# Dataset Documentation for Gridded estimates of Ellenberg N, R, F, and L indicators for Great Britain 1990 and 2015-2019

**Susan Jarvis**, UK Centre for Ecology & Hydrology, Lancaster, UK. https://orcid.org/0000-0001-5382-5135
**Hannah Risser**, UK Centre for Ecology & Hydrology, Lancaster, UK, https://orcid.org/0000-0001-9819-1092
**Simon Smart**, UK Centre for Ecology & Hydrology, Lancaster, UK, https://orcid.org/0000-0003-2750-7832
**Don Monteith**, UK Centre for Ecology & Hydrology, Lancaster, UK
**Pete Henrys**, UK Centre for Ecology & Hydrology, Lancaster, UK, https://orcid.org/0000-0003-4758-1482

**Version 1.0.**
**12th August 2022**

## Contents



## Background and rationale

The maps presented here are derived from an integrated model of three different vegetation datasets: Countryside Survey (CS), Natural England's Long Term Monitoring Network (LTMN) and the National Plant Monitoring Scheme (NPMS). Each dataset collects information on plant species present in vegetation plots, or quadrats, from across England. Two of the three datasets also collect data from other areas of GB (CS and NPMS).

The aim of the maps is to provide a consistent estimate of the spatial patterns in four vegetation indicators by combining data from all three datasets into an integrated model. This allows data from schemes with relatively low spatial coverage (e.g. LTMN) and data from schemes with unequal spatial coverage (NPMS) to be included even though it may be challenging to produce spatially explicit mapped predictions using these datasets singly. By integrating with a dataset with representative coverage from a stratified random sampling scheme (CS) we aim to provide a representative map of indicators, whilst allowing all available data to be used.

In this work we focus on a set of commonly used vegetation indicators: Ellenberg N (EBERGN), Ellenberg R (EBERGR), Ellenberg F (EBERGF) and Ellenberg L (EBERGL). These indicators are metrics which describe the relative nutrient content, acidity and wetness and light availability of the environment in which the vegetation is found and are widely used in conservation and restoration contexts.

Integrated maps are produced as a snapshot over the time period 2015-2019. In addition, for comparison we have produced maps for 1990 to investigate potential changes in indicator distributions over time. Maps for 1990 use data from a single dataset (CS) as this was the only survey operating at the time.

**Data collation and processing**

We collated data from all three schemes for the time period 2015-2019 and from CS for the time period 1990. For each scheme we calculated cover-weighted and non-cover weighted Ellenberg scores for each vegetation plot at each time point. For the 2015-2019 time period a total of 2,375 unique plots were recorded, with 445 plots from CS, 1,287 from LTMN and 643 from NPMS. For 1990, 2,282 plots were recorded, all from the CS scheme. For CS we considered only randomly located 2m x 2m “X” plots. For LTMN, we considered only 2m x 2m square (“VC”) plots which are distributed across sites on a stratified grid according to the vegetation types present. For NPMS, we considered only square 5m x 5m plots which are placed on intersections of a fixed grid wherever possible in order to minimise bias. For the NPMS, we only considered data collected during “Inventory” surveys. Further details on each data source can be found in the data availability section below.

We constructed indicators for each vegetation plot using only information on the vascular plants present. We calculated both plot-average and cover-weighted Ellenberg indicator values. To enable standardisation between datasets using different cover recording methods, cover values were converted to the Domin scale and the midpoints of each class used to calculate cover weighted indicators. Although indicator values are available for bryophytes the differences in bryophyte recording between schemes were large and therefore we did not think it was reasonable to include bryophytes in indicator calculation. We used the vegtaxon R package (https://github.com/NERC-CEH/vegtaxon contact hrisser@ceh.ac.uk) to facilitate taxon matching between schemes.

We also collated information on potential covariates to include which may explain some of the spatial patterns in indicators across GB. The covariates considered are broad habitat, nitrogen deposition, sulphur deposition, annual precipitation, maximum July temperature and minimum January temperature. Habitat classifications were assigned at plot-level by

surveyors of each scheme, and matched to the associated Biodiversity Action Plan Broad Habitat classification. Nitrogen and sulphur deposition were calculated as the mean value for the associated 5km square from the preceding 10 years to the year of survey, using the grid square average of deposition data over multiple land cover types taken from the Concentration Based Estimated Deposition (CBED) model (Tomlinson et al. 2020). Climate variables were calculated as the mean value for the associated 1km square from the preceding 10 years to the year of survey, using data obtained from the Met Office HadUK-Grid (Met Office 2021).

