A multi-scale assessment of habitat response of breeding farmland waders in Shetland, in order to enhance conservation outcomes

Anthony Waite

19 August 2020

# Abstract

This is the very first part of the book.e

# Acknowledgements

Something about me!

# Materials and Methods

## Study site

The Shetland archipelago was used as the study area. The remote islands of Foula and Fair Isle were removed from the study area.

## Breeding wader survey data

Two independent sets of survey data were used to provide spatial abundance for breeding waders. The Shetland Breeding Bird Survey (SBBS) is a citizen science project that is overseen by the Shetland Biological Records Centre. The SBBS is annual survey undertaken since 2002, whereby local volunteers select and walk a 1km Ordnance Survey GB square (OSGB36). It is important to note that to encourage volunteer uptake, these squares are not randomly allocated in space; typically the location of the survey is selected by the volunteer. Volunteers walk either a 2km transect around their chosen square, or two 1km transects that bisect the square. All breeding bird species observed 100m either side of the transect are recorded. All transects were walked twice, typically April to mid-May and mid-May to late June. All surveys were undertaken before 9am and in good weather (wind force four or less, and dry). The annual abundance recorded for each breeding species within a given square is the maximum of count from the two visits. n=?

The RSPB have undertaken annual field based surveys across a number of well-known Shetland breeding wader sites since 1987 (Campbell and Ellis [1987](#ref-Campbell1)). The initial survey covered 33,000 ha or 293 sites of *inbye* land; this is land that is lowland rough grassland, and is typically not farmed intensively. From the 1987 survey, 6900 ha or 122 sites were identified. The survey data used in this study was recorded between 2013 and 2018 and are taken from up-to 29 sites, a subset of the original 1987 sites. The survey method (O’Brien and Murray [1998](#ref-OBrien1998-iy)) required observers to walk to within 100m of every part of the survey site and to complete their survey between dawn and 12:00 hours. Three visits were made, one in each of the periods 15 April-30 April, 1 May-21 May and 22 May-18 June. Successive visits were separated by at least a week. Cold, wet or windy weather conditions were avoided. Each wader observed was recorded against a specific field. The method for calculating the breeding wader abundance recorded was species specific: Lapwing, Oystercatcher and Redshank abundance was recorded as the total numbers of birds seen in each field; Snipe—the number of drumming plus chipping birds in each field; for Curlew the mean count of individual Curlew (MNI) over three standard survey visits to each site was calculated, and the numbers of nesting pairs was calculated as 0.71 MNI + 0.10 (Grant et al. [2000](#ref-Grant2000-ns)).

## Estimating detectability

The two surveys in question assumed perfect detectability of the survey species. Due to the imprecise nature of field surveys this could result in a bias in species detectability, which could further result in inaccurate population estimates. To account for this bias, there are a number of sophisticated statistical techniques such as *distance sampling* that account for this, in order to reduce the standard error associated with a population estimate (Newson et al. [2008](#ref-Newson2008-oq)). Such techniques typically require more structured survey techniques such as precisely measuring the distance and angle, between the target and the transect line. Unfortunately this data was not available for the two survey datasets used within the study. Inorder to generate a probabiliy of detection, and therefore account for detectability bias, the r-package unmarked (Fiske and Chandler [2011](#ref-Fiske2011-xf)) was used to generate a probability of detection for each species. Unmarked adopts a generative modelling process whereby observations are modelled through a combination of (1) a state process determining abundance at each site and (2) a detection process that yields observations conditional on the state process. Probability of detection (detectability) was modelled using unmarked for each species by using count data across multiple sites within and years, using day of year as a covariate. An average was then taken across years to give an average detectability for each species.

