UK Centre for Ecology & Hydrology

Harmonisation and integrated modelling of UK long-term vegetation data: a case study focussed on heath & bog habitats

Draft report to Defra under the UKCEH-Defra Memorandum of Agreement project 07111 Task 04

Authors: Hannah Risser, Susan Jarvis, Peter Henrys, Lindsay Maskell, Sam Tomlinson, Bede West, Simon Smart & Don Monteith

Date 11/02/2021

Title	Harmonisation and integrated modelling of UK long-term vegetation data: a case study focussed on heath & bog habitats
Client	Defra
Client reference	Add Client reference
Confidentiality, copyright and reproduction	
UKCEH reference	UKCEH Project 07111 Task 04
UKCEH contact details	Don Monteith UK Centre for Ecology & Hydrology Lancaster Environment Centre Library Avenue Bailrigg Lancaster LA1 4AP email: dmonteit@ceh.ac.uk
Corresponding author	Monteith, Don
Approved by	Simon Smart
Signed	
Date	02/03/2021

Contents

Executive Summary

The primary aim of this project was to determine the potential to combine the vegetation data of some of the most established national monitoring and survey schemes within a single analysis in order to maximise our understanding of long-term vegetation change across the UK. Differences in methodologies between schemes has been an impediment in attempting to integrate data across them in the past.

The schemes of interest were the UK Environmental Change Network (ECN), the Countryside Survey (CS), Natural England's Long Term Monitoring Network (LTMN) and the National Plant Monitoring Scheme (NPMS).

The project focussed on heath and bog vegetation as a proof of concept.

Computer scripts were written at the outset in order to efficiently extract all vegetation data (i.e. not only from heath and bog) from the various scheme databases and other repositories. These scripts can be re-used to provide updated downloads whenever a scheme refreshes its data holdings.

The UKCEH project team worked with Natural England staff in standardising raw data formats for LTMN data to improve the efficiency of extraction of the LTMN data and assist NE more generally in the management and curation of these datasets.

Species codes used by each scheme were harmonised through the production of a common species dictionary. Code was produced as a package "*vegtaxon*", in the statistical programming language R. This matches Latin names of UK vascular plant species to the current accepted name. The R package has value beyond the current project, including potential application to Defra's developing UK APIENS project that reports on air quality impacts on ecosystems.

All plots from the fully integrated vegetation dataset that, at any point in their records, showed the necessary characteristics of either heaths or bogs, were then identified. This approach has already been re-applied under a separate 25YEP indicator project within the current UKCEH-Defra MoA, focussed on unimproved grassland.

A range of the most appropriate vegetation indicators were subsequently selected to characterise spatial and temporal variation in the heath and bog assemblages. These were Ellenberg R (soil acidity), Ellenberg N (soil fertility), Ellenberg W (soil wetness), and sets of both positive and negative Common Standards Monitoring (CSM) indicators.

Trends over time in the Ellenberg and CSM metrics were then modelled separately for the four schemes. Tight agreement was observed between most schemes in rates of change in the selected Ellenberg metrics, and particularly with respect to Ellenberg N and Ellenberg R, although there were significant differences in the average levels between schemes. In contrast, there was substantial disagreement between schemes in the temporal patterns of the CSM metrics.

An entirely novel vegetation sward simulation approach was developed. This allowed a virtual assessment of the importance of various differences in sampling protocols between schemes in influencing the value of vegetation metrics and their sensitivity to long-term shifts in vegetation assemblages. The simulations suggested that plot size was not important in determining mean Ellenberg scores, whether the vegetation data were cover weighted or not. The simulations provided further evidence that the calculation of CSM scores is much more sensitive to the specific scheme sampling methodologies.

Simulation of variation in the plot assemblages across an environmental gradient, as a surrogate for temporal change, demonstrated that that estimates of change in Ellenberg metrics were consistent

across schemes. While the simulation work has proved very successful, producing authentic results, more work is recommended to further improve realism, e.g. by including requirements for the spatial clustering of some species. It is envisaged that there will be widespread application of the method in developing better models for assessing vegetation survey data, and potential future survey design.

Informed by the assessment of similarities and differences in the temporal models for the individual schemes, and by the simulation work, appropriate structures were then developed for models of each vegetation indicator, integrated across schemes. The models therefore included random slopes and random intercepts to allow for differences between schemes.

The integrated models for the Ellenberg metrics provided consistent and precise estimates of change over time for all three of them. In all three cases, Ellenberg scores for heath and bogs were found to be increasing significantly over time.

We found little evidence for the value of combining data from multiple schemes for the assessment of change in the CSM metrics. The direction of trends in the integrated CSM models was highly uncertain as a result of conflicting trends in the individual datasets. Further work is required to better understand the reasons for the between-scheme disparities.

In the final round of modelling, spatially explicit environmental covariates representing air pollutant loads (sulphur and nitrogen) and climate, were introduced in an attempt to explain variation and change in the integrated Ellenberg signals. Our approach was unusual, in that it first standardised both response and explanatory variables in order to remove the potential influence of spatial effects, which then allowed a more robust test of hypotheses concerning the drivers of temporal change.

The attribution analysis provided new evidence demonstrating the strength of the link between nitrogen deposition and long-term change in Ellenberg N (soil fertility) in UK heath and bog vegetation. Ellenberg N was found to increase in CS, ECN and LTMN datasets at a remarkably similar rate. This is particularly striking, given that LTMN commenced only in 2010, and suggests that this trend towards more-nutrient loving species has been occurring over many years and appears to have continued until quite recently at least. It was not possible to determine whether the continuing rise in Ellenberg N represents a lagged floristic response to the historical nitrogen load, or whether there is sufficient contemporary deposition for these communities to be continuing to respond dynamically to continued soil N accumulation. This is an important distinction from an air quality policy perspective, but requires further investigation.

The attribution analysis also provided support for the hypothesis that changes in sulphur deposition have been a key driver of change in Ellenberg R (soil acidity), although it has not so far been possible to discount the importance of other drivers. We found some evidence that the response in Ellenberg R to the reduction in S deposition had been stronger in habitats with assemblages indicative of less acid conditions. This needs to be explored in further detail, but it is possible that this may at least in part be linked to the role of buffering of pollutant acidity by peaty soils that are rich in (natural) organic acids.

We found tentative evidence that inclusion of precipitation as an additional covariate strengthened the apparent role of sulphur deposition. However further work is required to investigate these relationships in more detail. In particular, there is a need to provide a more finessed range of potential climate covariates, including seasonal temperature and precipitation variables, and to consider more complex non-linear approaches.

Our attempts to model change in Ellenberg W (soil wetness) were more equivocal. Again a more detailed breakdown of climate variables, and possibly the introduction of additional metrics such reflecting, for example, water balances, will be necessary to make further advances here.

In summary, the attribution work adds considerable strength to a developing evidence base that suggests that terrestrial vegetation in the UK is changing progressively, albeit very gradually, in

response to long-term shifts in in the deposition and accumulation of air pollutants, but more work is required to pin down these relationships with greater confidence, particularly with respect to Ellenberg R. More work is also required to better understand the drivers of the national increase in Ellenberg W in heath and bogs.

Overall the project has demonstrated that there is considerable potential for integrating data for these surveys, and possibly others, in order to shed new light on the nature and causes of vegetation change. The fact that there is a clear regional shift in the Ellenberg metrics over recent decades, that can be linked directly to regional changes in drivers (particularly with respect to air pollutants), needs to be taken into account when assessing the potential impact of other more local drivers, including Environmental Land Management (ELM), on these heath and bog habitats at least, and possibly others.

The work also highlights the individual value of all these monitoring schemes, each established for a different reason, and the added value of bringing their various spatial and temporal strengths together in a single analysis. The differences in characteristics such as extent, frequency and co-location should be seen as a strength of the UK's diverse set of long-term environmental monitoring and observation assets.

While ECN plots are much fewer in number and much less geographically spread than the other surveys, our analysis showed that the trends identified in the data are still broadly consistent with the changes observed more widely across the UK. Since the ECN plots are monitored more frequently, and surveys are co-located with range of other environmental measurements, including air and soil chemistry and meteorology, they have particular potential for testing cause-effect hypotheses.

The project represents a major step forward in our ability to exploit the interoperability of what until now have been considered rather disparate sources of data. There is substantial potential to explore the signals we have begun to quantify, decipher causes of change within heath and bog habitats in much greater detail, and extend the approach to other habitats.

1 Introduction

The composition of vegetation provides a strong indication of the state of the immediate environment, including the structure, functioning and diversity of the ecosystem, and is widely applied in the assessment of habitat condition. Long-term monitoring of composition, therefore, provides valuable insights into ecosystem resilience to environmental pressures, and the causes and consequences of environmental change. Vegetation monitoring also conveys information on changes in the properties of natural capital, including biodiversity, and the delivery of ecosystem services, and is therefore fundamental in assessing the efficacy of policy measures developed to protect and/or restore ecosystems.

Defra have identified three current areas where assessment of vegetation monitoring data at a national scale should benefit policy development. These are: 1) Development of indicators for tracking progress in delivering the 25 Year Environment Plan (25YEP); 2) Support for condition assessment of SSSIs and other designated sites, and 3) Meeting the commitments of the National Adaptation Programme (NAP) for climate change.

There is mounting evidence that vegetation communities have been changing gradually over wide areas of UK semi-natural habitat in recent years. The spatial scale of change suggests that the key drivers operate over a similar range, implicating the effects of either, or both, changes in air pollutants and climate. There remains a need, however, to better quantify the nature and extent of these regional scale changes in vegetation, and to confidently attribute causes, in order that other more localised impacts, such as targeted environmental stewardship, can be also be evaluated appropriately.

A recent analysis of the long-term datasets of the UK Environmental Change Network, highlighted widespread increases in vegetation species richness across a range of semi-natural habitats (Rose et al. 2016). The changes were most closely linked with concomitant increases in Ellenberg R (a vegetation indicator of soil acidity) and consistent, therefore, with evidence for a national-scale response to declining soil acidity, a consequence of large reductions in acid deposition over recent decades. There was also an indication of increases in Ellenberg W (and indicator of soil wetness) in drier lowland habitats.

