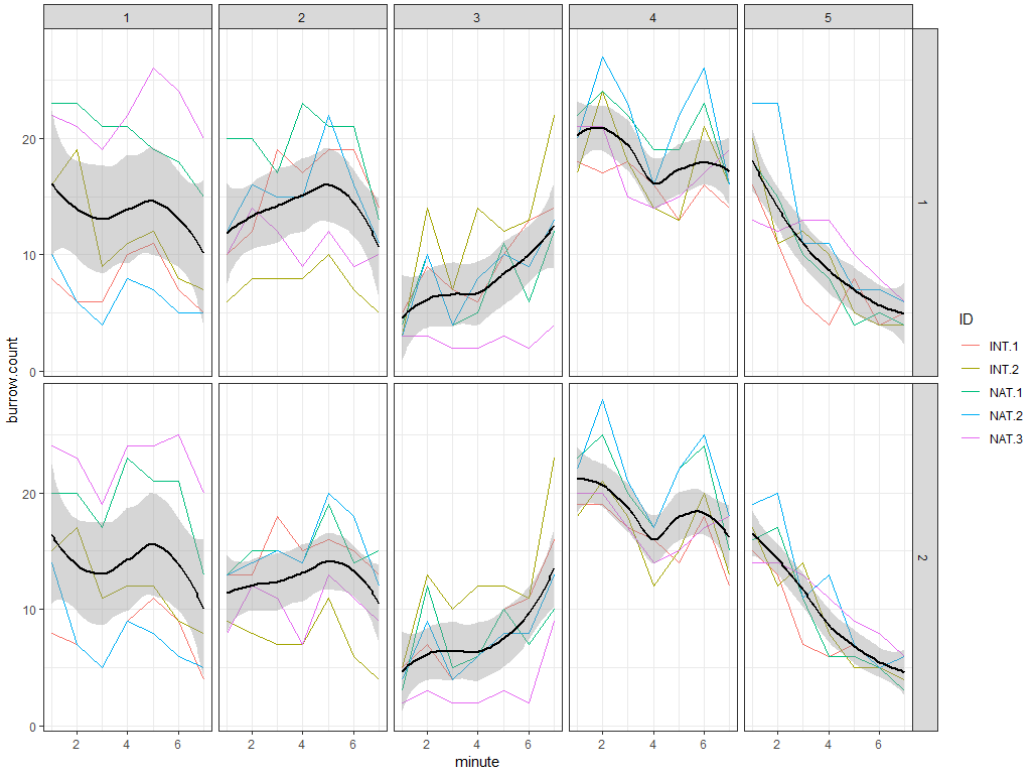


Figure 4.7.1 Minute (1 to 7) counts for all reviewers by station (1 to 5) and review number (1 to 2) with a loess smoother (black line). INT = international reviewer and NAT = national reviewer.

