Scottish Government
Riaghaltas na h-Alba
gov.scot

Marine Scotland

Seabird Survey Data Protocol

marinescotland

Seabird Survey Data – Automated Data Review 06/03/2017

This note is a guide for the R script (MS.seabird.survey.data.check v1.2.r) developed to undertake an automated review of a proposed standardised seabird survey data format. This data format and script were produced for Marine Scotland and followed discussions with JNCC, SNH, The Crown Estate and surveying companies. The script performs a series of checks of survey data and identifies departures from the expected formats or data content.

The script and associated files and folders are all contained within the following zip folder:

- MS.seabird.survey.data.check v1.2.zip

Once extracted this should have the following structure (folders in italics):

MS Seabird Data Check
└ MS.seabird.survey.data.check v1.2.r
└ *data template*
  └ Standardised seabird survey data structure protocol 13_12_2016.xlsx
└ *data requirement info*
  └ ESAS behaviour codes.xlsx
  └ ESAS.species.codes.xlsx
  └ euring.codes.txt
  └ expected.observation.fieldnames.txt
  └ expected.summary.fieldnames.txt
  └ expected.track.fieldnames.txt
  └ full.species.list.xlsx
└ *example data*
  └ Digital Aerial Still Example 27_01_2017.xlsx
  └ Digital Aerial Video Example 27_01_2017.xlsx
  └ ESAS Example 27_01_2017.xlsx
  └ Summary_unique_observations.csv
  └ Summary_unique_observations.txt

Data template

The template for the survey data is an Excel file (Standardised seabird survey data structure protocol 13_12_2017.xlsx) in the data template folder which contains three sheets:

- Summary,
- Track, and
- Observations.

The template has been developed to accommodate surveys conducted by boat (ESAS methods) and digital aerial imaging using either still or video methods. The differing nature of these survey methods means that not all data field types are relevant in each case. The file above indicates for each field type whether or not data should be recorded for each survey method. It should be noted that the intention is not to remove unused fields but rather enter suitable null values (e.g. ‘NA’). This will help to act as a prompt to ensure that all required data are included. The template structure is also set out in Appendix 1. It is also important to note that additional fields can be included in the data, but these will not be reviewed: only field which match the expected names are checked.

Automated data review

The R script loads data stored in an excel file with sheets named ‘**Summary**”, “**Track**” and “**Observations**” (these names must be used for the script to run correctly) and performs a series of tests on the contents of each. In order to complete these tests the following additional files are also (automatically) loaded into the R workspace:

- ‘expected.summary.fieldnames.txt’
- ‘expected.track.fieldnames.txt’
- ‘expected.observation.fieldnames.txt’
- ‘full.species.list.xlsx’
- ‘ESAS.species.codes.xlsx’

These files are saved in a folder (‘data requirement info’) which is included in the zip folder (‘MS.seabird.survey.data.check.r’) and contain lists of the expected formats and species names.

To run the R script the user needs to provide the following information within the script itself:

- Enter the data file folder location (this variable is called ‘*root.folder*’)
- This needs to follow R folder naming protocols, with quotation marks around the name and use of either a double backslash ‘\\’ or a single forward slash ‘/’ as the folder separator. The folder name must end with a folder separator, e.g.

```
root.folder="C:\\Users\\mtrinder\\MS\\Survey data protocols\\data structure and
script\\"
```

- Enter the excel file name (this variable is called ‘*data.file.name*’), e.g.

```
file.name = "ESAS Example 27_01_2017.xlsx"
```

No further input or editing of the script should be required. Once the above variables have been supplied at the top of the script, it can be run (e.g. if using RStudio click on ‘Source’ or use the ‘Ctrl-Shift-S’ shortcut combination). The script undertakes the following checks, with outputs to the R console indicated in italics.

'*CHECKING FIELD NAMES IN EACH DATA SHEET... '*

The column names in each of the three data sheets (summary, track, observations) are compared with the expected ones. For each sheet one of two messages is outputted to the R console, either:

1. '*all summary/track/observation fields present'*, or
2. '*the following summary/track/observation fields appear to be missing*: ', followed by a list of the expected field names which are missing.

