Report 1 - Data Availability

Anna Krystalli

24 December 2014

So far all mating system data, supplied by Andras and Gavin, as well as additional ecological etc data supplied by Tamas and the clutch size data published in the PLOS paper (DOI: 10.1371/journal.pbio.0060303) have been compiled onto the master data set.

Details of the datasets can be found on the data log on the project dropbox folder: <https://docs.google.com/spreadsheets/d/1xIp9pmQbqIx_ZuQlTC609TI6oQLWb6vZ0NumA5oYZp4/edit?usp=sharing>

The raw data can be found here: <https://drive.google.com/folderview?id=0B9BMeZ9H2nRxOElvUk1mZzA1clU&usp=sharing>

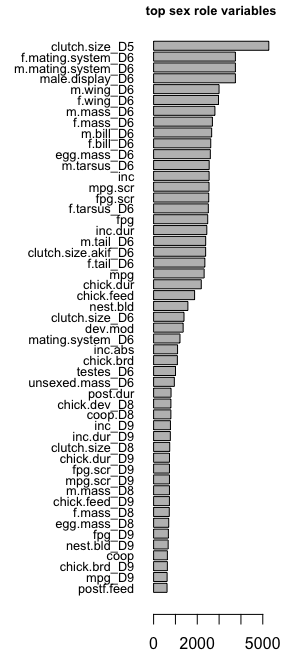
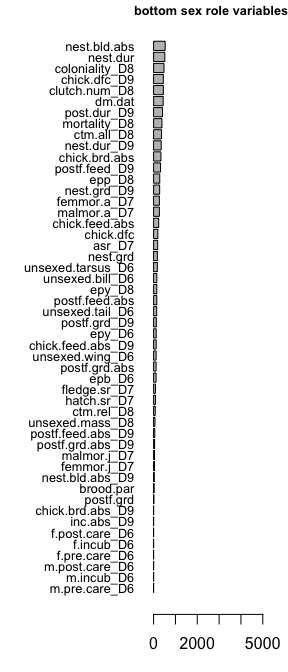
The up to date master spreadshet can always be found here: <https://drive.google.com/file/d/0B9BMeZ9H2nRxa3VOSXF3V2tCLTg/view?usp=sharing>

while descriptions of variables including details of scoring can be found here: <https://drive.google.com/file/d/0B9BMeZ9H2nRxeFlHamhfalBWZjA/view?usp=sharing>

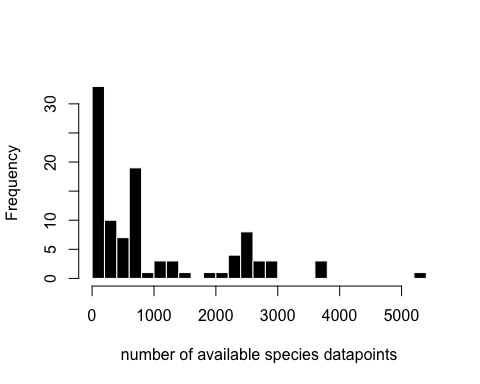
So far, collation has resulted in a total of 142,461 datapoints spread over 146 variables. 45 variables are related to data quality control (confidence), references and data sources leaving a total of 101 variables for analysis of sex roles so far.

Ornament data are yet to be added as these still have significant gaps. Global distribution data have been applied for from BirdLife and these should be available soon (hopefully within the next few days although it may now be after the holidays). These will then be used to extract environmental data for each species range and determine mean and variance values for each. The Bioclim environmental data have been downloaded so we are now only awaiting the BirdLife data to continue.

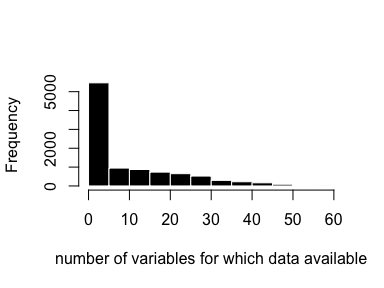
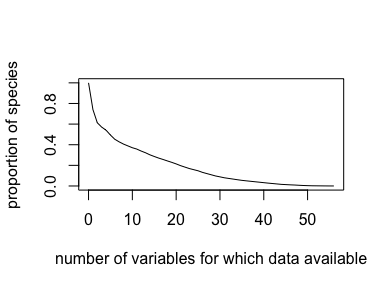
A more detailed report on the distribution of values across individual variables and repeatability of values across observers will follow shortly.

Frequency distribution of number of datapoints available per variable. The majority of variables have less than 1000 datapoints (most < 200) with a second small peak of variables of 2000-3000 datapoints.



For 2586 species there is currently no sex role variable data at all, with the maximum number of variables for an individual species being 56. The following plots show the frequency distribution of data availability across variables for individual species (top) and the proportion of data availability across species as a function of the number of variables for which data is available (bottom).

The top and bottom 20 species with respect to data availability are indicated below

## [1] "top 20"

## species variables  
## 6453 Parus\_major 56  
## 629 Fulmarus\_glacialis 56  
## 9308 Agelaius\_phoeniceus 55  
## 8427 Dendroica\_caerulescens 54  
## 6460 Parus\_caeruleus 54  
## 1534 Calidris\_maritima 54  
## 9248 Fringilla\_coelebs 53  
## 8925 Passer\_domesticus 53  
## 8700 Sialia\_mexicana 53  
## 7962 Ficedula\_hypoleuca 53  
## 1468 Charadrius\_alexandrinus 53  
## 7229 Troglodytes\_aedon 52  
## 5367 Malurus\_cyaneus 52  
## 1507 Philomachus\_pugnax 52  
## 1394 Haematopus\_ostralegus 52  
## 9670 Melospiza\_melodia 51  
## 8778 Turdus\_merula 51  
## 6554 Delichon\_urbicum 51  
## 1535 Calidris\_mauri 51  
## 328 Lagopus\_lagopus 51  
## 9986 Cardinalis\_cardinalis 50

## [1] "bottom 20"

## species variables  
## 2 Rhynchotus\_maculicollis 0  
## 5 Nothocercus\_julius 0  
## 6 Nothocercus\_nigrocapillus 0  
## 8 Nothoprocta\_curvirostris 0  
## 10 Nothoprocta\_taczanowskii 0  
## 12 Nothura\_chacoensis 0  
## 14 Nothura\_minor 0  
## 17 Taoniscus\_nanus 0  
## 26 Crypturellus\_berlepschi 0  
## 28 Crypturellus\_bartletti 0  
## 29 Crypturellus\_atrocapillus 0  
## 30 Crypturellus\_casiquiare 0  
## 33 Crypturellus\_duidae 0  
## 35 Crypturellus\_kerriae 0  
## 37 Crypturellus\_brevirostris 0  
## 38 Crypturellus\_ptaritepui 0  
## 47 Eudromia\_formosa 0  
## 61 Talegalla\_cuvieri 0  
## 65 Eulipoa\_wallacei 0  
## 72 Megapodius\_geelvinkianus 0

Distribution of available data across families. The first two figures show total number of datapoints across all variables available for each family. The last two image panels show a cross-tabulation of species-level variable data availability across families ie. the intensity of colour in each pixel indicates the number of species in a particular family that have data points across the number of variables indicated on the x axis. Red lines indicate the level below which families have zero data points.

