SPI-Birds:

STANDARD DATA QUALITY CHECK

Version 1.0

April 8, 2021

Produced by:

SPI-Birds team

Stefan J.G. Vriend, Liam D. Bailey, Chloé R. Nater,
Chris Tyson, Zuzana Zajková, Antica Culina,
& Marcel E. Visser

Table of Contents

SECTION 1

Introduction

This document is created as part of the **SPI-Birds** project. **SPI-Birds** aims to create a global network of Studies on Populations of Individuals - Birds (**SPI-Birds**), with the aim to archive data, improve data accessibility and transparency, and to facilitate collaboration. Within this project, we are building robust code pipelines that convert data stored in different formats (i.e. primary data) into data in a standard format (that consists of Brood data, Capture data, Individual data, Location data). This standard format (version 1.1) is aimed at facilitating greater collaboration by allowing data from multiple populations to be easily collated and compared.

In this document, we outline the standard data quality checks which are carried out as a part of the **SPI-Birds** workflow. The aim of the data quality check is to increase the integrity of the data by highlighting all the values within the dataset that are unlikely or impossible. As quality checks are performed on data that have been already converted into the standard format, and because all the datasets are checked in the same way, homogeneity in data quality across different datasets is further increased. Quality checks identify suspicious data records and flag them as *potential errors* or *warnings*. *Potential errors* are values that are considered impossible (e.g., negative values for clutch size) and *warnings* are values that are possible but are considered highly unlikely (e.g., a value for clutch size that is twice as large as a mean clutch size for a species).

The output of the standard data quality checks is a detailed quality check

report. Whenever a record is flagged as a 'potential error' or 'warning', a line is added to the report with information on the type of check that was violated and the row number of the corresponding record. The row number refers to the column Row in the corresponding data table in the standard format and does not refer to the row number in the primary data. In addition, the quality check procedure adds two new columns (Warning and Error) to each of the four data tables in the standard format to allow data users to easily identify potentially spurious records.

The standard data quality checks are run on three occasions. First, when the SPI-Birds team has created a new tailored pipeline and derived the data in the standard format, the quality checks are run on these standard data. This initial quality report is shared with the data owner. Second, when pipelines have been re-run following an update of the primary data by the data owner (e.g., when a new year of breeding information has been collected), the quality checks are re-run and the report shared with the data owner again. Data owners can decide to cross-check the records highlighted by the quality checks with their field notes, and if possible, correct some of the flagged entries (see 1.1 Verification of flagged records). Third, when a user's data request is approved, the quality checks are run on the requested data, and the corresponding quality check report sent to the user alongside the requested data.

1.1 Verification of flagged records

Some of the flagged records may be confirmed by a data owner to be true observations. We do not want these verified values to be flagged each time a new quality check is re-run. To avoid this, we have implemented an 'approve-listing' procedure that will prevent validated records (flagged but subsequently verified by the data owner) from appearing in future quality check reports.

SECTION 2

Check overview

The standard data quality checks are continuously updated to improve existing checks and include new checks. The following checks are currently included in **version 1.0** of the standard data quality checks.

2.1 Brood data

A more detailed description of *Brood data* checks can be found in section 3.

CheckID	Description
B1	Check format of each column in *Brood data*
B2	Compare `ClutchSize_observed` and `BroodSize_observed`
B3	Compare `BroodSize_observed` and `NumberFledged_observed`
B4	Compare `LayDate_observed` and `HatchDate_observed`
B5	Compare `HatchDate_observed` and `FledgeDate_observed`
B6	Compare breeding variables against reference values
B7	Compare `BroodSize_observed` and the number of chicks recorded in *Capture data*
B8	Check that `BroodID` is unique within populations
B9	Check that the order of `ClutchType_observed` is correct
B10	Compare the species of the parents of a brood
B11	Compare the species of the parents and the species of the brood
B12	Compare the species of the brood and the species of the chicks
B13	Check the sex of the mother
B14	Check the sex of the father

2.2 Capture data

A more detailed description of *Capture data* checks can be found in section 4.

CheckID	Description
C1	Check format of each column in *Capture data*
C2	Compare capture variables against reference values
C3	Check that chick age values are within the range of expectations
C4	Check that adults caught on a nest are recorded as the parents of that nest
C5	Check the age of subsequent captures

2.3 Individual data

A more detailed description of *Individual data* checks can be found in section 5.