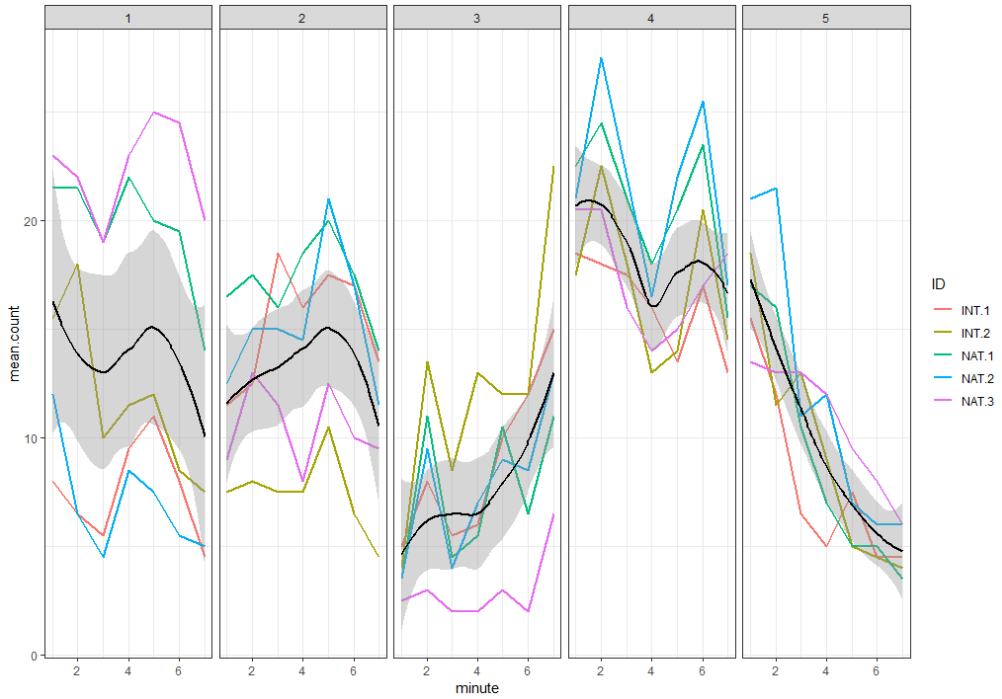


Figure 4.7.2 Mean counts by minute for each reviewer by station (1 to 5) with a loess smoother (black line) for the FU23-24 reference set. INT = international reviewer and NAT = national reviewer.

4.8 Results

The table below gives an overview of the reference counts work completed at the workshop. The use of Lins's CCC to develop reference counts appears to be a robust method. It is also transparent and easy to follow process.

FU	Number of National Reviewers	Number of International Reviewers	Reference Count Method	Comments
1	2	1	Annotated and consensus agreed	Reference Set Completed
3-4	2	4	See comment	Training set at workshop
6	4	1	Lin's CCC method	Reference Set Completed
11	3	1	Lin's CCC method	Reference Set Completed
15	5	1	Lin's CCC method	Intersessional work
20-21	5	1	Lin's CCC method	Reference Set Completed
23-24	3	2	See comment	Moving to HD system
30	3	2	Lin's CCC method	Reference Set Completed

5 Footage Review Groups

Group review sessions were held so that each reference set could be assessed in terms of quality (sledge speed, ground contact, visual clarity) and also to familiarise with the inherent characteristics of each survey area. The general practice was to review a small section of footage, commentating on what was being observed, and then to run several minute-by-minute counts to see how people were performing. When large discrepancies were observed further discussion around the rationale behind "accepted" counts ensued to harmonize the search pattern of reviewers. These plenary groups have been held at previous workshops where participants benefit from observing the different ground types.

The following summarizes the reviews of each survey area:

5.1 High Definition Format

Iceland (FU 1)

Footage quality clear and relatively easy to identify and count the *Nephrops* burrows which are quite large.

Skaggerak and Kattegat (FU 3-4)

Survey footage from 2017 to 2018 was used to train the national novice reviewer in a small group with experienced reviewers. Some of the footage was annotated where the group discussed burrows as they came into view. Survey footage presented was good in terms of visual clarity and speed. *Nephrops* burrow systems were easy to identify. Line lasers can sometimes bend around the outside of *Nephrops* burrows which was first noticed at the workshop in 2016 (ICES, 2017a) from this survey area. This "bending" effect and how to judge when a burrow is within the field of view when this occurs needs to be noted in the standard procedure for counting when using line lasers.

Screen grab below shows footage quality and line laser just touching a *Nephrops* burrow. In this case the *Nephrops* burrow would be counted as laser is touching the burrow structure.

Farne Deeps (FU 6)

Inshore area with many small burrow systems and medium density. Footage of good quality and relatively easy to identify burrows.

Gulf of Cadiz (FU 30)

Some problems related to the identification of small *Nephrops* burrow systems and other megafauna burrows were identified at the last WKNEPS (ICES, 2017a). WKNEPS 2017 agreed that Spain should establish a criterion about which burrow systems of small size should be counted as *Nephrops* burrows. In 2018, videos at different level of species interaction were reviewed in a group session and the size of the burrow systems was analyzed and discussed. Examples of small *Nephrops* and other megafauna inside of burrows helped to establish a reasonable doubt about some other small burrow systems. So if you are in doubt do not count. Other characteristics of the morphology of the burrows of the different species were taken account besides of size. The criterion of the experts in this FU was accepted. Some snapshots from footage in FU 30 are shown as examples in Figure 5.3.1.