**Modelling**

In all schemes multiple vegetation plots are measured in a single "site". The size of this site varies (a 1km square in CS and NPMS, variable in LTMN) as do the number of vegetation plots within each site. For 2015-2019 the number of sites visited was between 34 (LTMN) and 293 (NPMS). In 1990 450 sites were visited in CS.

In 2015-2019 706 (about a third) of plots were revisited within the time frame, all from the LTMN and NPMS surveys. NPMS plots can be visited twice every year, while LTMN and CS work on a rolling programme.

To construct a model of indicator distributions in each time period we needed to account for both the structure of sampling within each scheme (i.e. vegetation plots nested within sites) and the differences in protocols and designs between schemes. Previous work with these data suggested that it was not necessary to account for quadrat size when modelling Ellenberg indicators (Risser et al. 2021) therefore we did not explicitly account for differences in field protocols in our models. We only used the Inventory level of sampling from NPMS and assumed that there would be no bias in indicator values recorded from each scheme due to e.g. differing levels of surveyor expertise.

The three schemes also had different designs that determined where sites and plots were located. CS locations are chosen based on a stratified random scheme designed to ensure representative coverage of GB. We assume that this data is spatially unbiased. LTMN data are collected primarily from nature reserves and other areas of interest and are not designed to have representative coverage of GB. In addition, there are relatively few LTMN sites. We therefore cannot assume this data is unbiased as some landscapes (e.g. arable areas) are unlikely to be captured by this data. NPMS data are collected by community scientists who choose from an available list of squares which are weighted to be more likely to sample rarer habitats (Pescott et al 2019). This list is regionally weighted to try and ensure an even spread of sites across the UK. However, more squares are recorded where there is a higher concentration of community scientists, so this data can also not be assumed to be spatially unbiased. Within sites, CS plots are randomly located, LTMN plots located on a stratified random basis, and NPMS plots located on a stratified random basis where possible.

Our aim was to construct a model which estimates a shared spatial pattern, while accounting for the non-representative coverage of LTMN and NPMS. The weighting of NPMS squares to habitats of interest is by design, so we can simply account for this in the model by including this layer as weights that govern the relative contribution of each observation into the likelihood. However, the location of LTMN sites and NPMS uptake from the list of

available squares is not by design, and therefore we need to think about how best to estimate the patterns of uptake.

The vast majority of LTMN sites are situated on National Nature Reserves (NNRs) for scientific and practical reasons (Nisbet & Holdsworth 2017). The distribution of LTMN sites is therefore largely determined by the population of NNRs. To account for this in our modelling we obtained a layer of NNR locations and allocated all possible 1km squares to either NNR-present or NNR-absent based on overlap with the NNR layer. We then allocated weights of 1 to all squares NNR-absent squares and no. NNR-present squares/total no squares to NNR-present squares to downweight these squares. This weighting was only applied to the LTMN data.

We decided to estimate NPMS uptake directly from the NPMS data, rather than use covariates for uptake (e.g. human population density). This allows for complex patterns in uptake which may be more subtle than simple variations in the pool of potential observers. To do this we include a spatial term which is only fit to the NPMS data, implemented by the inclusion of a dummy variable. Because this spatial term is only fit to the NPMS data, it should be separable from a shared spatial term fit to all datasets.

The basic models are given by an additive function of covariate effects, random effects accounting for plot nested within sites nested within scheme, spatially explicit effects shared between all datasets and an additional spatially explicit effect for the NPMS uptake. This can be written in a simple model as:

Ellenberg ~ covariates + (1/scheme/site/plot) + s(East, North) + s(East, North, NPMS only)

More formally, the model can be described as follows. For observation $l$, in plot $k$, at site $j$, in scheme $i$:

$$y_{ijkl} = \alpha_{ijk} + \sum_{p=1}^{m} \beta_p \cdot f_p(X_{p;ijkl}) + \delta \cdot f_{s1}(x_{ij}, y_{ij}) + \gamma R_i \cdot f_{s2}(x_{ij}, y_{ij}) + \varepsilon_{ijkl}$$

Where $f$ denotes cubic regression spline functions, $x, y$ the spatial coordinates, $\beta$ the regression coefficients relating to the $m$ covariates denoted by $X$, $R_i$ is a dummy variable with value 1 if the observation was from the NPMS and 0 otherwise, and

$$\alpha_{ijk} = \alpha_0 + \mu_{ijk} + \mu_{ij} + \mu_i$$

Where the $\mu$ values specify random effects with uninformative normal priors.