'*CHECKING OBJECT CODES AGAINST ESAS CODES...*'

Observed objects (e.g. seabirds, marine mammals, etc.) are checked against the standard ESAS codes with output to the R console as follows:

1. '*Names of all objects in data*:' – this lists all the unique names in the object name field of the observations sheet.
2. '*Names of all objects in data which have an ESAS code*:' – this lists all the unique names in the object name field of the observations sheet which also have an ESAS code number (i.e. in the 'object.ID' field)
3. '*Official ESAS designation for objects in data - does this match previous output*?:' – this extracts the official (ESAS) species names for each of the ESAS codes in the data and outputs them to the R console. Comparison of these with the previous two outputs should allow the user to identify any mismatches between ESAS codes and species names.
4. '*If two lists are different then ESAS code or object name may have been incorrectly entered, check this table of unique objects and codes from the data*:' – this is a final output of object names and IDs to allow the source of any mismatches to be identified.

'*CHECKING FIELD CONTENT AGAINST EXPECTED TYPE*:'

The content of each field type (in all three sheets) is compared against the expected data format specified in the 'expected.summary.fieldnames.txt' (and the equivalent files for track and observations). The data formats have been defined as either 'numeric' or 'character'.

1. '*These data are identified as*:' – the survey.type in the Summary sheet is outputted to the console.
2. '*Given this survey type are the following mismatched field types a concern?*:' - this should confirm if the data are from a boat, digital still or digital video survey and thus which of the field types may be expected to have the wrong format (e.g. a boat survey will have 'NA' in the 'image.resolution' field which will be identified by R as 'character' instead of the expected 'numeric' given for this in the data requirement file).
3. '*The following SUMMARY/OBSERVATION/TRACK field types do not have the expected format (numeric, character, etc.) and may need to be checked*:' – the column headings for the data which do not match the expected format are outputted to the console, e.g.

```
                field.names data.field.type expected.field.type
```

```
1 Snapshot.window.length       character             numeric
2  Snapshot.window.width       character             numeric
3      Snapshot.interval       character             numeric
4     Snapshot.frequency       character             numeric
5               No.sides       character             numeric
6           No.observers       character             numeric
```

[note that the row number on the left is created as a default component of a dataframe and has no connection to the content of the tables under review].

4. The above example are Summary fields for a digital still survey, and all of the identified fields which have the wrong format are ones used for defining boat surveys. Therefore, this would not raise any concerns. However, if a time or date field was identified as 'character' rather than 'numeric' this would be a potential error which would need to be investigated. Any columns which contain more than one data type (e.g. numeric and character) are treated as character, thus this test will identify even a single entry which does not match the correct format.
5. A text file containing the unique values present in the main fields of the observation data is created ('*Summary_unique_observations.csv*') and saved in the same folder as the data. This will permit the source of any errors to be identified. This file excludes any fields that would have a unique entry in each row (e.g. image.ID, Latitude, etc.) as these would be as long as the data itself. As a csv file it opens by default in Excel, but can also be viewed in a text editor (e.g. Notepad).

'*CHECKING IF FLIGHT HEIGHT HAS BEEN RECORDED FOR BIRDS ON THE SEA*:' –

1. This section extracts observations with a flight height estimate (i.e. not 'NA' or blank) which also have 'y' entered in the sitting on the sea field ('On.sea').
2. If there are no instances of this mismatch then the following message is outputted to the console: '*no flight heights entered for birds recorded as 'on.sea*'' .
3. If this mismatch is detected then the following is displayed:
    a. '*There are # occasions when flight height was entered for a bird on the sea*.'
    '*check these rows of the Observations data:*'
    b. followed by the row numbers identified as errors.

All of the variables loaded into R and created by the script can be viewed via the 'Environment' tab in RStudio (e.g. by clicking on the variable name in the Environment tab, or entering 'View(*object name*)' at the command prompt in the R console). An optional output can also be generated if the variable 'save.top.tail' near the beginning of the script is given a value of 1 (this is set to 0 by default). Setting this to 1 causes the first and last 5 rows of each sheet to be outputted to the console and also to a csv file. The file names are prefixed a 'Summary_info_' followed by the data sheet (Summary, Observations or Track). These csv files can be opened by default using Excel.