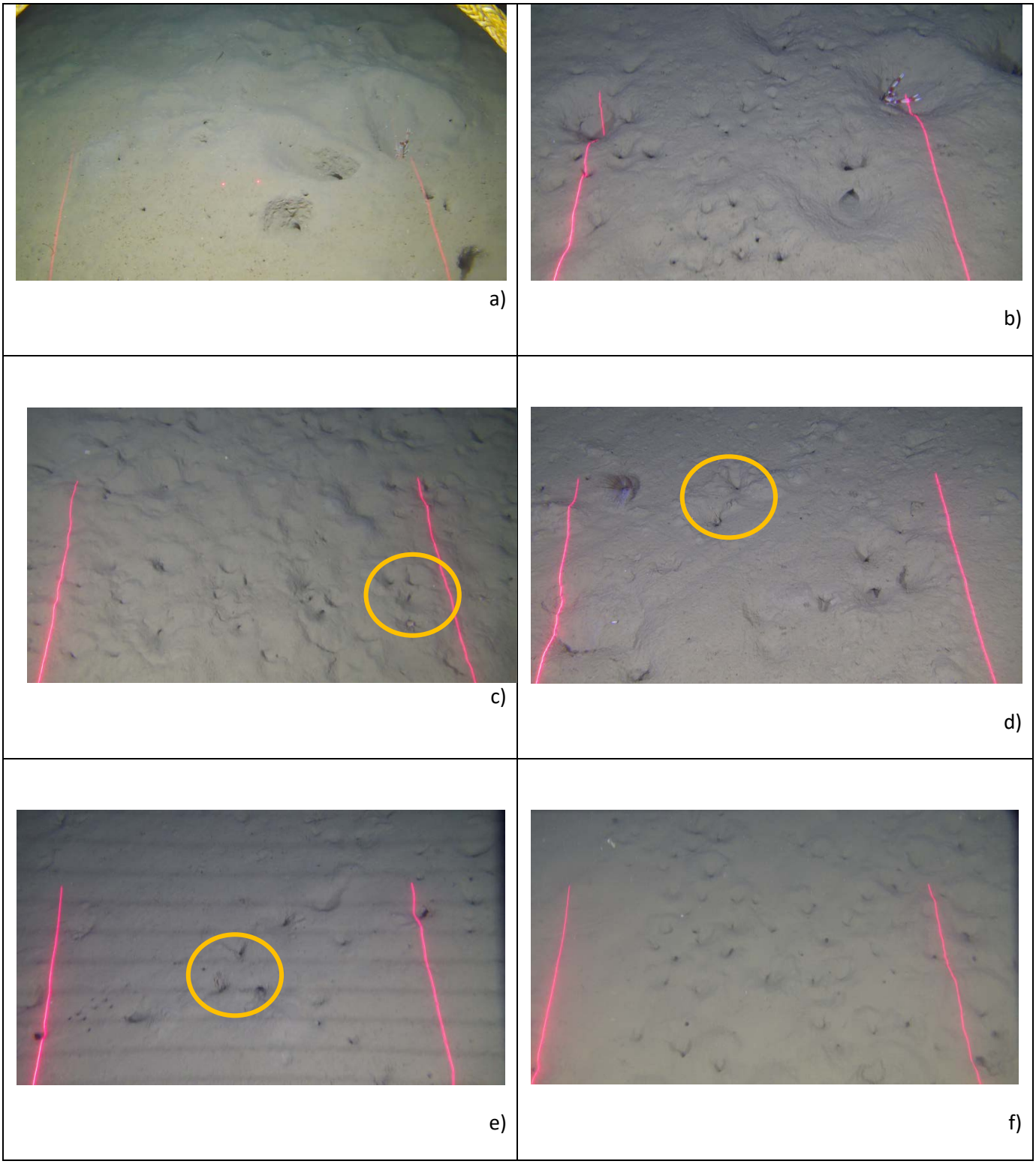


Figure 5.3.1. ***Nephrops*** **burrows (a & b); Squat lobster burrows (c & d); Crab burrows (e & f).**

5.2 Standard Definition Format

North Minch (FU 11)

Random minutes from all nine video clips from the North Minch (FU11) reference set were reviewed and discussed by the majority of the group. With the footage representing the wide variety of benthic environments and burrow densities (belonging to both Nephrops and other burrowing fauna), each clip was unique and generated productive discussions. Water clarity, burrow definition, lighting arrangements and burrow identification were all topics raised with constructive advice being well received. There were no major concerns raised by the group regarding the footage presented for this area.