A number of studies of vegetation across the UK and wider Europe have also highlighted long-term increases in Ellenberg N (an indicator of soil fertility and associated, in non-agricultural environments, with levels of atmospheric nitrogen deposition). Although Ellenberg R and Ellenberg N scores for individual taxa tend to be broadly correlated, the two indicators have been developed to reflect effects of different drivers, so assessment of changes in both is useful in determining the likely causes of vegetation change. In some circumstances common shifts in both could reflect related processes, such that a reduction in soil acidity (increase in Ellenberg R) could enable the return of more acid sensitive species that are better able to respond to a legacy of accumulated of soil nitrogen (thus an increase in Ellenberg N). Clearly, therefore, it is vital that both indicators are considered together, and that hypotheses are developed that best enable the discrimination of these two potential drivers of change.

The exploration of competing hypotheses surrounding the causes of vegetation change, and the separation of impacts of local management from effects of regional-scale pressures will be most effectively achieved through the integration of different types of data (biological, physical, chemical and management), both measured and modelled. In addition, efforts should made to draw in and integrate vegetation monitoring data from as many reputable sources as possible. Fine-scale intensive monitoring is particularly suited for investigation of short-term variability, and is often best placed to establish direct links between drivers and responses, but if possible should be combined with data collected from more occasional surveys capable of capturing much wider-scale patterns of change.

There are significant challenges, however, in bringing together vegetation datasets developed under different monitoring and survey programmes. There is no single established UK vegetation monitoring protocol, and most monitoring and survey schemes have unique methodological attributes that best suit their individual aims. Protocols differ, for example, with respect to the size and replication of survey plots, and approaches to assessing species cover. To complicate matters further, there are often differences between schemes in the names used for some species and the degree to which varieties, and species may be lumped into higher level taxonomic categories.

The efficient and robust integration of vegetation data from across schemes to enable a single integrated analysis, the ultimate aim of this project, therefore involves a number of steps. First, computer scripts are required to enable efficient extraction of data from host databases, and any potential programme-specific quirks in the consistency of formatting overcome. Second, taxonomic differences between programmes need to be addressed through the creation of a single species dictionary for the project. Third, decisions are required on the range of metrics to be applied in the analysis. Finally, the potential influence of programme-specific differences in sampling protocols on the calculation of the selected vegetation metrics and sensitivity of those metrics to detect vegetation change needs to be assessed, so that protocol-specific effects can be included, if necessary, in subsequent analysis.

This project embodies all of the steps outlined above, and the subsequent assessment and attribution of change in vegetation metrics in the integrated dataset. Four nationally applied vegetation monitoring/survey schemes were identified for integration, namely the UK Environmental Change Network (ECN), the Countryside Survey (CS), Natural England's Long Term Monitoring Network (LTMN) and the National Plant Monitoring Scheme (NPMS).

The work presented in this report covers three sub-work packages outlined in the original Memorandum of Agreement, namely:

Task 3.2: Data Integration

Develop an approach for extraction and harmonisation of data from a range of national plant monitoring programmes, and subsequent generation of indicator metrics.

Task 3.3: Understanding 25 YEP Indicator sensitivity to pressures and spatial scaling.

3.3a) Test compatibility of a range of vegetation indicators derived from each programme.

3.3b) Investigate metric performance, e.g. trend comparisons, signal to noise ratios; seasonal and inter-annual variation relative to long-term baselines.

Task 3.4: Understanding pressures on 25 YEP Indicators

3.4a) Analyse relative influences of air pollution and climate on terrestrial vegetation indicators

3.4b) Investigate the potential of resulting statistical models to predict vegetation response to these regional-scale drivers across a range of semi-natural habitats

It was agreed with Natural England at the project start-up meeting that, given the resources and timescale for the project, it would be necessary to restrict the scope of habitats assessed, but it was also recognised that boundaries between broad habitat groupings are often blurred and that vegetation monitoring plots sometimes cross these boundaries over time. Consequently both heaths and bogs were selected for specific attention, with a view that this project should be considered a "proof of concept" study, with a view to extending the work to other habitats in future.

The following chapters detail the series of steps set out above. Chapter 2 (Data Integration) describes the vegetation and driver data sources, development of the data extraction code, issues and recommendations to data providers regarding data recording and reporting protocols to facilitate efficient external data retrieval, taxonomic harmonisation and development of a project species

dictionary, and our approach to heath and bog plot selection. Chapter 3 (Indicator Selection) considers the potential range of indicators available to the project and the final rationale for selection. Chapter 4 (Comparison of trends in vegetation metrics between individual schemes) describes the development of programme-specific models to describe change in vegetation metrics and presents and compares the resulting trends. Chapter 5 (Vegetation Simulation) outlines the development of a virtual (i.e. computer based) vegetation sward simulator, and its subsequent application in exploring the influence of the different monitoring protocols on the calculation of metric scores and their relative sensitivity to change over time, in order to inform the need for terms accounting for protocol differences to be included within a final integrated model. Chapter 6 (The integrated vegetation model: temporal change) presents the outcomes of the integrated modelling with respect to testing for and quantifying overall trends in metrics over time, while Chapter 7 (The integrated vegetation model: trend attribution) is focussed on testing a range of hypotheses around the drivers of vegetation change via the fitting of environmental covariates. Finally, in Chapter 8, we summarise outcomes of the project, consider the primary research findings, assess the scientific and policy value of the work and make recommendations for further developments of the approach.

2 Data Integration

2.1 Introduction to the data integration exercise

We collated and integrated vegetation plot data from the Countryside Survey (CS), the Environmental Change Network (ECN), the Long Term Monitoring Network (LTMN), and the National Plant Monitoring Scheme (NPMS). Information on the history of each scheme, and a brief description of the types of data they record, are provided below.

The UKCEH CS has recorded plant and soil data in Great Britain since 1978. Plants are surveyed at over 591 1 km squares, chosen in an unbiased manner to represent all major UK habitat types across Great Britain (Carey et al. 2008). Historically, these squares have been surveyed on average every 10 years. Since 2019, surveys have taken place within a 5-year rolling programme. At each square, surveyors record, on average, 29 vegetation plots of different types, in addition to collecting data on soils, freshwater habitats and invertebrates, and land cover and landscape features. CS surveys are undertaken by professional surveyors.

The ECN has amassed data on plants, invertebrates, vertebrates, soils, air, water, weather, and climate at a limited number of terrestrial sites located across the UK since 1992 (Sykes and Lane, 1996). The 11 sites still in operation were established between 1992 and 1998. They cover a range of habitats and were selected on the basis of their size, the availability of past research at the site, known history of management and relative stability of land management. They provide broad geographical coverage, a range of environmental conditions and habitats, some guarantee of long-term physical/financial security, and are mostly considered to contain high-quality habitats. The ECN vegetation monitoring protocols comprise both high frequency (1 to 3 yearly) “fine grain” monitoring and lower frequency (9 year), “coarse grain” monitoring. ECN surveys are undertaken by ECN site managers with botanical training and other professional surveyors with detailed knowledge of the individual sites.

Natural England’s Long Term Monitoring Network has been recording weather, air quality, vegetation, soil, bird, butterfly and site management data at 37 sites across England since 2010 (Nisbet et al., 2017). The LTMN was designed to be complementary to the ECN, using some of the same protocols, including the coarse grain vegetation protocol. Most, but not all, LTMN sites are National Nature Reserves (NNRs) and were chosen to provide a wide geographical spread for a wide range of broad habitats. In common with ECN, therefore, LTMN sites tend to be under relatively stable management and of high environmental quality. LTMN surveys are undertaken mostly by Natural England staff drawn from a variety of roles and overseen by professional surveyors.

The National Plant Monitoring Scheme, launched in 2015, uses citizen science surveyors to record vegetation data at 1km squares across the UK (Pescott et al., 2019). The NPMS is unique in that it involves participation of recorders with a wide range of botanical skill levels. Recorders conduct surveys at one of three levels; wildflower, indicator or inventory. The wildflower and indicator levels survey vascular plant species from a specified list of indicator species for each habitat, whereas the inventory level involves the recording of all species. Recorders are encouraged to visit their plots twice yearly.

Although it had been decided at the outset to concentrate on plots representing heath and bogs only (see Chapter 1), it was first necessary to collate all available vegetation data so that the heath and bog subset could subsequently be defined robustly.

2.2 Data collation and cleaning

A significant amount of project time was spent collating and preparing the datasets for analysis, particularly the Natural England Long Term Monitoring Network data that had not previously been collated for the purposes of a multi-site analysis. Data from the CS and ECN were accessed directly from the UKCEH Oracle databases in which they are held. National Plant Monitoring Scheme data was received directly from Oliver Pescott in advance of it being published on the EIDC (now available at https://catalogue.ceh.ac.uk/documents/cdb8707c-eed7-4da7-8fa3-299c65124ef2). Data from Natural England's Long Term Monitoring Network is freely accessible online (http://publications.naturalengland.org.uk/category/5316639066161152). Data from all four schemes are publically accessible (under relevant licenses), apart from the CS location data which is confidential.

2.2.1 LTMN data cleaning

LTMN data is stored in the data.gov repository as separate MS Excel spreadsheets for each site and year of survey. Historically, the formatting of these spreadsheets had not been entirely consistent between sites and over time. At the beginning of this project, Natural England were in the process of migrating all data to new consistent templates. Consequently, some survey data had yet to be converted to the new format at the beginning of this project. This is important, as it meant that it was not possible at the outset to develop a single piece of computer code to extract all the data from this archive.

Files in the new format were shared with UKCEH via *Dropbox* as soon as they were available, and all data received through this channel were included in the final analysis. These data are also now available on the Natural England Access to Evidence website.