CheckID	Description
I1	Check format of each column in *Individual data*
I2	Check that `IndvID` is unique within populations
I3	Check that `BroodIDLaid` and `BroodIDFledged` match the brood in *Capture data*
I4	Check for uncertainty in `Sex_calculated`
I5	Check for uncertainty in `Species`
I6	Check that all individuals in *Individual data* are also present in *Capture data*

2.4 Location data

A more detailed description of *Location data* checks can be found in section 6.

CheckID	Description
L1	Check format of each column in *Location data*

3

SECTION

Brood data

3.1 Check B1

This check is redundant and will be removed.

3.2 Check B2

Compare clutch size (`ClutchSize_observed`) and brood size (`BroodSize_observed`) within each brood. We expect that clutch size is larger than or equal to brood size. Broods that do not meet this assumption are flagged. Broods that were not subject to experimental manipulation are flagged as *potential error*. Broods that were subject to experimental manipulation are flagged as *warning*, as clutch size might be smaller than brood size as a result of the experimental procedure.

3.3 Check B3

Compare brood size (`BroodSize_observed`) and number of fledglings (`NumberFledged_observed`) within each brood. We expect that brood size is larger than or equal to number of fledglings. Broods that do not meet this assumption are flagged. Broods that were not subject to experimental

manipulation are flagged as *potential error*. Broods that were subject to experimental manipulation are flagged as *warning*, as brood size might be smaller than number of fledglings as a result of the experimental procedure.

3.4 Check B4

Compare lay date (`LayDate_observed`) and hatch date (`HatchDate_observed`) within each brood. We expect that lay date is earlier than hatch date. Expected length of incubation period is not considered (e.g., hatch date one day after lay date would not be flagged). Broods that do not meet this assumption are flagged as *potential error*. This check does not flag records as *warning*.

3.5 Check B5

Compare hatch date (`HatchDate_observed`) and fledge date (`FledgeDate_observed`) within each brood. We expect that hatch date is earlier than fledge date. Expected length of fledging period is not considered (e.g., fledge date one day after hatch date would not be flagged). Broods that do not meet this assumption are flagged as *potential error*. This check does not flag records as *warning*.

3.6 Check B6a-d

Compare breeding parameters (i.e., `ClutchSize_observed`, `BroodSize_observed`, `NumberFledged_observed`, `LayDate_observed`) against expected upper and lower limits of each variable. These reference values are generated by population- and species-specific data if the number of observations is sufficiently large ($n \geq 100$). Population-species combinations that have too few records are evaluated for negative values or other values that are considered impossible.

For `ClutchSize_observed`, `BroodSize_observed` and `NumberFledged_observed`, records are considered unusual if they are larger than the 99^{th} percentile and will be flagged as *warning*. Records are considered impossible if they are negative or larger than 4 times the 99^{th} percentile and will be flagged as *potential error*. For population-species combinations that have too few observations ($n < 100$), records are considered impossible if they are negative and will be flagged as *potential error*.

For LayDate_observed, records are considered unusual if they are earlier than the 1st percentile or later than the 99th percentile and will be flagged as *warning*. Records are considered impossible if they are earlier than January 1st or later than December 31st for the current breeding season and will be flagged as *potential error*. For population-species combinations that have too few observations (n < 100), records are considered impossible if they are earlier than January 1st or later than December 31st for the current breeding season and will be flagged as *potential error*.

3.7 Check B7

Compare brood size (BroodSize_observed) of a specific brood with the number of chicks recorded in Capture data. We expect that these numbers should be equal. Records where BroodSize_observed is larger than the number of chicks recorded in Capture data are flagged as *warning*, because chicks might have died before ringing and measuring. Records where BroodSize_observed is smaller than the number of chicks recorded in Capture data are flagged as *potential error*, because this should not be possible.

3.8 Check B8

Check that brood identity (BroodID) is unique within a population. Records with non-unique brood identifiers are flagged as *potential error*. Brood identities are not required to be unique across populations. This check does not flag records as *warning*.

3.9 Check B9

Check that the order of clutch types (ClutchType_observed) per breeding female per season is correct. Replacement and second clutches can occur in any order, but never before first clutches. First clutches that are not the earliest brood in the season are flagged as *potential error*. This check does not flag records as *warning*.

3.10 Check B10

Check that the parents of a brood are of the same species. We expect that the parents of the majority of broods are marked as the same species. Common,

biologically possible multi-species broods (i.e., PARMAJ – CYACAE and FICHYP – FICALB) are flagged as *warning*. Other combinations of species are flagged as *potential error*.