Eastern Irish Sea (FU 14)

Moderate-high density footage, being the average 7 burrows per minute. The identification of *Nephrops* burrows can be challenging in some stations where the burrowing species *Calocaris macandreae* and *Goneplax rhomboides* are present. Quality of footage was good in terms of speed and ground contact.

Western Irish Sea (FU 15)

The group agreed that the footage was of good quality in terms of visual clarity and speed. Due caution in areas of high burrow densities and occurrence of non-target burrowing species in FU15 where *Calocaris macandreae* is present.

Labadie, Jones and Cockburn Banks (FU 20-21)

Footage difficult due to many undulations of the seabed which may be caused by oceanographic effects and / or other seabed life. High variability in sediment type and topography which can cause difficulty in identifying burrow structures as *Nephrops* or other burrowing megafauna.

Aran grounds (FU 17)

Reference footage was not presented for this FU at the workshop. However due to a change in burrow sizes observed on the 2018 survey this footage was shown to the group to discuss as there were very small burrows observed during the survey. It was agreed that these burrows were more than likely constructed by *Nephrop*s < 17 CL mm where this is the size of first UWTV selection (Leocádio *et al.*, 2018). The group agreed it would be useful to have a minimum burrow entrance size which could be used as a rule for counting *Nephrops* burrows.

Bay of Biscay (FU 23-24)

Following individual counts, a group review sessions was held. Several minutes from a selection stations were reviewed but some of the reviewers present did not agree on their counts given the many small burrows and the quality of the videos.

Resulting from the group review of the footage, it was advised that some of the counts should be more cautionary given the presence of *Munida* species which can cause some misallocation errors.

6 Recommendations

The workshop made recommendations in the following areas:

Reference Sets Compilation:

- To annotate and time-stamp the first two minutes of each reference station, which will be used as training material so that reviewers will understand what to count.
- To update the underwater TV SISP manual with standards for compiling the reference set.

Developing Reference Counts:

- To use Lin's CCC method to develop and calculate reference counts for any new reference sets where the process is to include at least one international reviewer.
- To support any requests for international reviews of reference sets.
- To present results of future work on reference count process to WGNEPS and relevant assessment working groups (WGBIE, WGCSE and WGNSSK).
- To update the underwater TV SISP manual with details for developing reference counts.
- New reference material, and associated counts, are to be established whenever there is a change in the UWTV equipment used or when it is recognized that there has been a significant change in the benthic environment.

Training Material:

- To develop and keep up-to-date training material such as photo libraries.
- To update the underwater TV SISP manual with details on how to count burrows as within the field of view if using a line laser so that reviewers will understand how to count burrows in such cases.

Research of burrows of other species:

The burrow entrances made by species such as *Goneplax rhomboides* or *Munida* species have some similar signatures to *Nephrops* burrows and can be misidentified. Further studies are needed to distinguish the different burrow systems and/or to estimate the relative abundance of the competing species. These studies could include the use of dredges, fixed cameras, new resin casts or long term monitoring of burrows using landers.

Automatic counting algorithms:

To investigate and support avenues in automatic counting algorithms where reference sets could be used as training sets.

7 References

ICES, 2008. Report of the Workshop and training course on *Nephrops* burrow identification (WKNEPHBID). ICES CM: 2008/ LRC, PGCCDBS.

ICES, 2017a. Report of the Workshop on *Nephrops* burrow counting, 9-11 November 2016. Reykjavík, Iceland. ICES CM 2016/SSGIEOM:34. 62 pp.