## Exploratory Data Analysis

An exploratory data analysis was undertaken on the SBBS dataset as per the protocol outlined in Zuur (Zuur, Ieno, and Elphick [2010](#ref-Zuur2010-kp)). In addition survey locations were plotted spatially across each survey year in order to gauge survey effort. Outliers were removed etc etc

## Environmental covariates

Environmental covariates were generated with values spatially assigned to each Ordnance Survey (OS) grid cell, at 1km resolution. The grid reference was generated to match the SBBS survey data. The spatial extent of the covariate grid excluded the remote islands of Foula and Fair Isle in order to reduce computational time when fitting spatial abundance models. The following environmental covariates were quantified for the OS squares and periods studied:

### EUNIS habitat classification

The EUNIS habitat classification is a European wide system for habitat classification (“EUNIS” [2014](#ref-noauthor_2014-bw)). For the purposes of EUNIS, a ‘habitat’ is defined as:

*‘a place where plants or animals normally live, characterized primarily by its physical features (topography, plant or animal physiognomy, soil characteristics, climate, water quality etc.) and secondarily by the species of plants and animals that live there’*

Each habitat type is identified by specific codes and a description. The classification of terrestrial habitat is hierarchical and comprises three levels. For the purpose of this study the level 1 classification was used so as not to overdisperse the response variable (breeding wader abundnace). The EUNIS dataset is available as a categorical GeoTiff file, at a resolution of 10x10m. The level 1 EUNIS categorisation is shown in Table: 1. (put this in the appendix?)

Table 1: EUNIS habiat classificaiton

|  |  |
| --- | --- |
| L1 Classification | Habitat Description |
| A | Marine habitats |
| B | Coastal habitats |
| C | Inland surface waters |
| D | Mires, bogs and fens |
| E | Grasslands and lands dominated by forbs, mosses or lichens |
| F | Heathland, scrub and tundra |
| G | Woodland, forest and other wooded land |
| H | Inland unvegetated or sparsely vegetated habitats |
| I | Regularly or recently cultivated agricultural, horticultural and domestic habitats |
| J | Constructed, industrial and other artificial habitats |
| K | Montane habitats |
| X | Habitat complexes |

A raster file containing the EUNIS habitat classification for Scotland in 2019 was cropped to the extent of the Shetland archipelago, and the percentage coverage of habitat classes *D, E and F* were calculated for each OS square (n=3000?).

### Improved grassland

Across Shetland crofters have *improved* parcels of low-lying grassland inorder to take an annual crop of haylage or silage. It is typically used as fodder for sheep and cattle during the long winter season. This habitat is an important environmental covariate for studying breeding waders, as it has been suggested that improved grasslands are an important feeding habitat for nesting waders and their chicks (McCallum et al. [2018](#ref-McCallum2018-gx)). Improved grassland is not within the EUNIS habitat classification, therefore remotely sensed satellite image data were used together with supervised machine learning techniques (Abdi [2020](#ref-Abdi2020-aa)) to generate a spatial classification for Shetland improved grassland.

A Sentinel-2 satellite image of the Shetland archipelago, comprising 11 different spectral bands (“Sentinel-2” [2020](#ref-sentinel-ovr)) , was used to produce an improved grassland raster layer. This was achieved by using the r programming language (R Core Team [2019](#ref-R-lang)) to train both a support vector machine (SVM) and extreme gradient boosting (XGBoost) machine learning algorithms. The algorithms used in conjunction with the tidymodels package (Kuhn and Wickham [2020](#ref-tidymodels)) to classify each raster cell, according to the spectral band values within it. Training datasets comprising polygons were created across five different habitat classes: improved grassland, unimproved grassland, upland, bare peat and crops. 1000 spatial samples were randomly drawn from each class training dataset; a sample comprising a value for each of the 11 spectral bands, for the specific sample location. The aggregate training dataset , comprising 5000 samples, was then split into training and testing data. The training data was subsequently resampled using 5-Fold cross-validation in order to reduce variability of the model classification. The hyperparameters of a radial basis SVM were tuned and the best model was selected based on an accuracy metric validated against the test data split. The model fit was then used to make a prediction for improved grassland over the entire Shetland archipelago.

### Median topsoil pH content

A shape file detailing the median topsoil pH content across Shetland was provided by the James Hutton Institute. A higher soil pH has been shown to be associated with greater earthworm abundance [McCallum2016-jt], which is one of the main food sources of breeding farmland waders and their chicks. The weighted median pH was calculated for each SBBS OS 1km square, using the spatial features package (Pebesma [2018](#ref-spatial-features)) within r.