We encountered various residual errors and inconsistencies within the newly formatted LTMN spreadsheets. These had to be identified and fixed before the data could be used in our analysis. Many of these problems were associated with the new template providing too much flexibility in the manner data could be entered by surveyors. For example, there were few limits on what types of data could be entered, and it was also possible for surveyors to add, remove or re-order columns within sheets. The main problems encountered are outlined below:

- Formatting errors: flexibility of data recording template meant that there were few limits on what could be entered or added to the spreadsheet.
  - Column names: These were not fixed and in some cases were altered.
  - Number of columns: It was possible to add/remove columns from the template. For example, at one site an extra column had been added with no explanation, and at another site several columns were accidentally duplicated.
  - Text spacing: Free entry of text led to formatting inconsistencies. For example, grid references could be entered as 'AA 11111 1111', 'AA1111111111', or 'AA11111 11111', or have an odd number of digits and therefore not be valid.
  - Dates: Entered in multiple formats, including: DD-MMM-YY, DD-MMM, DD/MM/YYYY.
  - Free text entry of presence/absence data: The presence/absence of a species is noted for each cell within each plot, with 1 indicating presence and 0 absence. There were multiple instances of accidental entry of incorrect numbers, e.g. 2 in place of 1.
  - Free text entry of site codes: Each Excel spreadsheet contained information for one site only. However, the "sitecode" associated with the plot-level data was not returned automatically and instead had to be entered manually. This led to accidental incorrect noting of site codes in more than one instance.

- Data in free text boxes. Plot-level information was sometimes stored in a free text box on the first sheet of each file if an appropriate place did not exist elsewhere in the document. For example, information about which plots were surveyed on that visit and whether these plots were relocated accurately could be found here. Such information was not included in our final dataset as it would have been incredibly time-consuming to extract. This valuable quality control information would benefit from being recorded in a more repeatable manner.
- Inconsistent plot numbering system. Plot numbering systems at several sites changed between years. For example, site B29 plot 1 in 2010 became B29 plot T1 in 2013, despite being at the same location. Data with this type of issue was fixed promptly by NE once discovered.
- Species occurrences without matching plot-level data. Some species presence data had no associated plot-level data. Plot-level data included survey dates, plot location, and other key details. These species occurrences were removed from our analysis.
- Lack of metadata. Clear metadata explaining the survey sheets and their contents was not made available to UKCEH.

The resulting need for careful manual quality checking and bespoke coding consumed considerably more staff time than had been initially factored in for this stage of the work. However, UKCEH and NE staff soon established an effective working relationship to enable a constructive and iterative exchange of information and proposed solutions, and data for the vast majority of LTMN surveys eventually made it through to the initial analysis stage.

At the January 2020 project meeting these data quality issues were discussed with NE, when it was proposed that NE create a new data entry template with fixed formatting requirements to limit flexibility and potential for error. UKCEH also suggested that all data be made available on a single webpage and ideally in one aggregated table in order to facilitate the most efficient extraction of data in the future.

2.2.2 ECN data cleaning

This data collation step highlighted gaps in the location data for some vegetation plots in the Environmental Change Network database. On discovering this, site managers were contacted and asked to provide exact location data for these plots.

2.3 Data filtering

Schemes differ in the types of data collected, regions of the UK covered, and the methods used for collection. A key part of preparing the integrated dataset involved making decisions on what data to include from each scheme. In order to analyse these datasets using an integrated approach, while minimising the complexity of the modelling structure, the structure of the data from the individual schemes must be broadly similar. The key differences between datasets are outlined below and a summary of the data included in the final modelling dataset from each scheme is provided in Table 2.1.

- **Location:** Only two of the schemes, the ECN and NPMS, collect data in Northern Ireland, and the ECN site does not include heath and bog plots. All data from the province were therefore excluded as coverage was deemed insufficient to be regionally representative.
- **Plot types:** All schemes survey more than one plot type, e.g. square, linear, woodland, and roadside plots. In order to minimise the effect of plot type in our analysis we chose to focus on square plots only, which are mostly 2m x 2m in size.

- **Bryophyte recording:** There were differences in the level of bryophyte recording between schemes, with the ECN and LTMN recording bryophytes to species level, the CS recording to broad groupings, and the NPMS not recording bryophytes at all. As a consequence we chose to limit the scope of our analysis to vascular plants only. This decision may ultimately have had a significant influence on the modelling outcomes since bryophytes are an important structural and ecological component of heath/bog habitats.
- **Survey type**: The NPMS allows recorders to choose from one of three survey levels according to past botanical experience. The Wildflower and Indicator levels involve the use of a limited habitat-specific species list, while the Inventory level involves recording all species present in the plot, not limited to a list. We used data from the Inventory level only on the grounds that this was most comparable to the survey methods in the other three schemes.
- **Cover:** Each scheme uses a different method for estimating cover. The LTMN protocol requires the estimation of percentage cover of each species in a plot on a continuous scale, the CS record cover at 5% increments, the NPMS use the Domin scale, and the ECN and LTMN both record presence/absence in each cell (40 x 40 cm) of a plot.
- **Temporal resolution:** Schemes differ greatly in the frequency of survey visits and the number of years for which plots have been surveyed. The CS and ECN have both monitored their plots for 25+ years, whereas the LTMN and NPMS were established much more recently, with records currently around 10 and 5 years in length respectively. Survey frequency of the plots taken forward to the modelling dataset ranged from twice yearly for the NPMS to once per decade for the CS.

Table 2.1. Summary of plot data included in the individual scheme and integrated analysis.

Scheme	Years	Countries	Plot types	Plot size	Survey frequency (years)
CS	1978, 1990, 2000, 2007, 2016-2019	England Scotland Wales	X plots (2x2m nest only) U plots	$4m^2$ (2m x 2m plot) $4m^2$ (2m x 2m plot)	10 (on average) 10 (on average)
ECN	1993-2015	England Scotland Wales	Coarse-grain (VC) Fine-grain (VF and VFA)	$4m^2$ (2m x 2m plot containing 25 cells, each 40cm x 40cm) $1.6m^2$ (ten 40cm x 40cm cells within larger 10m x 10m, or $100m^2$, plot)	9 3 (VF) 1 (VFA)
LTMN	2010-2019	England	Coarse-grain (VC)	$4m^2$ (2m x2 m plot containing 25 cells, each 40cm x 40cm)	4 (on average)
NPMS	2015-2019	England Scotland Wales	Square 5x5m plots at inventory level	$25m^2$ (5m x 5m plot)	1-5

2.4 Species name harmonisation

The four monitoring schemes were established at different times, and this partly accounts for the adoption of different names for the same species within scheme species lists. This can pose challenges not only for integration of data from multiple schemes but also for ensuring that taxa from across all schemes are linked consistently with appropriate plant indicator values. A common species dictionary was therefore created to solve this nomenclature problem. This enables vascular plant species names from each scheme to be matched to the name currently accepted by the Botanical Society of Britain and Ireland, and relies heavily on the Taxon name parser tool available on the BSBI website (https://database.bsbi.org/taxonnameparser.php). The method for creating the dictionary is outlined below.

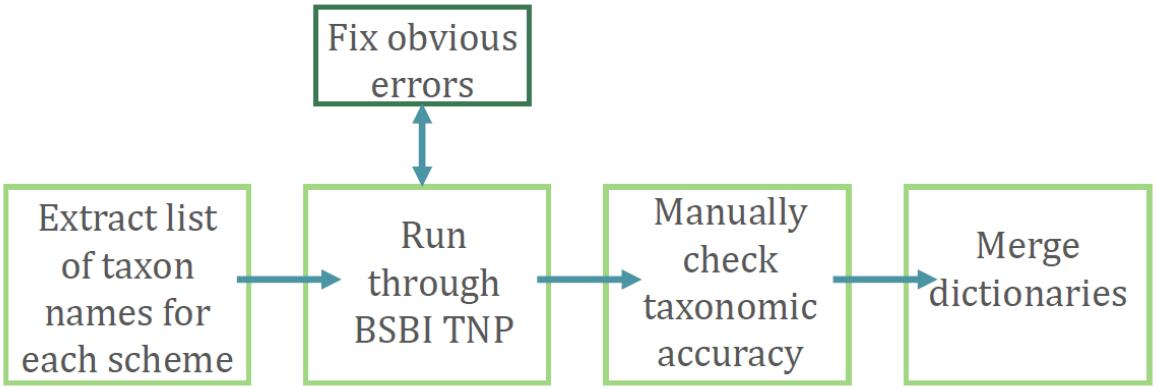


Figure 2.1. Flow diagram depicting species dictionary workflow. The process is described in the following text box.

1. Extract full list of observed plant species and their associated codes (e.g. BRC numbers, VESPAN codes).
2. Clean the species list to remove obvious spelling errors and non species-specific descriptors – e.g. "Acer campestre (c)" will need the (c) to be removed.
3. Input this cleaned species list to the Taxon name parser tool on the BSBI Distribution Database website (https://database.bsbi.org/taxonnameparser.php). The edited species names column created earlier must be entered as the column containing taxon names in the tool. This tool will match this list of species names to the current accepted name in the BSBI database. These accepted names are mostly matched to Stace 3rd edition (Stace, 2010).
4. Take output from the BSBI tool and check for errors where the species name entered does not have a match on the BSBI database. Where there are lots of errors, as found in the scheme datasets, go back to step 2 and clean the data further. If the reason for a name not matching to the database was not immediately obvious, we used the BSBI database and NBN Atlas to search for the species name and investigate why it did not match. Once the correct name was discovered, the edited species names were updated accordingly. Often it was immediately obvious why the name wasn't matching, for example if "var" has been used instead of "subsp" to indicate a subspecies, as the BSBI tool did not always recognise this.
5. Ensure that the BSBI database matches to sensu stricto/sensu lato/aggregate descriptors properly. In some cases, names did not match to the correct descriptor. With the guidance of botanical experts within the project team, we corrected these names manually. For example, 'Species x' may match to 'Species x sensu stricto' in the BSBI database when actually it should match to 'Species x sensu lato' because it was recorded before the change in nomenclature.
6. Re-format the file to keep key pieces of information, and join species names from each scheme to the new harmonised list of names.