ICES, 2017b. Report of the Working Group on *Nephrops* Surveys (WGNEPS). ICES CM 2017/SSGIEOM: 30. Ref: ACOM/ SCICOM.

ICES (in prep.). Manual for the *Nephrops* Underwater TV Surveys. Series of ICES Survey Protocols (SISP).

Leocádio, A., Weetman, A., and Wieland, K. (Eds). 2018. Using UWTV surveys to assess and advise on *Nephrops* stocks. ICES Cooperative Research Report No. 340. 49 pp. http://doi.org/10.17895/ices.pub.4370

Lin, L, I. 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 45, pp255-268.

Annex 1: List of Participants

NAME	ADDRESS	PHONE	EMAIL
Adrian Weetman (co-Chair)	Marine Scotland Science (MSS), Marine Laboratory, 375 Victoria Road, Aberdeen, UK AB11 9DB	+44 (0)131 2444142	weetmana@marlab.ac.uk
Anders Wernbo	Swedish University of Agricultural Sciences (SLU), Department of Aquatic Resources, Institute of Marine Research, Turist-gatan 5 S- 453 30 Lysekil, Sweden	+46(0)10 4784048	anders.wernbo@slu.se
Anna Ragnheiður Grétarsdóttir	Marine and Freshwater Research Institute (MFRI), Skulagata 4, 121 Reykjavik, Iceland	+354 575 2000	anna.ragnheidur.gretars-dottir@hafogvatn.is
Auður Bjarnadottir	Marine and Freshwater Research Institute (MFRI), Skulagata 4, 121 Reykjavik, Iceland	+354 575 2000	audur.bjarnadottir@hafogvatn.is
Candelaria Burgos	Instituto Español de Oceanografía (IEO), Centro Oceanográfico de Cádiz Puerto Pesquero, Muelle de Levante s/n E-11006, Cádiz, Spain	+34 956 294189	caleli.burgos@ieo.es
Ewen Bell	Centre for Environment, Fisheries and Aquaculture Development (CEFAS) Pakefield Road Lowestoft Suffolk, UK NR33 0HT	+44 (0) 1502 562244	ewen.bell@cefas.co.uk
Gráinne Ryan	Marine Institute (MI) Rinville, Oranmore, Co. Galway, Ireland	+353 91387200	grainne.ryan@marine.ie
Jean-Jacques Rivoalen	IFREMER Station de Lorient - 8, rue François Toullec – 56100, Lorient, France	+33 (0)297873845	jean.jacques.rivoalen@ifremer.fr

NAME	ADDRESS	PHONE	EMAIL
Jean-Philippe Vacherot	IFREMER Station de Lorient - 8, rue François Toullec – 56100, Lorient, France	+33(0)297 873813	jean.philippe.vacherot@ifremer.fr
Jennifer Doyle (co-Chair)	Marine Institute (MI) Rinville, Oranmore, Co. Galway, Ireland	+353 91387200	jennifer.doyle@marine.ie
Katie Boyle	Marine Scotland Science (MSS) Marine Laboratory, 375 Victoria Road, Aberdeen, UK AB11 9DB	+44(0)131 2444162	katie.boyle86@gmail.com
Michelle Inglis	Marine Scotland Science (MSS) Marine Laboratory, 375 Victoria Road, Aberdeen, UK AB11 9DB	+44(0)131 2442500	michelle.inglis@gov.scot
Mikel Aristegui-Ezquibela	Marine Institute (MI) Rinville, Oranmore, Co. Galway, Ireland	+353 91387200	mikel.aristegui@marine.ie
Peter McCorriston	Agri-Food and Biosciences Institute (AFBI) 18a Newforge Lane, BT9 5PX Belfast, Northern Ireland (UK)	+44(0)2890 255153	peter.mccorriston@afbini.gov.uk
Rosana Ourens	Centre for Environment, Fisheries and Aquaculture Development (CEFAS) Pakefield Road Lowestoft Suffolk, UK NR33 0HT	+44 (0) 1502 562244	rosana.ourens@cefas.co.uk
Rosemarie Butler	Marine Institute (MI) Rinville, Oranmore, Co. Galway, Ireland	+353 91387200	rosemarie.butler@marine.ie
Sinéad O'Brien	Marine Institute (MI) Rinville, Oranmore, Co. Galway, Ireland	+353 91387200	sinead.o'brien@marine.ie

