# Citizen Science Butterfly Monitoring and Spatial Statistics

Malcolm Connolly, supervised by Rachel McCrea
m.connolly4@lancaster.ac.uk, Lancaster University

## Citizen science

Citizen science (CS) refers to any scheme in which data is collected by volunteer members of the public. The UK Butterfly Monitoring Survey (UKBMS), started in 1976, is a CS scheme where volunteers count the number of butterflies of each species encountered along a fixed route (transect) in a Pollard walk. The transect data is collected weekly between the start of April to the end of September. The observation region in a Pollard walk is an imagined $5m^3$ observation box along the walker's path, see figure 1. CS presents challenges for statistical modelling, including missing data and bias. However, CS gives us some of the most comprehensive and compelling evidence we have to inform conservation efforts.

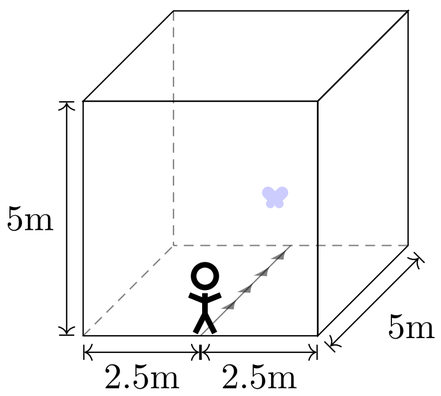


*Figure 1: A $5m^3$ Pollard walk box for observations.*

## Neither capture nor recapture

The population is not closed throughout the survey period, and individuals are not identified in the counts. Models based on capture-recapture in open populations, can be generalised to a stopover model for the UKBMS counts [2]. The stopover model is parametrised by probabilities of entry, detection and retention.

**Stopover model**

$T =$ number of sampling occasions.
$S =$ number of sites.
$y_{ij} =$ count at site $i$ occasion $j$.
$N_i =$ the superpopulation at given site $i$.
$\beta_{i,j-1} =$ proportion of $N_i$ new to survey at time $j$ (**entry**).
$p_{ij} =$ probability of **detection** at site $i$ on occasion $j$.
$\phi_{ja}^{(i)} =$ probability of **retaining** an individual at site $i$ from $j$ to $j+1$ given presence on $a$ previous occasions.

The mean of the count, $\mathsf{E}[Y_{ij}] = \lambda_{ij}$, is given by

$$\lambda_{ij} = N_i \left[ \sum_{b=1}^{j} \beta_{i,b-1} \left( \prod_{k=b}^{j-1} \phi_{k,k-b+1}^{(i)} \right) \right] p_{ij}. \tag{1}$$

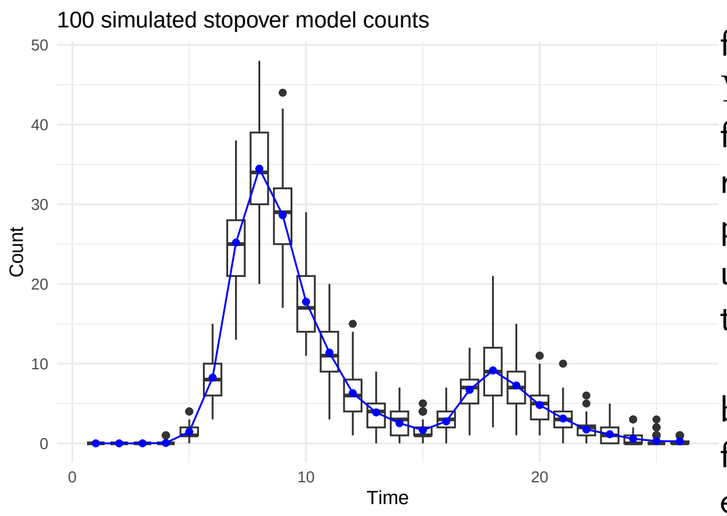


*Figure 2: Entry modelled with normal mixture with two components of unequal weights.*

Counts could be chosen to be from a Poisson distribution so that $Y_{ij} \sim \mathsf{Pois}(\lambda_{ij})$. The expression for the mean in equation (1) is derived by taking the sum over all possible entry times of an individual detected at time $i$, weighted by the probability of their retention.

The entry probabilities $\boldsymbol{\beta}$ can be used to model arrival quanta from distinct broods of butterfly eggs over a season using a mixture of $B$ normal distributions.

In figure 2 we use two normal distributions and take the probabilities to be the integral of the mixture density between successive ordinates, and when $j = 1$ and $T$ use the tails. In place of $T$ parameters for $\boldsymbol{\beta}$, we need only estimate the means and variances of the mixture components. Restricting the entry parameters to lie on parametrised families of curves thereby reduces the number of parameters to estimate.

