Summary of UKBMS data capture, processing, validation and reporting

Version: 31st May 2018

Data capture

The primary method for capturing UKBMS data, including the Wider Countryside Butterfly Survey, is through the online capture system available at www.ukbms.org/mydata. This includes site details (e.g. location, habitat and management information), species counts through transect walks and other survey methods (e.g. timed counts, egg/larval counts).

A proportion of data is also captured via the Transect Walker software package or via spreadsheets.

Data is processed on an annual basis. The majority of data is from surveys conducted in the previous summer, but data from previous years is also often collated. All data is processed in the same way.

Standardisation and harmonisation of the UKBMS dataset

All UKBMS data is collated into a single dataset to enable analysis and reporting. As of 2018, the dataset comprises over 7 million butterfly counts. Data is standardised to conform the UKBMS database structure, including: standardised species nomenclature, data integrity checks to ensure that all mandatory information is captured, valid date and time information, accurate geographic location information.

Data verification

The UKBMS online data capture system is built using the Indicia software tools and links to the iRecord verification system (www.brc.ac.uk/irecord) to enable review of the data by experts approved by Butterfly Conservation or other National Recording Schemes (for records for non-lepidoptera). To support verification, iRecord applies automated data checks against known species distributions (e.g. derived from the Butterflies for the New Millennium recording scheme) and timing of adult flight periods. Experts can use these checks and other information to confirm observations.

The UKBMS online data capture system (www.ukbms.org/mydata) also provides data summaries to enable UKBMS Branch Co-ordinators to review all transect data for their area and make corrections.

Further review and correction is undertaken by staff at Butterfly Conservation and the Centre for Ecology & Hydrology at the end of each field season, including the following checks that are discussed with Branch Co-ordinators and/or transect recorder:

- Counts outside of known distribution
- Counts outside of the standard flight period for a species
- Species newly recorded on a transect site
- Species recorded on a transect site after being absent for more than five years
- Potentially data input errors or misidentifications. All counts of specialist butterfly species are closely scrutinised. Summary tables for generalist species are reviewed for anomalies.

Transect visits which are undertaken outside the criteria for butterfly activity (e.g. based on weather conditions and time of day) are flagged and excluded from the main data analyses; data is retained within the database for use in other analysis.

Data analysis

4a. Classification of separate generations for bivoltine species

For bivoltine species, separate generations are identified by defining the time of year where there is a gap between generations. Classification of generations is supported by visual inspection of the seasonal pattern of counts through the season at each transect site.

4b. Calculation of phenology metrics

Algorithms are applied to butterfly counts throughout the season for each species at each site to estimate phenology metrics for each year (and separately for each generation of bivoltine species). The following metrics are calculated for each site, year, species (generation):

- Number of generations
- Date of gap between generations
- Date of first positive count (for each generation)
- Date of last positive count (for each generation)
- Date of highest positive count (for each generation)
- Count at date of highest positive count (for each generation)
- Mean date of flight period (for each generation), as defined as the weighted date of counts (Brakefield, 1987)
- Length of flight period (for each generation), as defined as the standard deviation of counts (Brakefield, 1987)

Long-term and decadal phenology trends are calculated for each species (and generation) at each site, where sufficient data is available, using linear regression models on the timing and duration phenology metrics.

4c. Calculation of abundance indices for each species, site, year

Algorithms are applied to butterfly counts throughout the season for each species at each site to estimate a total abundance for the year (and separately for each generation of bivoltine species). This can be interpreted as the area under the flight period distribution curve. The following metrics are calculated for each site, year, species (generation):

- Number of observations, including zero counts
- Number of positive counts
- Sum of observed counts

The following based on the methods described in Rothery and Roy (2001):

- Index of abundance calculated by Trapazoidal rule fitted to counts
- The smoothing parameter used for the Generalized Additive Model (GAM) fitted to counts
- Sum of fitted counts from GAM
- Sum of Imputed counts (observed or GAM fitted counts)
- Sum of Imputed counts (observed or Trapazoidal estimate)
- Highest seasonal count is a GAM estimate (yes/no)

- Index of abundance estimated via a GAM (GAM Index)
- Proportion of GAM index contributed by estimated counts versus observed counts

Long-term and decadal abundance trends are calculated for each species at each site, where sufficient data is available, using linear regression models on the site indices.