NAME	ADDRESS	PHONE	EMAIL
Svend Koppetsch	Swedish University of Agricultural Sciences (SLU), Department of Aquatic Resources, Institute of Marine Research, Turist-gatan 5 S- 453 30 Lysekil, Sweden	+46(0)10-4784048	svend.koppetsch@slu.se
Yann Coupeau	IFREMER Station de Lorient - 8, rue François Toullec – 56100, Lorient, France	+33(0)297 873842	yann.coupeau@ifremer.fr
Yolanda Vila	Instituto Español de Oceanografía (IEO). Centro Oceanográfico de Cádiz Puerto Pesquero, Muelle de Levante s/n, E-11006, Cádiz, Spain	+34 956294189	yolanda.vila@ieo.es

Annex 2: Agenda

Day 1 – 02/10/18

09:00 Welcome
Adoption of agenda
ToRs
Room allocation and reviewing arrangements
Background history of *Nephrops* TV workshops
Summary of data to be reviewed and desired outcomes from meeting
Welcome from MSS Coastal and Offshore Fisheries Programme Manager

10:30-10:45 Coffee break

10:45-13:00 Presentations (ToR a):
Methodology in generating North Minch (FU 11) reference footage. A. Weetman.
Methodology in generating Labadie (FU 20-21) reference footage. J. Doyle, M. Aristegui
Methodology in generating Western Irish Sea (FU 15) reference footage. P. McCorriston.
Methodology in generating Gulf of Cadiz (FU 30) reference footage. Y. Villa.

13:00-14:00 Lunch.

14:00-15:00 Subgroups undertaking international reviews where this was not possible prior to the meeting (ToR a):
FU 1 internationally reviewed by Spain
FU 6 internationally reviewed by Scotland.
FU 23-24 internationally reviewed by England and Ireland.
FU 30 internationally reviewed by France and Ireland.

15:00-15:15 Coffee break

15:15-18:00 Presentation:
Methodology in generating Bay of Biscay (FU 23-24) reference footage. J-P. Vacherot.

Break-out subgroups (ToR a):
SLU received training by staff from Iceland and AFBINI
CEFAS, AFBINI and MSS group discussion Farne Deeps (FU 6).

18:00 Adjourn.

Day 2 – 03/10/18

09:00-09:30 Discuss any issues experienced the previous day

09:30-10:30 Group review of reference footage from North Minch (FU 11) and Bay of Biscay (FU 23-24, ToR b).

10:30-10:45 Coffee break.

10:45-13:00 Group review of Western Irish Sea (FU 15, ToR b).

Continue international counting of reference footage from Iceland (FU 1), Bay of Biscay (FU 23-24) and Gulf of Cadiz (FU 30, ToR a).

13:00-14:00 Lunch.

14:00-15:00 Group discussion on reviewing methodologies and possible technological developments (ToR b).

15:00-15:15 Coffee break.

15:15-18:00 Drafting text for report.

18:00 Adjourn.

Day 3 – 04/10/18

09:00-09:30 Discuss progress and any issues experienced the previous day

09:30-12:00 Drafting text for report.
Group discussion on observations made during the process of generating and reviewing reference footage (ToR a).

12:00-13:00 Presentations (ToR a):
New UWTV equipment used in the Gulf of Cadiz (FU 30).
Methodology in generating Iceland (FU 1) reference footage. A. Ragnheiður/ A. Bjarnadottir.

13:00-14:00 Lunch.

14:00-15:00 Presentations (ToR a):
Reference count results using R from Iceland (FU 1) reference footage. A. Bjarnadottir
Reference count results using R from Gulf of Cadiz (FU 30) reference footage.
Y. Villa/C. Burgos.
Reference count results using R from Farne Deeps (FU 6) reference footage.
R. Ourens/E. Bell.