The likelihood is then given by:

$$L(\theta) = \pi(y\,;\,\theta)$$

Where $\theta$ represents the full set of unknown parameters.
Accounting for the weights, $W$, assigned to the observations

$$\dot{L}(\theta) = W \cdot \pi(y\,;\,\theta)$$

$$L\dot{L}(\theta) = \log\left(W \cdot \pi(y\,;\,\theta)\right)$$

Which across all $N$ observations is given by

$$L\dot{L}(\theta) = \sum_{l=1}^{N} \log\big(w_l \cdot \pi(y_l;\, \theta)\big)$$

Note that an effect of survey year within 2015-2019 is not estimated and we assume stationarity over this time period with data from individual years counted as independent replicates.

## Predictions

Predictions were made for both time periods for every 1km cell in GB using the same climate and deposition input data and the dominant target class from the Land Cover Map (LCM). We considered 1km an appropriate scale for prediction as the aim was to capture broad spatial patterns in indicators across GB, rather than provide detailed local level guidance. The 2015 LCM was used for the most recent time period as it is the most recent product which includes the 1km dominant target class. For 1990 we used the 1990 LCM 1km dominant target class. Separate uncertainty maps were also produced for each indicator.

## Quality control

To give an indication of how well the model predicts indicator scores we assess the fit of the model by looking at the root mean square error (RMSE). Evaluation of integrated models is complicated, particularly when some datasets contribute less information (e.g. the LTMN and NPMS data in our model, which are down-weighted relative to CS). We chose to evaluate the models based only on CS data which had the highest weighting, provided the best spatial representation and was the only dataset available in both time periods.

Validation indicated that our generally models performed well overall (RMSE of less than 1, indicating that models predicted indicator scores within +/- 1 unit on average; Table 1). However, models of cover weighted scores performed less well in general, perhaps reflecting the greater variation in cover recording techniques between the three datasets which may have contributed to greater variation in the observed indicator values.

We would therefore recommend using the non-cover weighted values unless cover weighted values are specifically required. We also provide uncertainty (as standard error) maps alongside each of the predicted maps for a spatial assessment of indicator uncertainty. These reflect that certain parts of the country, particularly northern Scotland, have higher uncertainty and therefore estimates for these regions should be treated with more caution.

We also note that although our models were constructed to account for differences between schemes, there could be more subtle differences that mean the final maps are not the same as would be achieved with a single, high coverage unbiased dataset. However, in the absence of such data our method allows us to make best use of the data that is available. We will seek to update the maps going forward as new data becomes available.

**Table 1.** RMSE values of each indicator model

| Indicator | RMSE of 1990 model | RMSE of 2015-2019 model (CS-only) |
|---|---|---|
| **Ellenberg N** | 0.76 | 1.18 |
| **Ellenberg N - cover weighted** | 1.44 | 1.54 |
| **Ellenberg R** | 0.69 | 0.94 |
| **Ellenberg R - cover weighted** | 1.54 | 1.52 |
| **Ellenberg F** | 0.55 | 0.70 |
| **Ellenberg F - cover weighted** | 1.28 | 1.42 |
| **Ellenberg L** | 0.44 | 0.59 |
| **Ellenberg L - cover weighted** | 1.68 | 1.74 |

## Calculation of change

To calculate change between 1990 and 2015-2019 we subtracted the estimated indicator values for 1990 from the estimated values for 2015-2019. As a rough indication of uncertainty around this change we calculated the standard error of the difference as

$$SE_{2015\text{-}2019\text{-}1990} = \sqrt{(s.e._{1990}^2 + s.e._{2015-2019}^2)}$$

Whilst this cannot be regarded as a true variance estimate due to the fact that covariance is not accounted for, it provides a relative measure of uncertainty in the change.

## Data structure

Three sets of 2 band tiff files are provided with the following naming structure:

[variable]_90.tif: predicted mean values of each indicator in 1990 in band 1, standard error around predictions in band 2

[variable]_15-19.tif: predicted mean values of each indicator in time period 2015-2019 in band 1, standard error around predictions in band 2

[variable]_difference.tif: predicted differences in indicator values between 1990 and 2015-2019 in band 1, standard error around predicted differences in band 2

Where [variable] contains either EBERGN, EBERGR, EBERGF or EBERGL with either "_site" indicating non-cover weighted indicator maps or "_cwt" indicating cover weighted indicator maps. A total of 24 files are provided.