The harmonisation work within this project has proved a valuable outcome in its own right. We have received interest from other projects (including the Defra-supported UK APIENS project) both to use the current dictionary and in further development that would enable incorporation of other vegetation datasets. The species dictionary has been developed into an R package, '*vegtaxon*', which aims to assist with the harmonisation of vascular plant survey data from UK datasets. Similar packages have been developed in R, but none are specific to UK vascular plants. The package is built using the underlying integrated species dictionary, which matches the Latin names of UK vascular plant species to the current accepted name. This is done mostly to classifications provided by Stace's Flora (Stace, 2010). The package '*vegtaxon*' also provides functions to easily calculate commonly applied indicator values such as Ellenberg scores.

The package 'vegtaxon' is now publically available and can be downloaded from Github (https://github.com/NERC-CEH/vegtaxon). It is hoped that it will encourage increased, and more efficient, use of these vegetation datasets, as well as integration of them with others. It should therefore become a valuable resource for the wider ecological community and further work will go into developing its functionalities under the NERC funded programme UK-SCaPE. We also plan to create a separate package which implements National Vegetation Classifications from the Modular Analysis of Vegetation Information System software (MAVIS; Smart et al., 2016) in R. A function which maps user's vegetation species names to the species names used by MAVIS will be added to the '*vegtaxon*' package.

2.5 Cluster analysis

Because of differences in the way habitats are classified between schemes, we devised a clustering method to assist with the selection of heath and bog plots.

We calculated a Bray-Curtis dissimilarity matrix (using the 'vegdist' function from the R package 'vegan') to estimate the pairwise dissimilarities between vegetation plots. This method is commonly used to quantify differences between ecological samples. The non-Euclidean measure assumes that plots are of equal size, however our use of binary presence/absence data rather than count data minimises this issue.

We ran k-medoids clustering on the matrix using the 'pam' (partitioning around medoids) function from the R package 'cluster', imposing a maximum of 12 clusters. A medoid is the most centrally located object of a cluster, i.e. it is the point with the minimum sum of distances to other points within the cluster. Every plot from every visit in all four surveys was assigned separately to one of the 12 clusters. Countryside Survey habitat data was used to guide our selection of clusters, whereby we chose the clusters that contained the majority of CS bog, shrub heath, and acid grassland plots (Table 2.2). Clusters 6, 7, 8 and 10 were subsequently chosen to represent heath and bog habitats. Plots which fell in any of these four clusters at any point in time were included in subsequent analysis. For example, if a hypothetical CS plot was classified as 'bog' in 1978, but as 'bracken' in subsequent survey years, all incidences of this plot regardless of habitat classification were included in the heath/bog dataset. This method of data selection was designed to be broad, in order to encompass the majority of plots present in heaths/bogs in all schemes, including marginal plots in which the characteristics of the assemblage has either shifted towards or away from heath/bog habitat over time.

Table 2.2. Proportion of CS plots in a selection of relevant habitats represented by each cluster

Cluster	Acid.Grass	Arable	Bog	Bracken	Broadleaf	Calcareous.Grass	Fen	Montane	Shrub.Heath
1		0.02		0.05	0.32		0.02		
2	0.03	0.26		0.07	0.15	0.03	0.06		
3				0.08	0.14		0.01		
4		0.55					0		
5		0.06		0.02	0.02	0.01	0.08		
6	0.08	0.07	0.58	0.07	0.03		0.07		0.23
7	0.27		0.24	0.09	0.06	0.03	0.31	0.09	0.19
8	0.43		0.05	0.28	0.09	0.15	0.1	0.55	0.19
9	0.05		0.01	0.24	0.06		0.01	0.18	0.21
10	0.03		0.11				0.02	0.09	0.16
11	0.01			0.01	0.02	0.51			0.01
12	0.09	0.03	0.01	0.09	0.11	0.26	0.32	0.09	0.02

2.6 Covariate data

2.6.1 Climate variables

We obtained climate data from the Met Office HadUK-Grid (Met Office, 2019), which provides a range of gridded climate variables at 1 km resolution, derived from more widely spaced land surface observations. We extracted the annual total precipitation, maximum July temperature, and minimum January temperature for each 1 km square in the UK which contained vegetation data from the CS, ECN, LTMN or NPMS. Met Office data was received in NetCDF format. Climate data for all 1 km squares in the UK containing vegetation scheme data was extracted from this NetCDF file. The HadUK-Grid dataset does not provide data for 1 km squares which intersect with the coastline. To obtain climate data for plots in coastal locations, we used a nearest neighbour approach to find the closest 1 km square for which we had data. If the closest square with data was less than 10 km away, we assigned this data to the coastal square. Where no nearby data was available, climate variables for that plot were set to NA (not available).

Values for temperature and precipitation for each year and grid square were time-standardised as the difference between annual estimates and the long-term mean, such that if the mean across all years for a grid cell was 10 and the value in the year of vegetation survey was 15 then the value attributed to the vegetation plot would be 5.

At the time of access, climate data were only available for 1862-2017. Plots sampled in 2018-2019 were assigned climate variables from 2017.

2.6.2 Atmospheric Deposition

The Defra Nitrogen futures scenario method (Dragosits et al., 2020) involved comparing a FRAME (Fine Resolution Atmospheric Multi-pollutant Exchange) model run for the future to a contemporary year FRAME run, and applying those differences to CBED (Concentration Based Estimated Deposition) model data derived for the same contemporary year. We did this in reverse, using a 1970 LTLS FRAME and a contemporary FRAME. This used a grid average time series of CBED data from 1986 to 2018 and LTLS FRAME data for 1970 and 2010. The resulting datasets contain yearly grid average total nitrogen and total sulphur deposition for 1970 to 2018 on a 5 km grid covering the UK.

Values of deposition variables were standardised before use in analysis, and the rationale for the standardisation is explained in further detail at the start of Chapter 6.2. Nitrogen deposition (1970-2018) was represented as the plot-specific 2003-2005 average Nitrogen deposition. This value is used to represent peak nitrogen deposition load for the grid square across the time series, as in RoTAP (2012). Sulphur deposition to each grid square in each year (1970-2018) was time-standardised as the difference from the long-term grid square mean sulphur deposition. No deposition data was available for 2019, and so plots surveyed in 2019 were assigned deposition variables from 2018.

2.7 Indicator values

In order to summarise the compositional structure of the vegetation plots in relation to environmental pressures, and quantify change in attributes over time, we used a range of vegetation indicator metrics. We focussed on Ellenberg N, Ellenberg R, Ellenberg F/W, CSM positives and CSM negatives. Further detail on the selection of indicators is given in Chapter 3.

Ellenberg indicator values were sourced from the LUS_SP_LIB_AND_TRAITS table in the Countryside Survey Oracle database and integrated with the other schemes using the species dictionary.

As this analysis is looking broadly at heaths and bogs rather than a specific strict habitat type, we compiled generalised Common Standards Monitoring (CSM) positive and negative indicator lists by aggregating the CSM scores for the four most relevant CSM habitat types; lowland dry heath, lowland raised bog and lowland blanket bog, alpine dwarf-shrub heath, and upland blanket bog and valley bog.

Indicator values were calculated for each plot, in each site, in each year. The number of CSM positive and negative species in each plot were totalled to provide the CSM scores. Neither Ellenberg nor CSM indicators were cover-weighted due to differences in cover recording between the schemes.

2.8 Creation of modelling dataset

Data from the four individual schemes were filtered as outlined in Section 2.3 and joined together in a single dataset with the help of the species dictionary described in Section 2.4. Data were aggregated to plot-level, with indicator calculation taking place as outlined in Section 2.7.

The final integrated dataset included 8490 individual plots, with CS data comprising by far the largest proportion of the data (Table 2.3). This integrated dataset cannot be made publically available due to the confidential nature of the CS data.

Table 2.3. Summary of the aggregated plot data used in the analysis.

Dataset	Proportion of plots in integrated dataset (%)
CS	65
ECN	15
LTMN	15
NPMS	5

3 Indicator selection

3.1 Introduction

The selection of indicators was carried out in association with other work in the MoA on the 25 YEP indicator D1: Habitat quantity, quality and connectivity on the habitat quality component (Maskell et al. 2020, 2021). D1 sits within the "wildlife" theme and the "thriving plants and wildlife" goal and contributes to headline 7 "Changes in nature on land and water that support our lives and livelihoods". It is an indicator for all terrestrial habitats: Priority habitats[1] as well as habitats less rich in wildlife (contributing to the matrix of a habitat network) forming an environmental system providing wider benefits.

As part of that work, potential indicators were reviewed from a range of sources; those currently being used to monitor habitat condition within different organisations and those under development. A long list of potential indicators was used as a basis for a stakeholder workshop discussion around the development of quality indicators held in October 2019.

Representatives from Natural England, Defra, BTO, RSPB, National Trust, CEH, Environment Agency, Forestry Commission, JNCC and the Woodland Trust attended the workshop. Many organisations have been working on metrics of habitat condition/quality, and the aim of the workshop was to share ideas and develop a consensus on appropriate measures of habitat quality for this indicator.

Following the workshop an indicator framework was created using outputs from the workshop and subsequent discussions with stakeholders. Six functional elements were identified; nutrient status, plant species composition, vegetation structure, naturalness of hydrology, habitat heterogeneity and soil health. Within each habitat, appropriate indicators for each functional element were considered to create a shortlist. For the purposes of this task (task 4) we used the indicators for Bog and Heathland and selected the most appropriate to reflect the potential driver-linked changes in those habitats. These indicators were Ellenberg N, Ellenberg R, Ellenberg W, CSM positive richness and CSM negative richness.

3.2 Indicators

Ellenberg scores (Ellenberg et al. 1991) allow the assessment of environmental conditions without direct measurements (Diekmann 2003). They were developed as an indicator system for vascular plants of central Europe (Ellenberg, 1979; Ellenberg et al., 1991) and are based on a simple ordinal classification of plants according to the position of their realized ecological niche along an environmental gradient. They describe the response of individual species to a range of ecological conditions (light, temperature, fertility, moisture, pH). Ellenberg scores were adapted for application to UK plants (Hill et al. 2000) and are available for a wide range of higher plants and bryophytes. Preferences are pre-classified and an average score is calculated for each study community on the basis of individual species preferences. Scores can be weighted by cover (hence, relative dominance) of individual species, however, we have not done this because of the difficulties associated with estimating cover for the different schemes. The advantages of Ellenberg scores are that they are relatively easy to calculate, and reflect environmental conditions that allow, or restrict, the occurrence of species at a site over long time periods (Bartelheimer and Poschod 2016).