In general take the mean count to be $\lambda_{ij} = N_i a_{ij}$, for any distribution and arbitrary seasonality $a_{ij}$. For a particular species, and consistent choices for distribution and seasonality, the **generalised abundance index** (GAI) is then the mean of the abundance estimates of the sites.

$$\mathsf{GAI} = \frac{1}{S} \sum_{i=1}^{S} \hat{N}_i \ . \tag{2}$$

## Towards a more spatially representative average

We can re-weight the mean in (2) to account for the distance between sites. In general, given data $\{Y(\mathbf{x}_1), Y(\mathbf{x}_2), \ldots, Y(\mathbf{x}_S)\}$, the problem is to infer $\hat{Y}(\mathbf{x_0})$ as a weighted sum,

$$\hat{Y}(\mathbf{x_0}) = \sum_{i=1}^{S} w_i Y(\mathbf{x}_i), \quad \text{where} \quad \sum_{i=1}^{S} w_i = 1. \tag{3}$$

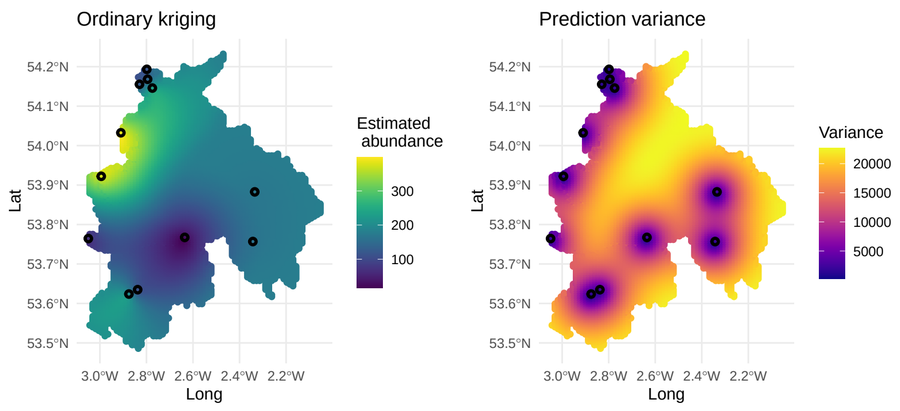


*Figure 3: There is high uncertainty away from surveyed sites.*

**Kriging** enjoys the desirable statistical property of minimising the mean-squared error of prediction, though does make distributional assumptions on $Y$ [1]. **Inverse distance weighting** chooses weights in (3) such that $w_i \propto d_i^{-\beta}$, using the Euclidean distance $d_i = d(\mathbf{x}_0, \mathbf{x}_i)$. That is, When $\beta = 0$, the mean is obtained. In any case, we should exercise caution with extrapolation.

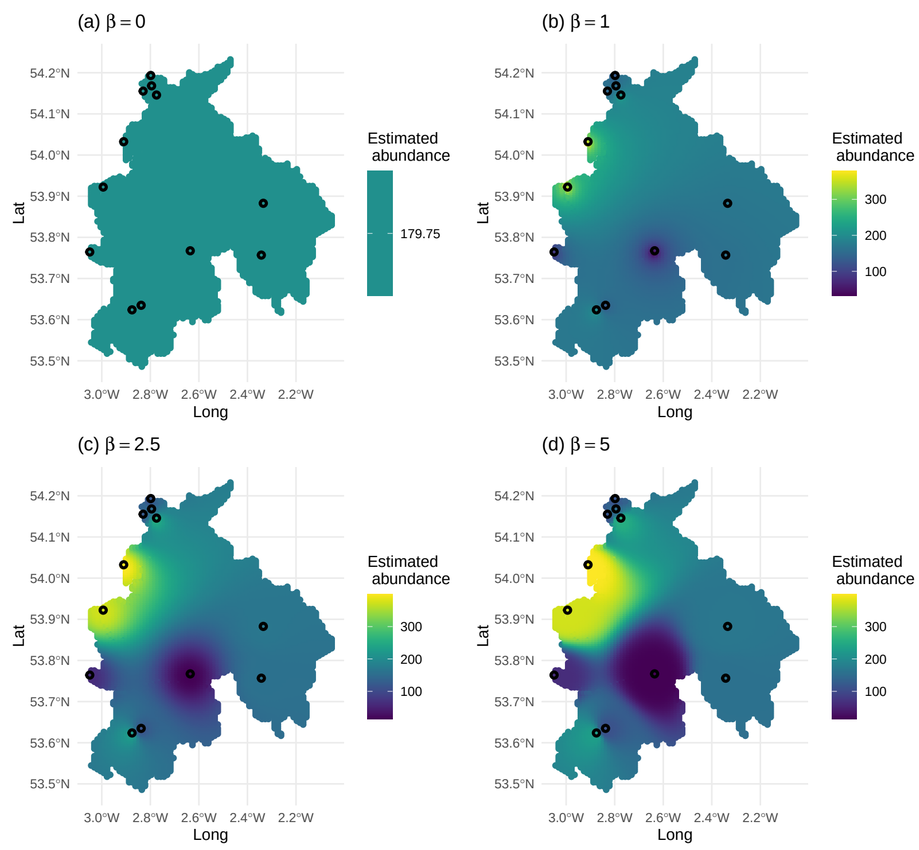


*Figure 4: (IDW) Further sites exert influence on the prediction as $\beta$ increases.*

## References

[1] Noel A. C. Cressie. *Statistics for spatial data*. Wiley Series in Probability and Statistics. Wiley-Interscience Publication, New York, revised edition. edition, 1993. ISBN 1-119-11515-9.

[2] Eleni Matechou, Emily B Dennis, Stephen N Freeman, Tom Brereton, and Jason Matthiopoulos. Monitoring abundance and phenology in (multivoltine) butterfly species: a novel mixture model. *The Journal of Applied Ecology*, 51(3):766–775, 2014. ISSN 0021-8901.

EPSRC Engineering and Physical Sciences Research Council
STOR-i Lancaster University