### Topsoil organic carbon content

A shape file containing the topsoil organic carbon content was downloaded from the James Hutton Institue (Lilly, A and Baggaley, N and Donnelly, D. [2012](#X281cfb5c30af3cb11602389e4468c737eca2860)). The topsoil of grasslands store significant amount of Carbon (Eze, Palmer, and Chapman [2018](#ref-Eze2018-pd)). Therefore fertile grasslands favoured by waders may have significant carbon stored within the topsoil, whilst grassland that is heavily grazed by livestock may have relatively low carbon storage.The mean topsoil organic carbon content was calculated for each SBBS OS 1km square, using the spatial features package (Pebesma [2018](#ref-spatial-features)) within r.

### Available water capacity

Available water capacity (AWC) is the amount of water a soil can provide for plants, and so is a useful indicator of the ability of soils to grow crops. The amount of moisture in the soil is thought to be an important aspect of wader abundance (Smart et al. [2006](#ref-Smart2006-dt)). A shape file detailing the AWC across Shetland was downloaded from the James Hutton Institute (Gagkas, Z., Lilly, A., Baggaley, N. & Donnelly, D. [2019](#Xc1035fe64dbdd93986e190df212fc806af046b3)), and the mean AWC across each SBBS 1km square was calculated, using the spatial features package (Pebesma [2018](#ref-spatial-features)) within r.

### Bare peatland

Shetland has a significant amount of peatland that has been cut for use as fuel. It is expected that the resulting bare peat is negatively associated with breeding waders. Scottish Natural Heritage have produced a shapefile of bare peat, at a resolution of 10x10m (Blake [2020](#ref-Blake2020-zj)). The percentage cover of bare peat for each SBBS 1km square was calculated, using the spatial features package (Pebesma [2018](#ref-spatial-features)) within r.

### Distance to sea

Proximity of breeding wader territory to the coast as in-land breeding sites could be more sheltered from Shetland’s maritime weather. For each SBBS 1km square the mean distance to the coast was calculated using a raster for the Shetland landmass.

### Elevation

the r package elevatr (Hollister and Tarak Shah [2017](#ref-elevatr)) was used to generate a mean elevation above sea level for each SBBS 1km square.

## Bootstrap analysis of SBBS survey squares

As SBBS survey volunteers were able to choose which squares to survey, they are not randomly allocated. As such this may not be a true representation of the distribution of habitat types across Shetland. In order to analyse any potential bias in the habitat types surveyed, a bootstrap analysis was performed across all OS 1km squares, using the EUNIS habitat categorisation. This was then compared against a bootstrap of surveyed 1km squares from a single SBBS survey year (2018). This provides a quantification of how biased survey samples are with respect to the overall Shetland landmass.

## Wader food response

## Breeding wader density response to environmental covariates

To investigate how breeding wader abundance responds to environmental covariates a generalised additive model (GAM) was used. The approach is an extension of work undertaken to look at breeding Curlew density undertaken by the BTO (Franks et al. [2017](#ref-Franks2017-co)). Species count data from the SBBS was split into two time periods, each covering nine years; 2002-10 and 2011-19. Then for each species and each surveyed square within the two time periods, a mean count was derived for each of the two time periods and joined with the respective environmental covariates, as a function of spatial location.

Breeding wader counts were modelled using a log link function with a *poisson* distribution in R using the gam function in the mgcv package (Wood [2003](#ref-mgcv)). Location coordinates of survey squares were converted to planar OSGB36 national Grid coordinates as normalized easting and northing (in meters). Spatial variability in wader abundance was modelled as a spatial smooth, with the number of knots constrained (k = 20) and environmental covariates included as a parametric term. A different gam was generated for each environmental covariate, in order to explore univariate associations between response and covariate. To explore absolute abundance of breeding wader species per 1km the natural log of the average detectability per species was included as an offset within each gam model. The gam structure is summarised in equation (1).

$$

, where *X* is the environmental covariate, *Y* is species count per surveyed square, *Z* is average species detectability. Associations between response and the spatial effect of environmental covariate were determined by the resulting p value of the gam fit. Only those models with p < 0.05 were considered statistically significant.