[1] https://jncc.gov.uk/our-work/uk-bap-priority-habitats/

3.2.1 Ellenberg fertility score

The Ellenberg fertility (or Ellenberg N) score for a vegetation plot is determined as the mean of the Ellenberg fertility scores of all species present, providing an overall community score on a scale of nutrient poor (1) to nutrient rich (10) (Sutton et al. 2004). The sensitivity of Ellenberg fertility metric has been tested in nitrogen addition experiments. Ellenberg fertility scores have been shown to reflect plant and soil nitrogen status (Hill, 2000; Smart et al.,2003; Diekman, 1995), and used in many different contexts, from local (Pitcairn, 1998) to regional and national scales (Smart et al., 2003). An increase in an Ellenberg fertility score of a community indicates a floristic shift consistent with eutrophication, and scores can therefore be used to attribute changes to eutrophication compared to other potential drivers, such as understanding the impact of atmospheric nitrogen deposition on vegetation (RoTap, 2012). Ellenberg fertility scores might be expected to be more sensitive to subtle vegetational shifts than standard CSM monitoring (Emmett et al. 2011), and may therefore reveal ongoing damage even at sites judged to be in favourable condition.

3.2.2 Ellenberg R

The soil reaction gradient (Ellenberg R) ranges from “strong acidity, never moderately acidic or alkaline” (R-number 1) to “alkaline and calcareous conditions, only calcareous soils”(R number 9) (Bartelheimer and Poschod, 2016).

3.2.3 Ellenberg Moisture

Ellenberg moisture (or Ellenberg W) scores range from very dry conditions with a low moisture score (1) to wet soils (9); scores 10-12 represent aquatic communities. Ellenberg moisture scores can therefore be used to track the impacts of hydrological change on vegetation over time. More direct indicators of wetness, e.g. soil moisture, have the disadvantage of greater short-term variability and sensitivity to antecedent weather conditions.

3.2.4 Ellenberg Light

Plant species are assigned an Ellenberg Light (Ellenberg L) indicator values on a scale from 1 to 10. This indicates position along an environmental gradient from heavy shading in late successional habitats (1), e.g. underneath woodland canopy, to strongly illuminated (often disturbed) open habitats (10).

3.2.5 Positive and negative habitat quality indicators

In Common Standards Monitoring, positive and negative indicator species were originally selected on the basis that they are typical or distinctive for the habitat; are useful for determining site condition; are not so scarce that they will rarely be observed; and occur across a wide geographic range (Rowe et al., 2016).

Positive and negative habitat indicators are useful for assessing how the vegetation studied compares to a reference habitat type. Negative indicators may indicate that abiotic conditions are changing, perhaps indicating eutrophication or disturbance. Dominance by a small number of negative indicators can have significant impacts on a site. However, there is a degree of subjectivity in the choice of indicator species, and more work would be useful to agree the species lists with habitat representatives.

A study by Rowe et al. (2016) combined qualitative semi-structured interviews with conservation professionals specialising in grasslands, heathlands and mires, and quantitative ranking of example habitat communities, in order to determine the best metric for assessing habitat quality. The specialists’ rankings were compared to metrics calculated from the data, including total species

richness, positive and negative indicators, % forbs, sphagnum cover, DSH cover, and a metric based on Ellenberg fertility. The number of positive indicator-species was the metric most consistently associated with specialists' rankings, although there was also some agreement with respect to Sphagnum cover in bog habitats.

Positive habitat quality indicators have been used in reporting from the Glastir Monitoring and Evaluation project as an indicator of habitat quality (https://gmep.wales/) and the Natural Capital maps produced for NE[2], and will be used for the current work under ERAMMP (https://erammp.wales/en).

The number of CSM positive and negative indicators have been derived by extracting indicator species used in Common Standards Monitoring guidance for Sites of Special Scientific Interest (JNCC) and refined in consultation with the Botanical Society of Britain & Ireland to create a list of plants indicative of habitats of high conservation value (Maskell et al., 2019).

[2] https://eip.ceh.ac.uk/naturalengland-ncmaps/reports/diversity_report.pdf

4 Comparison of trends in vegetation metrics between individual schemes

4.1 Introduction to the scheme comparison

The overall aim of this project is to explore the potential for combining long-term vegetation monitoring and survey data, and associated covariate data, within single statistical analyses in order to maximise our understanding of vegetation change across the UK over recent decades. We have focussed on the integration of data from the four vegetation schemes (Countryside Survey, Environmental Change Network, Long Term Monitoring Network, National Plant Monitoring Scheme), and heath and bog habitats as the target for this case study.

Our integrative approach is novel, but imposes some constraints on the complexity of trends that can be quantified. Because the programmes above were initiated at different times we were only able to assess linear changes, i.e. working on the assumption of linear increases, decreases or stability over time. We consider this reasonable given that that the dominant drivers of change over recent years are likely to have exerted largely uni-directional effects on vegetation composition. For example, progressive reductions in acid deposition ever since the 1970s are expected to be leading to gradual, if possibly lagged, reductions in soil acidity, while there is an assumption that atmospherically deposited nitrogen has continued to accumulate in terrestrial ecosystems, resulting in a gradual eutrophying effect. Likewise, climate change is likely to be exerting a gradual, although possibly stepped, influence on air and soil temperatures. In some circumstances, particularly with respect to change in CSM scores, it is possible that trends will be more nuanced, involving periods of both increases and decreases, and possibly more geographically dependent. However, it was not possible to investigate these relationships with the datasets and modelling structure used.

Before beginning to build integrated models it is important to investigate datasets individually to inform any integration. There are two key questions we need to ask of the data:

1. What characteristics of each scheme need to be accounted for when modelling individually? For example, if the scheme design includes the nesting of plots within larger areas such as sites or 1 km squares, we need to account for this structure in the individual dataset models and in the integrated models.
2. Do our individually fitted models for each indicator of interest support the assumption of a common trend across schemes? This is the most important question to decide whether we can integrate the different schemes or not. If we believe that the assumption of a shared underlying trend is not met (e.g. because the datasets show very different trends over time) then it would not be valid to combine the datasets.

To answer these questions we conducted exploratory data analysis and fitted some simple temporal models to each dataset. The purpose of these models is to assess similarity in trends between schemes rather than to attribute change, so no covariates are included within them. We would not expect our estimated trends to be identical to those reported previously for these schemes due to our exclusion of covariates at this stage and the custom selection of heath and bog plots defined in Section 2.

4.2 Exploratory data analysis

We explored the five indicator metrics selected according to the rationale set out in in Chapter 3. These indicators were Ellenberg N, Ellenberg R, Ellenberg W, CSM positive richness and CSM negative

richness. Ellenberg scores for each plot were calculated as the unweighted means of scores for all species present for which indicator scores were available. CSM richness was calculated as the number of unique species per plot, assigned either positive or negative indicators on the basis of the lists described in Chapter 3. For each scheme we selected heath and bog plots as outlined in Section 2 and calculated the five indicators for each vegetation plot.

Data exploration involved six key elements which are described briefly here. Full details of exploration for each scheme can be found in Appendix 1. For each scheme we assessed:

1. The distribution of each selected indicator using histograms. Inspection of these plots informed the choice of distribution for the trend models
2. Plots of each indicator over time. These plots informed our treatment of time in the models e.g. as a continuous or discrete variable. It also allowed us to identify any gaps in the time series and consider how to account for these in autocorrelation structures
3. Plots of each indicator by site. It might be necessary to account for site identity in our models by including a random effect. Plotting indicators against site allowed us to evaluate the potential variation explained by site identity
4. Plots of each indicator over space. These plots allowed us to assess the spatial coverage of each scheme and give a preliminary indication of potential spatial patterns in indicator values
5. Summaries of the number of repeat visits to each plot to inform how to build the models.
6. Summaries of the number of plots per site. The schemes vary substantially in the maximum number of plots that occur within a site. It may be wise to fit a site level random effect for schemes with high numbers of plots per site. For schemes with very low numbers of plots per site, a site level random effect is not likely to provide any benefit.

Examples of the plots described in points 1-4 are given below for a selected scheme and indicator (Ellenberg R in the LTMN; Figure 4.1).

On the basis of the exploratory plots and summaries, we concluded that the same basic model structure could be fitted to each scheme dataset (Eq 1).

$$\text{Indicator} \sim \text{Year} + (1|\text{site/PlotID}) \qquad \text{(Eq 1)}$$

This model accounts for similarity between multiple revisits to the same plot, and similarity between plots within the same square. Ellenberg indicators were modelled with a normal distribution, while CSM indicators were modelled with a Poisson distribution. Ideally an autocorrelation term conditional on plot identity would be fitted. This term would account for the fact that when a plot is revisited, the vegetation recorded is likely to be similar to the last time it was visited. Similarity would be expected to decay over time i.e. the longer the time between visits the less likely the vegetation would be similar. Unfortunately, we could not fit this model to all schemes because of the relatively small number of repeat visits to each plot. Even for CS, the scheme with the longest history, only 2% of plots were revisited 5 times (i.e. every survey).