Appendix 1. Data structure

Table 1 - Summary	Ship	Digital still	Digital video	Data type	Note
Master.id	Y	Y	Y	Combination	Site name and year (e.g. Beatrice_2016, robin_rigg_2015, etc.). This is part of key to link tables.
Survey.id	Y	Y	Y	Text	Two parts, first digit is survey number (e.g. Of series of monthly surveys), second digit is day of survey (e.g. If it takes >1 day then this is captured). Third survey, day one = 3_1 etc. This is part of key to link tables.
Survey.target	Opt	Opt	Opt	Text	Object of investigation e.g. 'Seabirds', 'seabirds_marine_mammals', 'marine_mammals'
Survey.type	Y	Y	Y	Text	Combination of platform and data collection method (e.g. Plane_digital_video, plane_digital_still, ship_observer, etc.)
Altitude	Y	Y	Y	Decimal	Of plane or boat viewing platform above msl (m). Average or target height.
Geodetic.ref.system	Y	Y	Y	Text	Geodetic reference system e.g. 'Wgs_84'
Image.width	Na	Y	Y	Decimal	Only for digital aerial (m). Refers to individual camera. Combined width in 'strip.width'
Strip.width	Y	Y	Y	Decimal	Transect (ship or plane) = total width, e.g from all cameras, or total width of observed band (ship, both sides); digital image grid = individual image area
Image.resolution	Na	Y	Y	Decimal	Only for digital aerial surface (cm) gsd
Study.area	Y	Y	Y	Decimal	Total area covered by survey (i.e. Total within site boundary including area between images, transects etc.)
Observed.area	Y	Y	Y	Decimal	Digital = total area of images, ship = area within strip.width for all transects
Year.start	Y	Y	Y	Integer	Four digits
Month.start	Y	Y	Y	Integer	Two digits
Day.start	Y	Y	Y	Integer	Two digits
Year.end	Y	Y	Y	Integer	Four digits
Month.end	Y	Y	Y	Integer	Two digits
Day.end	Y	Y	Y	Integer	Two digits
Hour.start	Y	Y	Y	Integer	Two digits (24hr clock) utc
Minute.start	Y	Y	Y	Integer	Two digits

Table 1 - Summary	**Ship**	**Digital still**	**Digital video**	**Data type**	**Note**
Second.start	Y	Y	Y	Decimal	Two digits (minimum)
Hour.end	Y	Y	Y	Integer	Two digits (24hr clock) utc.
Minute.end	Y	Y	Y	Integer	Two digits
Second.end	Y	Y	Y	Decimal	Two digits (minimum)
Snapshot.window.length	Y	Na	Na	Decimal	Only for ship
Snapshot.window.width	Y	Na	Na	Decimal	Only for ship
Snapshot.interval	Y	Na	Na	Text	Only for ship: "distance" or "time"
Snapshot.frequency	Y	Na	Na	Decimal	Only for ship, metres or seconds (depends on 'snapshot.separator')
No.sides	Y	Na	Na	Integer	Only for ship
No.observers	Y	Na	Na	Integer	Only for ship
No.cameras	Na	Y	Y	Integer	Only for digital aerial
Notes	Opt	Opt	Opt	Combination	Extra information if required

Table 2 - Track	**Ship**	**Digital still**	**Digital video**	**Data type**	**Notes**
Master.id	Y	Y	Y	Combination	Site name and year (e.g. Beatrice_2016, robin_rigg_2015, etc.). This is part of key to link tables.
Survey.id	Y	Y	Y	Text	Two parts, first digit is survey number (e.g. Of series of monthly surveys), second digit is day of survey (e.g. If it takes >1 day then this is captured). Third survey, day one = 3_1 etc. This is part of key to link tables.
Transect.id	Y	Y	Y	Integer	Numerical transect identifier
Transect.length	Y	Y	Y	Decimal	Length of transect
Camera.id	Na	Y	Y	Integer	Camera id reference number
Image.id	Na	Y	Y	Text	If still = unique image id, video = reel.id_frame.id (if multiple cameras then add rows)

Table 2 - Track	**Ship**	**Digital still**	**Digital video**	**Data type**	**Notes**
Image.area	Na	Y	Na	Decimal	Individual image area on surface (still). Not required for video as strip.width (summary sheet) and transect.length can be combined to calculate survey coverage. For boat this may be area surveyed in time interval (as per esas)
Image.quality	Na	Y	Y	Integer	Logical: 0/1 = poor/good
Latitude	Y	Y	Y	Decimal	Position of vessel / plane (centre of image) using geodetic ref identified in summary sheet.
Longitude	Y	Y	Y	Decimal	Position of vessel / plane (centre of image) using geodetic ref identified in summary sheet.
Year	Y	Y	Y	Integer	Four digit
Month	Y	Y	Y	Integer	Two digits
Day	Y	Y	Y	Integer	Two digits
Hour	Y	Y	Y	Integer	Two digits (24hr clock) utc
Minute	Y	Y	Y	Integer	Two digits
Second	Y	Y	Y	Decimal	Two digits (minimum)
On.survey	Y	Y	Y	Logical	Logical, indicates on/off survey (e.g. "y" or "n" or 0/1)
Heading	Y	Y	Y	Decimal	Direction of travel of vessel/plane, decimal degrees
Seastate	Opt	Opt	Opt	Integer	Standard coding (e.g. Use esas)
Visibility	Opt	Opt	Opt	Integer	Standard coding
Sunglare	Opt	Opt	Opt	Integer	Not usually collected by ships, but no reason not to. Intensity of the reflection of the sun in regular intervals (for details see additional chart) and after a change of direction or transect, ascertained for the analysed part of the image. 0 = no or very little glare (image not affected) 1 = low amount / intensity of glare which covers less than 25% of image 2 = medium amount / intensity of glare covering less than 50% of image 3 = severe amount or intense of glare covering over 50% of image
Altitude	Na	Y	Y	Decimal	Of plane at time image captured
Speed	Opt	Opt	Opt	Decimal	Snapshot or average
Notes	Opt	Opt	Opt	Text	Optional