Plots of density (abundance per 1km) against each environmental covariate were generated in order to visualise the trend of any association between response and covariate.

## Breeding wader population change as a response to environmental covariates

Given models for breeding wader density response against environmental covariates across two distinct time ranges, a new set of models were created to model the change in abundance between 2002-10 and 2011-19 by using the count in 2011-19 as the response, and the natural log of the 2002-10 count as an offset. It is assumed that detectability remained constant within each square, which in turn means that changes in count reflect changes in density, against a given environmental covariate. Only SBBS squares that were surveyed in both periods and where breeding waders were present in 2002-10 where used, as it is not possible to take the natural log of zero.

the population change ration was then plotted against each environmental covariate, in order to determine the nature of any popultion decline (or increase) as a function of the given environmental covariate.

## Habitat spatial complexity response

In order to investigate an association between breeding wader abundance and the complexity of habitat spatial patterns within the landscape, metrics using information theory (IT) (Nowosad and Stepinski [2019](#ref-Nowosad2019-bo)) were applied to quantify habitat pattern complexity of all SBBS squares across Shetland. Results from each metric were then incorporated into gam density response and population change models as covariates, in order to explore if varying degrees of habitat spatial pattern are statistically correlated with breeding wader abundance.

The r package landscapemetrics (Hesselbarth et al. [2019](#ref-landscapemetrics)) was used to generate all IT metrics, using the EUNIS habitat raster for Shetland as a spatially distributed categorisation of landscape habitat. A description of the different metrics incorporated in the analysis is as follows:

* *Marginal Entropy, H(X)* - measures the compositional diversity of habitats in space; from mono-thematic patterns (lower H(X)) to multi-thematic patterns (large (H(X)). So landscapes with one dominating habitat (for example, peatland on the Shetland island of Yell) have a relatively low H(X), whilst landscapes with multiple different types of habitat that are evenly distributed (such as a village with a patch work of gardens, housing and undeveloped land) have a relatively high H(X). This can be seen in Figure 1, with H(X) shown above each landscape sample.

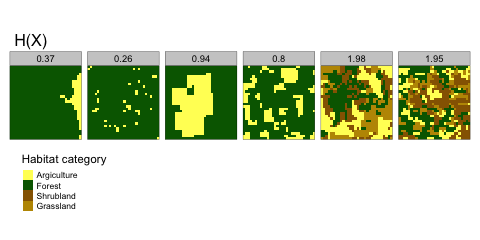


Figure 1: Marginal Entropy

* *Conditional Entropy, H(Y|X)* - quantifies the geometric intricacy of a spatial pattern within a landscape. If habitat type A is predominantly adjacent to habitat type B, H(Y|X) will be relatively low. Conversely if habitat type A is adjacent to many different habitat categories, then H(Y|X) will be relatively high. Figure 2 gives some examples of this.

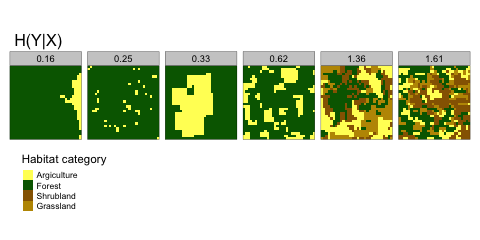


Figure 2: Conditional Entropy

* *Joint Entropy, H(X,Y)* - this provides a measure of the uncertainty in determining the habitat category of a focus cell and an adjacent cell. So landscapes with high H(X,Y) are typically spatially complex with many different habitat types. Note that joint entropy is not capable of distinguishing between patterns that have high spatial aggregation. The variation can be seen in Figure 3 .