3.11 Check B11

Check whether the parents of a brood and the brood itself are of the same species. We expect that the parents and their broods are marked as the same species. Broods with a combination of species for which brood fostering is known to exist (i.e., PARMAJ – CYACAE and FICHYP – FICALB) are flagged as *warning*. Other combinations of species are flagged as *potential error*.

3.12 Check B12

Check that the chicks in a brood and the brood itself are of the same species. We expect that the chicks and the broods are marked as the same species. Broods with a combination of species for which brood fostering is known to exist (i.e., PARMAJ – CYACAE and FICHYP – FICALB) are flagged as *warning*. Other combinations of species are flagged as *potential error*.

3.13 Check B13

Check that the individual listed as the mother of a brood is female. Broods where the mother is male are flagged as *potential error*. This check does not flag records as *warning*.

3.14 Check B14

Check that the individual listed as the father of the brood is male. Broods where the father is female are flagged as *potential error*. This check does not flag records as *warning*.

4

SECTION

Capture data

4.1 Check C1

This check is redundant and will be removed.

4.2 Check C2a-b

Compare chick and adult capture measurements (i.e., `Mass`, `Tarsus`) against expected upper and lower limits of each variable. For `Mass` in adults and `Tarsus` in adults and chicks, reference values are generated by population- and species-specific data if the number of observations is sufficiently large ($n \geq 100$). Population-species combinations that have too few records are evaluated for negative values or other values that are considered impossible. For Mass in chicks, reference values are calculated for each age (in days). A logistic growth model is fitted to determine reference values for each day. If the model fails, reference values are generated per age if the number of observations for chicks of that age is sufficiently large ($n \geq 100$). Population-species combinations that have too few records are evaluated for negative values or other values that are considered impossible.

Records are considered unusual if they are smaller than the 1st percentile or larger than the 99^{th} percentile and will be flagged as *warning*. Records

are considered impossible if they are negative or larger than 4 times the 99^{th} percentile and will be flagged as *potential error*. For population-species combinations that have too few observations (n < 100), records are considered impossible if they are negative and will be flagged as *potential error*.

4.3 Check C3

Check that chick age values (in number of days since hatching) are within the expected duration of the nestling period. Possible values currently include any number of days between 0 and 30 days since hatching. Values that are outside this range are flagged as *potential error*. This check does not flag records as *warning*.

4.4 Check C4

Check that adults captured on a nest are listed as the parents of that nest in *Brood data*. As non-parent adults can be caught close to the nest, adults caught on a nest (i.e., `LocationType` of capture is a nest, 'NB') that are not marked as the parents of that nest are flagged as *warning*. This check does not flag records as *potential error*.

4.5 Check C5

Check that the observed age of chronologically ordered captures is correct. The age recorded in an individual's subsequent capture is expected to be equal when the capture was in the same year or increase when the capture was in a later year. Records of an individual caught as an adult before records of the same individual caught as a chick are flagged as *potential error*. Other records where the observed age of a capture is larger than the age of a subsequent capture are flagged as *warning*.

SECTION 5

Individual data

5.1 Check I1

This check is redundant and will be removed.

5.2 Check I2

Check that individual identities (`IndvID`) are unique. Records with non-unique individual identifiers within populations are flagged as *potential error*. Individual identities are not required to be unique across populations. This check does not flag records as *warning*.

5.3 Check I3

Check that brood identities for an individual caught as a chick (`BroodIDLaid`, `BroodIDFledged`) match the correct nest in Capture data. Chicks caught on a nest that are not associated with corresponding brood identities are flagged as *potential error*. This check does not flag records as *warning*.

5.4 Check I4

Check for uncertainty in the sex of an individual (`Sex_calculated`). Individuals that have been recorded as both male ('M') and female ('F') in *Capture data* are marked as conflicted ('C') by the pipeline and flagged as *potential error* by the check. This check does not flag records as *warning*.

5.5 Check I5

Check for uncertainty in the species of an individual (`Species`). Individuals who have been recorded as different species in Capture data are marked as conflicted ('CCCCCC') by the pipeline and flagged as *potential error* by the check. This check does not flag records as *warning*.

5.6 Check I6

Check that all individuals recorded in *Individual data* also appear in *Capture data*. *Individual data* is usually a direct product of *Capture data*, so this should never occur. Should it occur nonetheless, it is an indication of problems in the underlying pipelines and missing individuals are flagged as *potential error*. This check does not flag records as *warning*.

SECTION 6

Location data

6.1 Check L1

This check is redundant and will be removed.