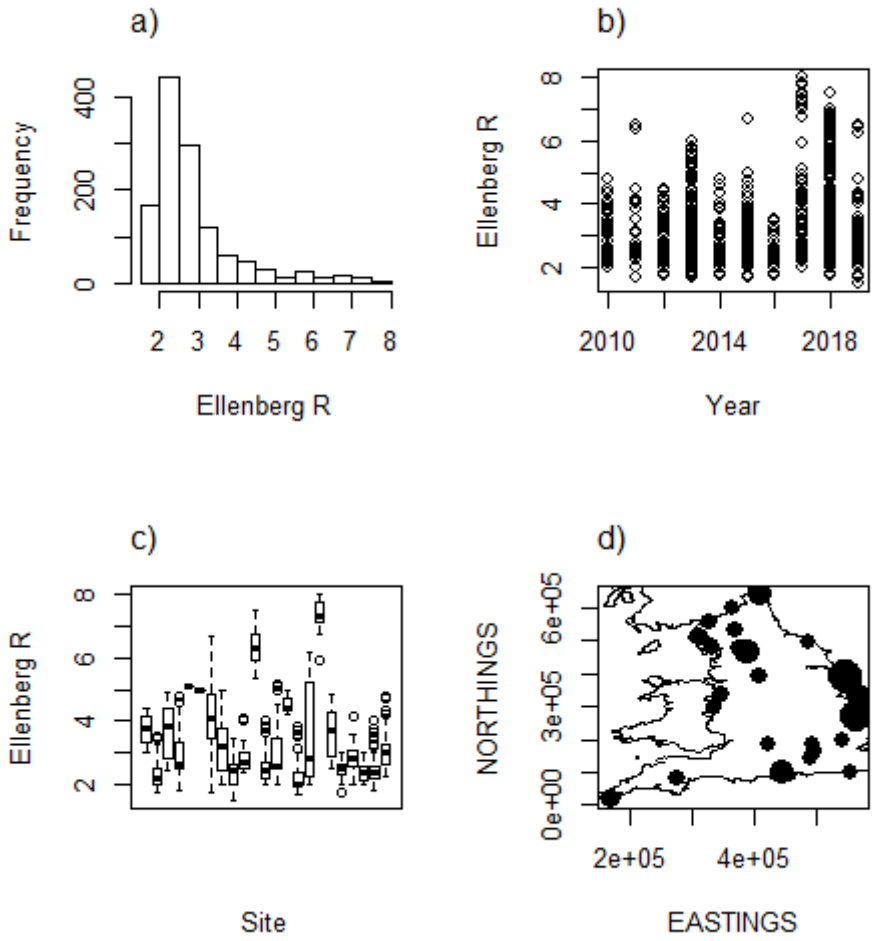


Figure 4.1. Examples of a) indicator histograms, b) plots of indicators of time, c) boxplot of indicator against site, d) exploration of geographic spread of indicators where larger dots indicate greater mean indicator value. All plots are shown for Ellenberg R from the LTMN scheme.

4.3 Results of individual scheme models

By fitting the model described in Section 4.2 to each of the five selected indicators for each scheme individually, we calculated trends over time that could then be compared between schemes. In these models we have included the variable "year" as the year since the onset of the scheme rather than the calendar year.

4.3.1 Countryside Survey models

Models of the CS data indicated a very small increasing trend in Ellenberg N and R (of 0.006 and 0.007 units per year respectively; Figure 4.2). This is suggestive of gradual shifts in heath and bog assemblages towards those indicative of more nutrient rich and less acid environments respectively.

There was also some evidence of an increasing trend in Ellenberg W of 0.002 units per year, indicating a trend towards species with a preference for wetter conditions. Both CSM indicators showed a decreasing trend over time, which could potentially suggest an overall decrease in richness.

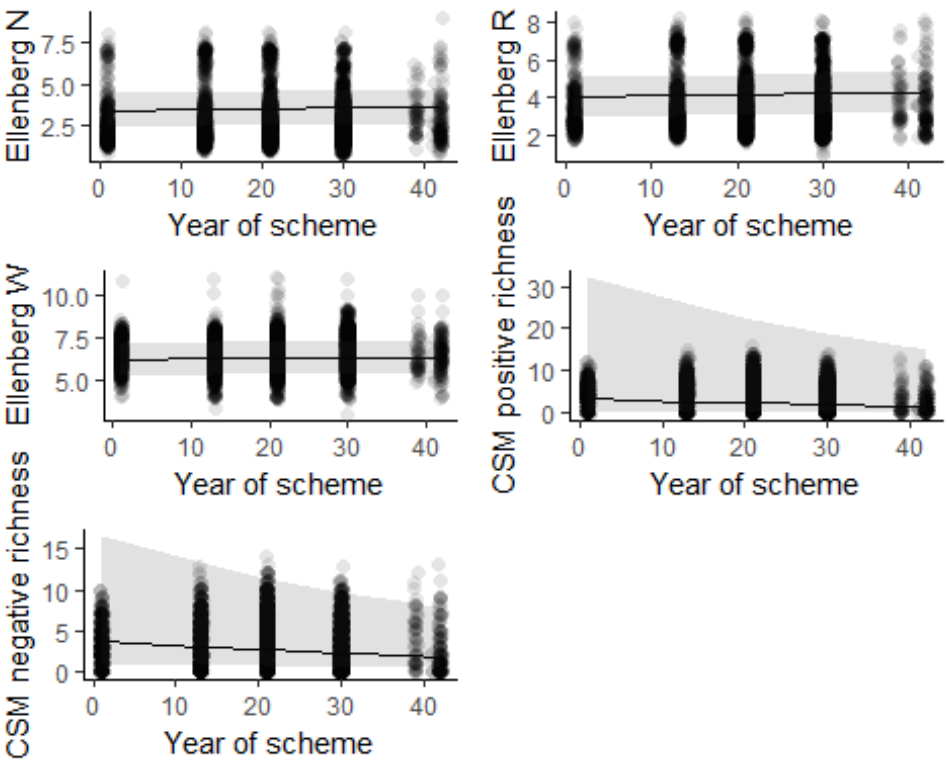


Figure 4.2. Estimated trends in each indicator over time in the CS dataset with prediction intervals.

4.3.2 Environmental Change Network models

The ECN models also provided an increasing trend in Ellenberg N of 0.006 units per year, (i.e. a very similar trend to that observed in the CS data). However, no significant trends were observed in the other indicators (Figure 4.3).

4.3.3 Long Term Monitoring Network

In common with CS, the LTMN models again yielded very gradual positive trends in both Ellenberg N and Ellenberg R (Figure 4.4). The increase in both Ellenberg indicators was around 0.008 units per year, a comparable rate to that estimated by the Countryside Survey dataset (and in the case of ECN, Ellenberg N only). Trends in both indicators were statistically significant.

4.3.4 National Plant Monitoring Scheme

The NPMS data used in this work only cover 5 years (from 2015 to 2019) and therefore it is no surprise that there was no evidence of trends over time (Figure 4.5).

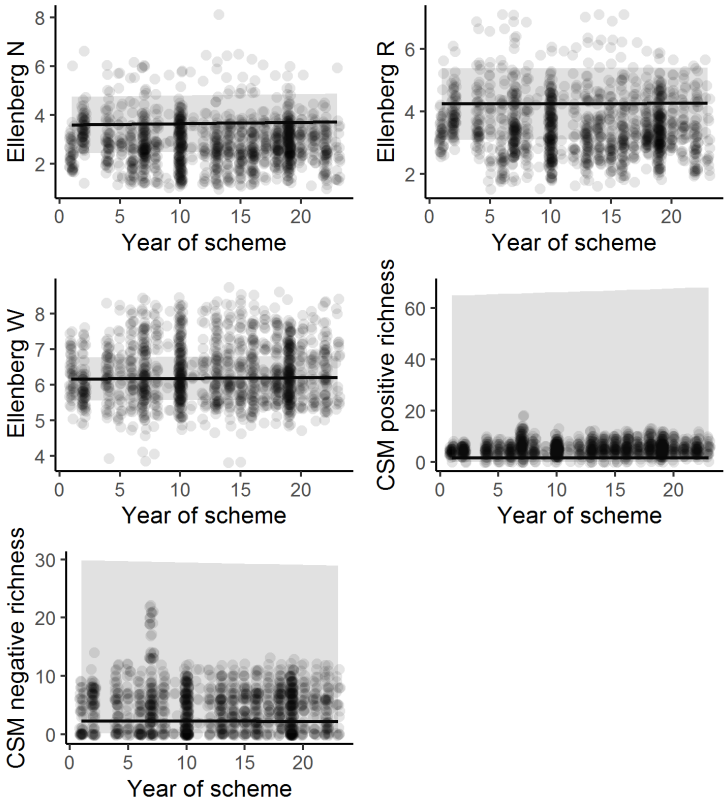


Figure 4.3. Estimated trends in each indicator over time in the ECN dataset with prediction intervals.

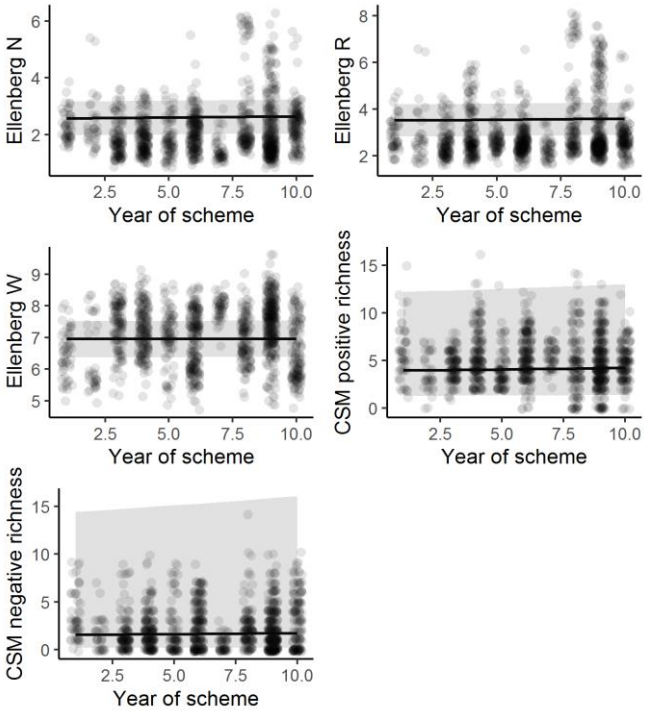


Figure 4.4. Estimated trends in each indicator over time in the LTMN dataset with prediction intervals.