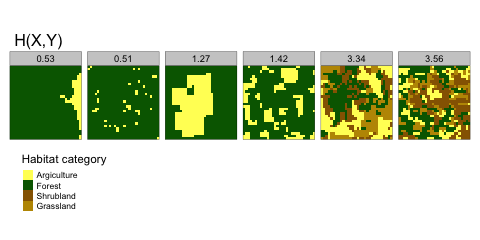


Figure 3: Joint Entropy

* *Relative mutual information, U* - quantifies the degree of aggregation (clumpliness) of spatial habitat categories from fragmented patterns (lower U) to consolidated patterns (higher U). A landscape comprising a loch within a forest would have a relatively high U, whilst a landscape comprising many different crop types spread across many small fields would have low U. Figure 4 gives a mutual entropy landscape

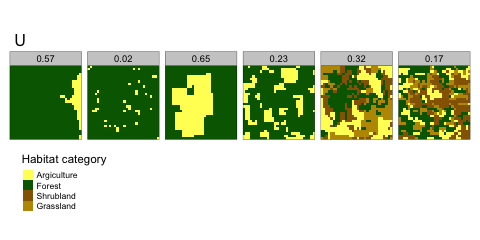


Figure 4: Relative Mututal Entropy

## Wader population abundance modelling

Given the set of environmental covariates for each SBBS 1km square and the associated SBBS count data, a machine learning approach using rand the tidymodels (Kuhn and Wickham [2020](#ref-tidymodels)) package was taken to fit a regression model. The resulting model enabled an annual population to be generated for each breeding wader species across all of Shetland, from the 2002 to 2019, as a function of environmental covariates. In order to fit the model the count data for each species were adjusted according to their mean detectability. All count data for a given SBBS 1km square was joined with the associated environmental covariate data. The joined data were thens plit such that 70% was allocated to model training and 30% to model testing. The training data were then processed so that covariates that have 80% absolute correlations with other covariates are removed; this ensures that possible adverse outcomes due to co-linearity are minimised (Zuur, Ieno, and Elphick [2010](#ref-Zuur2010-kp)). The covariate data within the training dataset were then normalised to a mean of zero and a standard deviation of one. An extreme gradient boosting (xgboost) algorithm was used to fit a regression model using a poisson distribution. 10-fold cross validation was used to resample the data in order to ensure any bias in SBBS survey squares was minimised. Hyperparameters used to fit the xgboost model were selected using a grid containing 20 random variations for each hyperparameter:

* min\_n - the minimum number of data points in a node that are required for the node to be split further)
* trees - the number of trees contained in the ensemble
* mtry - the number of predictors (covariates) that will be randomly sampled at each split when creating the tree models.

A model was fit for each hyperparameter combination and for each species; giving 200 different models. The best model fit for each each species was selected according to the lowest root mean squared error (rmse). The hyperparamters associated with each best fit were then used to further tune the model for a final fit. The rmse for the final fit was then evaluated against the test dataset, for each species. Finally, to generate an annual population account across all over Shetland for each species, the final model fit was used across all SBBS 1km squares (n=?). This gave a mean estimate for abundance for each square, plus a lower and upper confidence interval. Summing up these values across all squares, for each species by year, gave an overall population estimate across Shetland for the years 2002 to 2019.

## Wader spatial distribution modelling

Having generated an abundance prediction for each SBBS square, the spatial abundance can be plotted using the spatial coordinates for each SBBS square. Spatial distributions were plotted for 2002 and 2019 in order to visualise spatial abundance distribution between the two periods, for each species. The net change between the two periods was also plotted so as to visualise the areas of Shetland that have seen a net increase or decrease in species abundance between 2002 and 2019.

## Landscape connectivity

# Results

## Explororatory Data Analysis

### Initial spatial view

### Outliers

### Changes over two periods

### Zero inflation

### Coutn data distribution

## Survey Bootstrap

## Improved grassland classification

Spectral response charts

## Landcover histograms

By EUNIS type

## Density plots

## Environmental covariate response

## IT covariate response

## Abundance spatial distribution

## Abundance population trends

# References

Abdi, Abdulhakim Mohamed. 2020. “Land Cover and Land Use Classification Performance of Machine Learning Algorithms in a Boreal Landscape Using Sentinel-2 Data.” *GISci. Remote Sens.* 57 (1): 1–20.

Blake, Duncan. 2020. “Scotland Bare Peat.” https://github.com/duncansnh/Bare-peat.

Campbell, DR, LH Bird, and PM Ellis. 1987. “Breeding Waders on Croft and Farmland in Shetland.” *Unpublished Report*.

“EUNIS.” 2014. <https://www.eea.europa.eu/data-and-maps/data/eunis-habitat-classification>.