## Acknowledgements

Many thanks to Oli Pescott for providing NPMS weights and advice on using NPMS data and to Ed Rowe for interpretation of the results. This work was supported by the Natural Environment Research Council award number NE/R016429/1 as part of the UK-SCAPE programme delivering National Capability.


## Data availability

This data is made available under the Open Government Licence. When using the data you must cite: Jarvis, S.G.; Risser, H.; Smart, S.M.; Monteith, D.; Henrys, P.A. (2022). Gridded estimates of Ellenberg N, R, F, and L indicators for Great Britain, 1990 and 2015-2019. NERC EDS Environmental Information Data Centre. https://doi.org/10.5285/0a9900f2-8556-4487-bc13-9c2fdc05082c

Scripts to replicate the modelling process are available (access on request; contact susjar@ceh.ac.uk) at https://github.com/NERC-CEH/T1.1-indicatormaps. Input data are publically available (locations listed below) although locations of CS plots are only available on request.

**Countryside Survey vegetation plots:** Smart, S.M.; Andrews, C.; Fitos, E.; Garbutt, R.A.; Gray, A.; Henrys, P.A.; Koblizek, E.; Pallett, D.W.; Robinson, D.A.; Rose, R.J.; Rowe, R.L.; Scarlett, P.; Towill, J.; Wagner, M.; Williams, B.; Wood, C.M. (2020). Vegetation plot data from the UKCEH Countryside Survey, Great Britain, 2019. NERC Environmental Information Data Centre. https://doi.org/10.5285/fd6ae272-aeb5-4573-8e8a-7ccfae64f506

**LTMN vegetation plots:**
http://publications.naturalengland.org.uk/category/5316639066161152

**NPMS vegetation plots:** Pescott, O.L.; Walker, K.J.; Day, J.; Harris, F.; Roy, D.B. (2020). National Plant Monitoring Scheme survey data (2015-2019). NERC Environmental Information Data Centre. https://doi.org/10.5285/cdb8707c-eed7-4da7-8fa3-299c65124ef2

**Land Cover Map 1990:** Rowland, C.S.; Marston, C.G.; Morton, R.D.; O'Neil, A.W. (2020). Land Cover Map 1990 (1km dominant target class, GB) v2. NERC Environmental Information Data Centre. https://doi.org/10.5285/f5e3bd00-efd0-4dc6-a454-aa597d84764a

**Land Cover Map 2015:** Rowland, C.S.; Morton, R.D.; Carrasco, L.; McShane, G.; O'Neil, A.W.; Wood, C.M. (2017). Land Cover Map 2015 (1km dominant target class, GB). NERC Environmental Information Data Centre. https://doi.org/10.5285/c4035f3d-d93e-4d63-a8f3-b00096f597f5

**Climate data:** Met Office; Hollis, D.; McCarthy, M.; Kendon, M.; Legg, T.; Simpson, I. (2021): HadUK-Grid Gridded Climate Observations on a 1km grid over the UK, v1.0.3.0 (1862-2020). NERC EDS Centre for Environmental Data Analysis, 08 September 2021. doi:10.5285/786b3ce6be54468496a3e11ce2f2669c.

**Deposition data:** Tomlinson, S.J.; Carnell, E.J.; Dore, A.J.; Dragosits, U. (2020). Nitrogen deposition in the UK at 1km resolution, 1990-2017. NERC Environmental Information Data Centre. https://doi.org/10.5285/9b203324-6b37-4e91-b028-e073b197fb9f

**NPMS weights:** access on request, described in Pescott OL, Walker KJ, Harris F, New H, Cheffings CM, Newton N, et al. (2019) The design, launch and assessment of a new volunteer-based plant monitoring scheme for the United Kingdom. PLoS ONE 14(4): e0215891. https://doi.org/10.1371/journal.pone.0215891

**NNR map:** https://data.gov.uk/dataset/726484b0-d14e-44a3-9621-29e79fc47bfc/national-nature-reserves-england

## Additional references

Nisbet, A., Smith, S.J., & Holdsworth, J., (Eds) 2017 Taking the long view: An introduction to Natural England's Long Term Monitoring Network 2009 – 2016. Natural England Report NERR070.

Risser, Hannah; Jarvis, Susan; Henrys, Peter; Maskell, Lindsay; Tomlinson, Sam; West, Bede; Smart, Simon; Monteith, Don. 2021 Harmonisation and integrated modelling of UK long-term vegetation data: a case study focussed on heath & bog habitats. Lancaster, UK Centre for Ecology & Hydrology.