Table 3 - Observations	**Ship**	**Digital still**	**Digital video**	**Data type**	**Notes**
Master.id	Y	Y	Y	Combination	Site name and year (e.g. Beatrice_2016, robin_rigg_2015, etc.). This is part of key to link tables.
Survey.id	Y	Y	Y	Text	Two parts, first digit is survey number (e.g. Of series of monthly surveys), second digit is day of survey (e.g. If it takes >1 day then this is captured). Third survey, day one = 3_1 etc. This is part of key to link tables.
Year	Y	Y	Y	Integer	Four digit
Month	Y	Y	Y	Integer	Two digits
Day	Y	Y	Y	Integer	Two digits
Hour	Y	Y	Y	Integer	Two digits (24hr clock)
Minute	Y	Y	Y	Integer	Two digits
Second	Y	Y	Y	Decimal	Two digits (minimum)
Transect.id	Y	Y	Y	Integer	Numerical transect identifier
Image.id	Na	Y	Y	Text	Still = unique image id, video = reel id (if multiple cameras then add rows)
Object.name	Y	Y	Y	Text	Species etc. Can include abiotic objects
Object.id	Y	Y	Y	Integer	Species coding - euring
Object.id.confidence	Y	Y	Y	Text	Quality of object id. Can also be notes for abiotic if necessary
No.individuals	Y	Y	Y	Integer	If observation refers to flock, pod, etc.
Latitude	Y	Y	Y	Decimal	Of object
Longitude	Y	Y	Y	Decimal	Of object
Object.height.estimate	Y	Y	Y	Decimal	Bird flight height (central value)
Object.height.min	Y	Y	Y	Decimal	Bird flight height (lower estimate)
Object.height.max	Y	Y	Y	Decimal	Bird flight height (upper estimate)
Height.method	Y	Y	Y	Integer	0 = observer judgement, 1 = laser range finder (boat), 2 = parallax, 3 = body length, 4 = lidar

Table 3 - Observations	Ship	Digital still	Digital video	Data type	Notes
Object.heading	Na	Y	Y	Integer	Only for digital aerial, degrees
On.sea	Y	Y	Y	Text	Logical "y", "n"
In.flight	Y	Y	Y	Text	Logical "y", "n"
Behaviour	Opt	Opt	Opt	Integer	Standard coding (e.g. Esas)
Surfacing	Opt	Opt	Opt	Integer	Added field bird/mammal on the water surface. For birds only relevant if they show diving behaviour 0 = surface has been breached 1 = below the surface 2 = not clearly visible; if not applicable record leave blank
In.snapshot	Y	Na	Na	Text	Logical "y", "n"
Body.length	Na	Y	Y	Decimal	Measurement from the beak to the end of the tail
Wing.span	Na	Y	Y	Decimal	Measurement of the wing span
Animal.age	Opt	Opt	Opt	Text	Standard coding (e.g. A = adult, im = immature, juv =juvenile)
Animal.sex	Opt	Opt	Opt	Text	M/ f
Plumage	Opt	Opt	Opt	Text	Standardised specification of the plumage of the bird
Association	Opt	Opt	Opt	Text	With vessel, or other structure; logical "y", "n"
Association.note	Opt	Opt	Opt	Text	Details if 'association' = y (e.g. With fishing vessel, buoys, etc.)
Distance.from trackline	Y	Na	Na	Decimal	Only for ship for animals on sea surface, can be distance band or actual distance if estimated
Notes				Text	Optional