Eze, Samuel, Sheila M Palmer, and Pippa J Chapman. 2018. “Soil Organic Carbon Stock in Grasslands: Effects of Inorganic Fertilizers, Liming and Grazing in Different Climate Settings.” *J. Environ. Manage.* 223 (October): 74–84.

Fiske, Ian, and Richard Chandler. 2011. “Unmarked: An R Package for Fitting Hierarchical Models of Wildlife Occurrence and Abundance.” *Journal of Statistical Software, Articles* 43 (10): 1–23.

Franks, Samantha E, David J T Douglas, Simon Gillings, and James W Pearce-Higgins. 2017. “Environmental Correlates of Breeding Abundance and Population Change of Eurasian Curlew Numenius Arquata in Britain.” *Bird Study* 64 (3): 393–409.

Gagkas, Z., Lilly, A., Baggaley, N. & Donnelly, D. 2019. “Map of Available Water Capacity of Soils in Scotland.”

Grant, M C, C Lodge, N Moore, J Easton, C Orsman, and M Smith. 2000. “Estimating the Abundance and Hatching Success of Breeding Curlew Numenius Arquata Using Survey Data.” *Bird Study* 47 (1): 41–51.

Hesselbarth, Maximilian H. K., Marco Sciaini, Kimberly A. With, Kerstin Wiegand, and Jakub Nowosad. 2019. “Landscapemetrics: An Open-Source R Tool to Calculate Landscape Metrics.” *Ecography* 42: 1648–57.

Hollister, Jeffrey, and Tarak Shah. 2017. *Elevatr: Access Elevation Data from Various Apis*. <http://github.com/usepa/elevatr>.

Kuhn, Max, and Hadley Wickham. 2020. *Tidymodels: Easily Install and Load the ’Tidymodels’ Packages*. <https://CRAN.R-project.org/package=tidymodels>.

Lilly, A and Baggaley, N and Donnelly, D. 2012. “Map of Soil Organic Carbon in Top Soils of Scotland. Map Prepared for EU Project GS-SOIL.”

McCallum, Heather M, Kirsty J Park, Mark G O’brien, Alessandro Gimona, Laura Poggio, and Jeremy D Wilson. 2018. “Soil pH and Organic Matter Content Add Explanatory Power to Northern Lapwing Vanellus Vanellus Distribution Models and Suggest Soil Amendment as a Conservation Measure on Upland Farmland.” *Ibis*, June, 677–87.

Newson, Stuart E, Karl L Evans, David G Noble, Jeremy J D Greenwood, and Kevin J Gaston. 2008. “Use of Distance Sampling to Improve Estimates of National Population Sizes for Common and Widespread Breeding Birds in the UK.” *J. Appl. Ecol.* 45 (5): 1330–8.

Nowosad, Jakub, and Tomasz F Stepinski. 2019. “Information Theory as a Consistent Framework for Quantification and Classification of Landscape Patterns.” *Landsc. Ecol.* 34 (9): 2091–2101.

O’Brien, Mark, and Stuart Murray. 1998. “Estimating the Breeding Wader Populations on Farmland in Northern England in 1993.” *BULLETIN-WADER STUDY GROUP* 85: 60–65.

Pebesma, Edzer. 2018. “Simple Features for R: Standardized Support for Spatial Vector Data.” *The R Journal* 10 (1): 439–46. <https://doi.org/10.32614/RJ-2018-009>.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.

“Sentinel-2.” 2020. <https://sentinel.esa.int/web/sentinel/missions/sentinel-2/overview>.

Smart, Jennifer, Jennifer A Gill, William J Sutherland, and Andrew R Watkinson. 2006. “Grassland-Breeding Waders: Identifying Key Habitat Requirements for Management.” *J. Appl. Ecol.* 43 (3): 454–63.

Wood, S. N. 2003. “Thin-Plate Regression Splines.” *Journal of the Royal Statistical Society (B)* 65 (1): 95–114.

Zuur, Alain F, Elena N Ieno, and Chris S Elphick. 2010. “A Protocol for Data Exploration to Avoid Common Statistical Problems.” *Methods Ecol. Evol.* 1 (1): 3–14.