Initial group discussion of footage from Bay of Biscay (FU 23-24, ToR a).
Subgroup to undertake further international reviewing of Bay of Biscay (FU 23-24, ToR a).
Group discussion Skaggerak and Kattegat (FU 3-4) reference footage (ToR a).

Discussion regarding practical ways to gather samples of benthic fauna to relate to observations recorded from UWTV footage (ToR a).

15:00-15:15 Coffee break.

15:15-18:00 Further development and use of R scripts to generate reference values from data generated from counts.

18:00 Adjourn.

Day 4 – 05/10/18

09:00-09:30 Discuss progress and any issues experienced the previous day.

09:30-13:00 SLU received training by staff from Iceland, AFBINI and Marine Institute (ToR a).
Presentation (ToR a):
Reference count results using R from Western Irish Sea (FU 15) reference footage.
P. McCorriston.

Drafting text for report.

13:00-14:00 Lunch.

14:00-15:00 Discussion about methods to annotate video for training purposes (ToR a).

15:00-15:15 Coffee break

15:15-16:30 Presentations:
Reference count results using R from North Minch (FU 11) reference footage. A. Weetman.
Presentation on the UWTV survey from Bay of Biscay (FU 23-24). J-P. Vacherot.

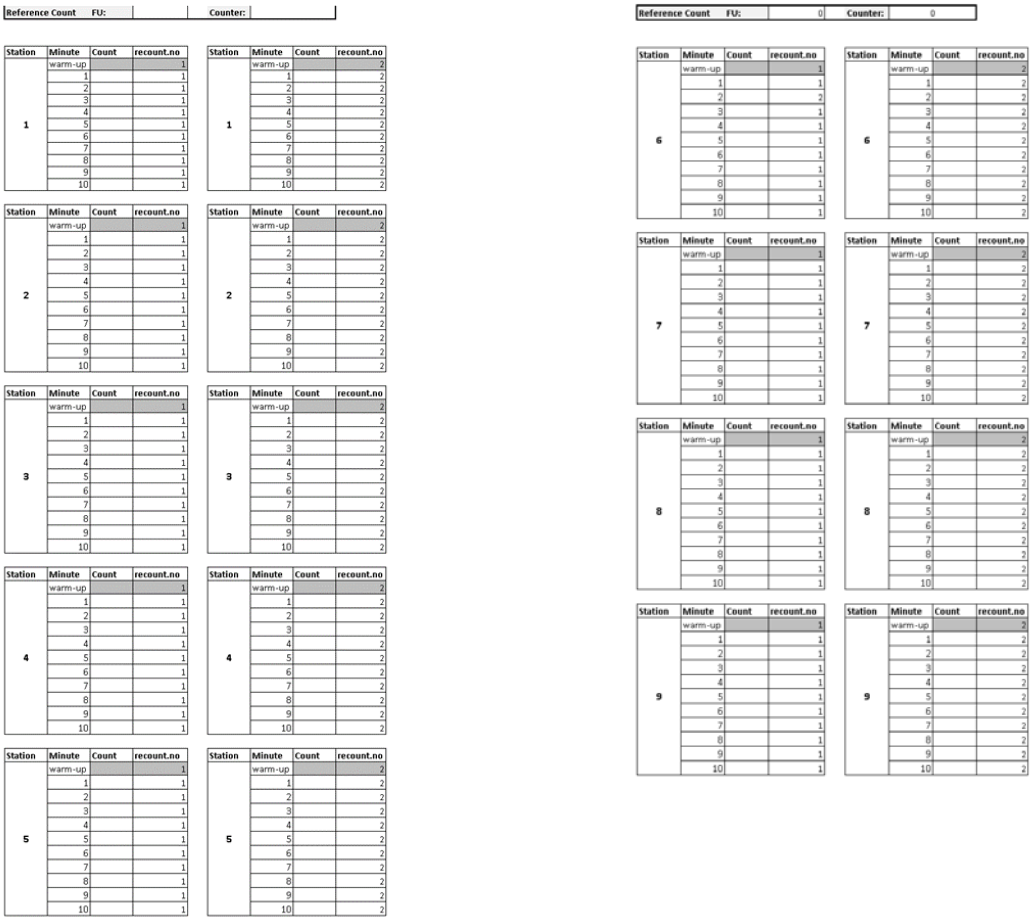
Final group discussion of reference footage from Bay of Biscay (FU 23-24).
Review Bay of Biscay (FU 23-24) reference values.