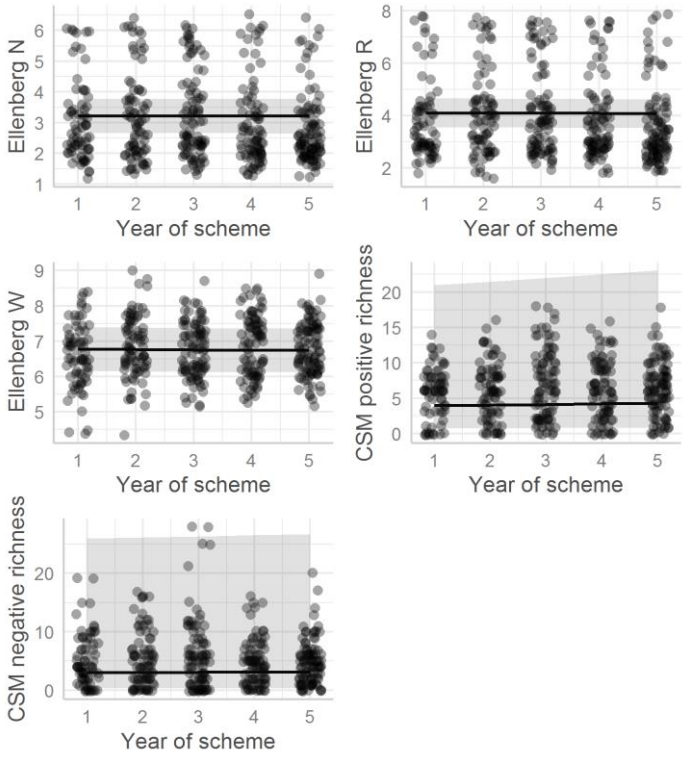


Figure 4.5. Estimated trends in each indicator over time in the NPMS dataset with prediction intervals.

4.4 Summary of individual scheme trend models

We found that trends between schemes in the five selected indicators were generally very similar (Table 4.1) for all but the shortest lived scheme (NPMS). In particular, three out of four schemes showed a positive trend in Ellenberg N over time, with rates of between 0.006 and 0.008 units per year. The NPMS scheme did not show a positive trend in Ellenberg N over time, but this is not surprising as only five years of data were available.

There was more divergence in trends observed for the other indicators. Ellenberg R increased by around 0.008 units per year in the CS and LTMN schemes, but did not change significantly in the ECN scheme. Only CS reported significant trends in CSM indicators, with both positive and negative indicators declining over time. The CS was also the only scheme to show a significant trend in Ellenberg W. Previous analysis of the ECN vegetation data bulked into “upland”, “lowland” and “woodland” categories indicated an increase in Ellenberg W in the lowland category only (largely comprised of unimproved and improved grassland). Our analysis estimated an average increase of 0.002 Ellenberg W units per year in the ECN data, but this was not statistically significant (P = 0.06).

Overall the results of the individual trend analyses provide the justification for creating an integrated model, particularly with respect to Ellenberg N and R. No schemes directly contradicted each other in terms of observed trends, and in some cases the estimated trend coefficients were almost identical between datasets.

Table 4.1. Summary of the trend estimates for each indicator from each scheme modelled individually. Estimates of the year coefficient are given followed by standard errors in brackets and P values. Comparison of trends between schemes

Indicator	CS	ECN	LTMN	NPMS
Ellenberg N	0.004 (0.001) *P* < 0.001	0.005 (0.001) *P* < 0.001	0.006 (0.001) *P* < 0.001	-0.001 (0.007) *P* = 0.832
Ellenberg R	0.007 (0.001) *P* < 0.001	0.001 (0.001) *P* = 0.534	0.008 (0.003) *P* = 0.003	-0.009 (0.011) *P* = 0.394
Ellenberg W	0.002 (0.001) *P* = 0.036	0.002 (0.001) *P* = 0.057	0.002 (0.003) *P* = 0.573	-0.010 (0.014) *P* = 0.452
CSM positive indicators	-0.018 (0.001) *P* < 0.001	0.002 (0.001) *P* = 0.259	0.007 (0.005) *P* = 0.182	0.237 (0.018) *P* = 0.177
CSM negative indicators	-0.018 (0.001) *P* < 0.001	-0.001 (0.002) *P* = 0.493	0.012 (0.008) *P* = 0.109	0.007 (0.019) *P* = 0.717

In addition to assessing similarity in terms of estimated trend coefficients (see Table 4.1.1), it is also possible to plot estimated trends for all schemes. To do this, we refitted the models above replacing 'Year of scheme' with calendar year to allow direct comparisons.

4.4.1 Comparison of Ellenberg N trends

As expected from comparisons of model coefficients, we found that plotted trends of Ellenberg N over time were very similar between schemes (Figure 4.6). Plotting the data also demonstrates differences in uncertainty around the trends, with confidence intervals around the CS trend being much smaller than those around the NPMS trend, reflecting differences in the amount of data available for each scheme. Intercepts also differed between schemes, with the average Ellenberg N score lowest in LTMN plots.

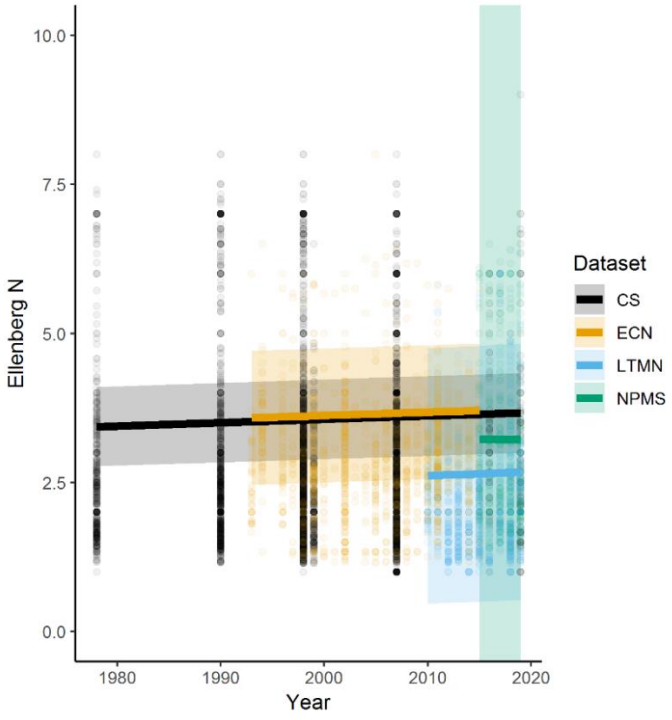


Figure 4.6. Trends over time in Ellenberg N for each vegetation monitoring scheme. Confidence intervals are shown around the fitted trend.

4.4.2 Comparison of Ellenberg R trends

Similar to our observations regarding Ellenberg N trends, the plotted trends over time in Ellenberg R showed a good degree of similarity between schemes (Figure 4.7). Confidence intervals were much greater for NPMS than the other schemes, reflecting high uncertainty in estimated trends with only 5 years of data. LTMN plots tended to have lower Ellenberg R scores (more acidic) on average than plots from other schemes.

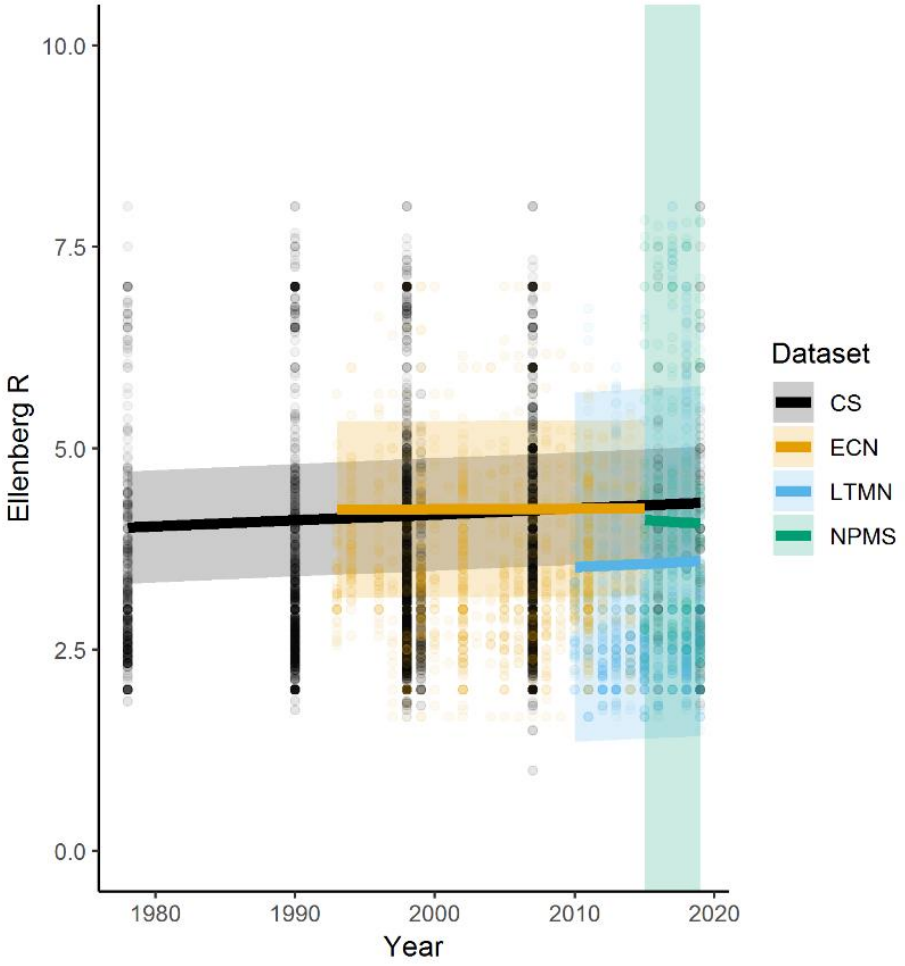


Figure 4.7. Trends over time in Ellenberg R for each vegetation monitoring scheme. Confidence intervals are shown around the fitted trend.

4.4.3 Comparison of Ellenberg W trends

The plot of Ellenberg W trends provides less evidence of consistent cross-scheme change in this metric in comparison to Ellenberg N and Ellenberg R scores (Figure 4.8). The models of Ellenberg W suggested smaller changes over time in Ellenberg W values, suggesting this indicator may be more stable.