4d. Estimation of zero index for species, site, year

Zero indices are not produced by the GAM models as it only deals with counts data. Where a species is not recorded at a site in a given year there is no count (no data). This may mean that the species was not seen, but could simply be because the site was not walked enough during the flight period of that species. We run a series of automated and manual checks to determine where site indices of zero are considered likely.

4e. Calculation of collated indices (regional index of abundance for each year) and trends

A range of methods are available to analyse UKBMS data to derive regional and national collated indices and be used to estimate trends over time. Two main methods are used to calculate collated indices for the UKBMS:

4e(i). Analysis combining site indices (does not include WCBS)

Although a relative measure, site indices can be combined to derive regional and national collated indices. However, this collation is not a straightforward calculation because not all transect sites in the UKBMS dataset have been recorded each year; some transect sites have operated for twenty years or more but the great majority have not and some have only been recorded for a few years. A statistical model is therefore needed to produce a regional or national index of how butterfly populations have changed each year. In common with most butterfly and bird monitoring schemes in Europe (ter Braak et al. 1994), a log-linear Poisson regression model is used. In this approach, the expected count at a particular site in a given year is assumed to be a product of a site and a year effect. Put more simply, the model attempts to take account of the fact that some years are generally better than others for numbers of a particular butterfly species (the year effect), e.g. if weather is generally favourable. Similarly, the model accounts for some sites supporting higher numbers of a particular species than other locations (the site effect), e.g. if habitat conditions are highly suitable. In this way, for years where a transect site has not been recorded, the model imputes an estimated site index that allows for the general conditions of the year in question and the how favourable the site is. The national collated index is then calculated as the mean (on a log scale) of the imputed and recorded site indices for each year. Long-term and decadal trends are calculated for each species at UK and country level where sufficient data is available, using linear regression models on the collated indices.

4e (ii). Analysis combining individual transect counts (including WCBS)

Since 2013, a suite of methods have been developed analyse individual transect counts, including the Wider Countryside Butterfly Survey (WCBS). Briefly, the methods (Dennis et al. 2012, Dennis et al. 2016) adopt a two-stage approach. Firstly, all butterfly counts in a season from both traditional UKBMS transects and WCBS are used to estimate the seasonal pattern of butterfly counts for that year, either via a Generalised Additive Model or other statistical model

of the flight period pattern. This stage relies heavily on the traditional UKBMS transect data with good coverage throughout the season. A second stage of the model is then applied to the full set of annual counts, accounting for where the counts occur within the flight season, to then calculate annual population indices using a statistical model to accounting for sites and years in a comparable way described above. Long-term and decadal trends are calculated for each species at UK and country level where sufficient data is available, using linear regression models on the collated indices.

4f. Calculation of multi-species (composite) indices and trends

Biodiversity indicators use multi-species (composite) indices of abundance for different groups of butterflies e.g wider countryside and habitat specialist species, and butterflies in different habitats e.g farmland and woodland. Composite indices are calculated following methods developed for UK birds, derived by calculating the geometric mean index across each species assemblage. Trends and confidence intervals in these indicators are then assessed by structural time-series analysis using the program Trendspotter. These indicators are updated and published annually and can be viewed at: http://jncc.defra.gov.uk/page-1824

References

Brakefield, P.M., (1987). Geographical variability in, and temperature effects on, the phenology of Maniola jurtina and Pyronia tithonus (Lepidoptera, Satyrinae) in England and Wales. *Ecological entomology*, *12*(2), pp.139-148.

Dennis, E.B., Freeman, S.N., Brereton, T. & Roy, D.B. (2013) Indexing butterfly abundance whilst accounting for missing counts and variability in seasonal pattern. *Methods in Ecology and Evolution*, 4, 637-645

Dennis, E.B., Morgan, B.J., Freeman, S.N., Brereton, T.M. and Roy, D.B., (2016). A generalized abundance index for seasonal invertebrates. *Biometrics*, 72(4), pp.1305-1314.

Rothery, P. and Roy, D.B., 2001. Application of generalized additive models to butterfly transect count data. *Journal of Applied Statistics*, *28*(7), pp.897-909.

ter Braak, C.J.F., van Strien, A.J., Meijer, R., & Verstrael, T.J. (1994). Analysis of monitoring data with many missing values: which method? In Bird Numbers 1992: Distribution, monitoring and ecological aspects. (eds W. Hagemeijer & T. Verstrael), pp. 663-673. SOVON, Beek-Ubbergen, Netherlands.