Group discussion of burrow size recorded from Aran grounds (FU 17, ToR b).
Group discussion of footage from Labadie, Jones and Cockburn Banks (FU 20-21) (ToR b).

Plenary on draft report and recommendations (ToR d).

16:30 Official closure.

Annex 3: Excel template for review data



Reference Count	FU:		Counter:	

Station	Minute	Count	recount.no
1	warm-up		1
	1		1
	2		1
	3		1
	4		1
	5		1
	6		1
	7		1
	8		1
	9		1
	10		1

Station	Minute	Count	recount.no
1	warm-up		2
	1		2
	2		2
	3		2
	4		2
	5		2
	6		2
	7		2
	8		2
	9		2
	10		2

Station	Minute	Count	recount.no
2	warm-up		1
	1		1
	2		1
	3		1
	4		1
	5		1
	6		1
	7		1
	8		1
	9		1
	10		1

Station	Minute	Count	recount.no
2	warm-up		2
	1		2
	2		2
	3		2
	4		2
	5		2
	6		2
	7		2
	8		2
	9		2
	10		2

Station	Minute	Count	recount.no
3	warm-up		1
	1		1
	2		1
	3		1
	4		1
	5		1
	6		1
	7		1
	8		1
	9		1
	10		1

Station	Minute	Count	recount.no
3	warm-up		2
	1		2
	2		2
	3		2
	4		2
	5		2
	6		2
	7		2
	8		2
	9		2
	10		2

Station	Minute	Count	recount.no
4	warm-up		1
	1		1
	2		1
	3		1
	4		1
	5		1
	6		1
	7		1
	8		1
	9		1
	10		1

Station	Minute	Count	recount.no
4	warm-up		2
	1		2
	2		2
	3		2
	4		2
	5		2
	6		2
	7		2
	8		2
	9		2
	10		2

Station	Minute	Count	recount.no
5	warm-up		1
	1		1
	2		1
	3		1
	4		1
	5		1
	6		1
	7		1
	8		1
	9		1
	10		1

Station	Minute	Count	recount.no
5	warm-up		2
	1		2
	2		2
	3		2
	4		2
	5		2
	6		2
	7		2
	8		2
	9		2
	10		2

Reference Count	FU:	0	Counter:	0

Station	Minute	Count	recount.no
6	warm-up		1
	1		1
	2		1
	3		1
	4		1
	5		1
	6		1
	7		1
	8		1
	9		1
	10		1

Station	Minute	Count	recount.no
6	warm-up		2
	1		2
	2		2
	3		2
	4		2
	5		2
	6		2
	7		2
	8		2
	9		2
	10		2

Station	Minute	Count	recount.no
7	warm-up		1
	1		1
	2		1
	3		1
	4		1
	5		1
	6		1
	7		1
	8		1
	9		1
	10		1

Station	Minute	Count	recount.no
7	warm-up		2
	1		2
	2		2
	3		2
	4		2
	5		2
	6		2
	7		2
	8		2
	9		2
	10		2

Station	Minute	Count	recount.no
8	warm-up		1
	1		1
	2		1
	3		1
	4		1
	5		1
	6		1
	7		1
	8		1
	9		1
	10		1

Station	Minute	Count	recount.no
8	warm-up		2
	1		2
	2		2
	3		2
	4		2
	5		2
	6		2
	7		2
	8		2
	9		2
	10		2

Station	Minute	Count	recount.no
9	warm-up		1
	1		1
	2		1
	3		1
	4		1
	5		1
	6		1
	7		1
	8		1
	9		1
	10		1

Station	Minute	Count	recount.no
9	warm-up		2
	1		2
	2		2
	3		2
	4		2
	5		2
	6		2
	7		2
	8		2
	9		2
	10		2

To the left, the front page and on the right the second page of the Excel table used to record the counts from the first and second passes by each reviewer.