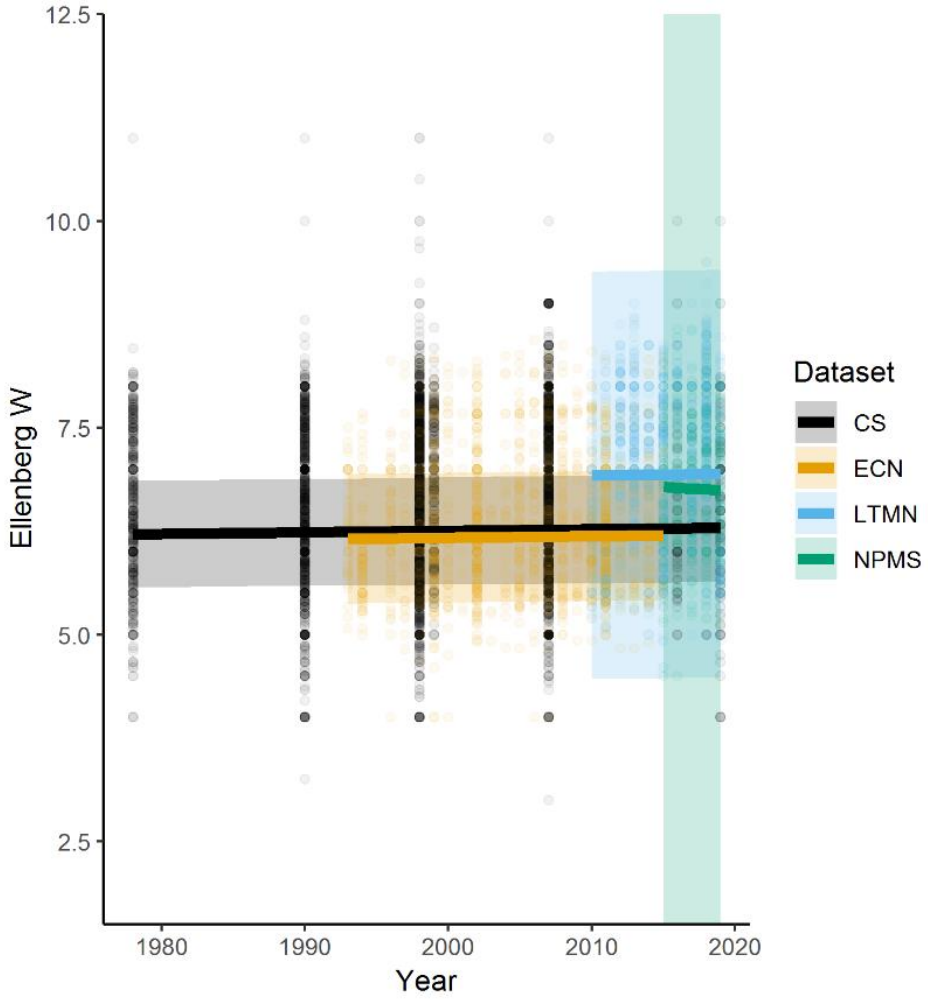


Figure 4.8. Trends over time in Ellenberg W for each vegetation monitoring scheme. Confidence intervals are shown around the fitted trend.

4.4.4 Comparison of trends in CSM positive indicator richness

Although Countryside Survey was the only scheme to show a significant trend in CSM positive indicator richness over time, when plotted it is clear there is quite a lot of dissimilarity in the relationships occurring in different schemes (Figure 4.9). Uncertainty around the trend in CSM positive richness was highest for ECN, where the model estimated no change in CSM richness over time. CS estimated a strong decrease in richness, whereas NPMS estimated a strong increase despite only five years of data being available.

4.4.5 Comparison of trends in CSM negative indicator richness

Trends in CSM negative indicators also differed markedly between schemes, with a negative trend in CS not mirrored by any of the other schemes (Figure 4.10).

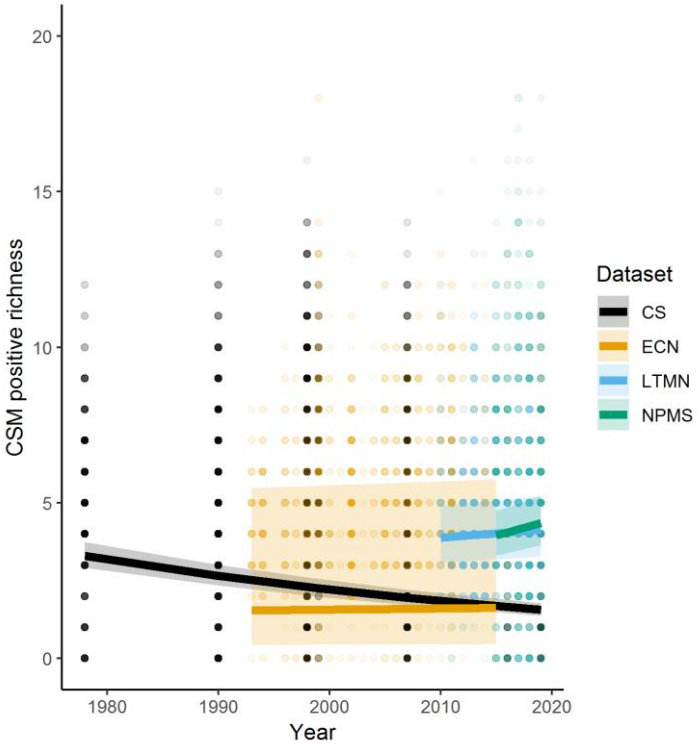


Figure 4.9. Trends over time in CSM positive indicator richness for each vegetation monitoring scheme. Confidence intervals are shown around the fitted trend.

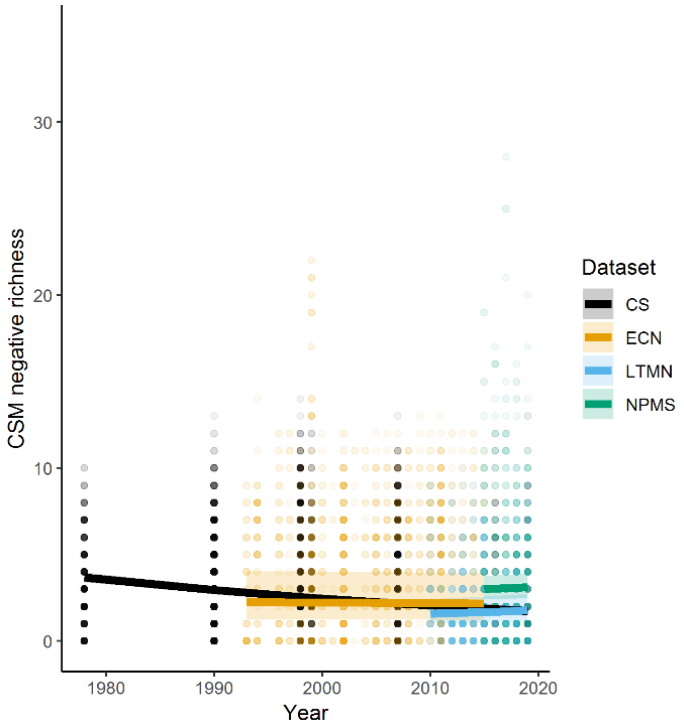


Figure 4.10. Trends over time in CSM negative indicator richness for each vegetation monitoring scheme. Confidence intervals are shown around the fitted trend.

5 Vegetation simulation

5.1 Introduction to the simulation of vegetation communities

When bringing together data from multiple monitoring schemes in an integrated analysis, it is important to account for any differences in sampling protocols across the schemes that may influence the distribution of the observations of interest. This is because any integrated analysis will assume that at least one model component, such as the global mean, trend, variance, or other covariate effects, is common across all schemes. If significant differences in the distribution of the data between schemes are not accounted for, these jointly estimated components may not be robust and could be substantially biased.

Across the schemes considered here, namely the Countryside Survey (CS), National Plant Monitoring Scheme (NPMS), Environmental Change Network (ECN) and the Long Term Monitoring Network (LTMN), a range of different vegetation survey protocols have been adopted. Prior to any integrated analysis we need to understand the impact these have on the distribution of any derived indices. Some differences may have little or no impact on the data obtained and the derived indicators, some may have significant impact on species level data but less influence on derived indicators, and others still may influence both the species level data and derived indicators. Those protocols that do not affect the indicators under consideration here can be overlooked within any analysis, whereas those that have substantial impact must be accounted for.

To investigate differences in the influence of the protocols on observations, one could conduct a large-scale experiment whereby a substantial number of sites are independently surveyed according to the various protocols. The resulting data could then be compared across protocols using the sites as replicate samples to aid comparison. There are two major drawbacks to this however, the first being the financial cost. With potentially hundreds of replicate samples needed and multiple protocols to apply independently (i.e. ideally at separate visits by separate survey teams), the surveyor effort and practical cost becomes prohibitive. Second, distinguishing the particular effect of any individual protocol from another via physical survey can be very challenging. To do so would require a multi-factorial experimental design with a large number of factor levels.

An alternative mechanism for establishing differential impacts of the various sampling protocols is to simulate realistic vegetation patches and then “pseudo sample” from that according to different hypothetical protocols. This computational approach offers the advantages of being relatively cheap, efficient and quick to perform compared to field survey alternative, as it enables comparison of individual protocols, sets of protocols, and any such mixture. It does, however, rely on an ability to simulate realistic vegetation patches in the first place.

Here, we used a simulation-based approach to determine the impact the various sampling protocols adopted across the CS, ECN, LTMN and NPMS schemes would be expected to exert on indicator responses. We sought to quantify any effects, whilst also understanding whether it was possible to account for them within an integrated model and whether some indicators are more robust to these differences than others.

5.2 The conceptual approach

To focus the simulation study around a few critical elements of the sampling protocols across schemes, we considered the aspects of: i) plot size ii) cover estimates iii) relocation error, and iv) accuracy of surveyors. Plot sizes vary across the four schemes with the CS, LTMN and ECN (coarse-scale plots)

covering a 2 x 2 m area, whereas the NPMS uses a 5 x 5 m quadrat size and the ECN fine grain plots cover a 10 x 10 m area. Plot size was hypothesised to have a significant impact on habitat condition indicators due to the well documented, explicit species-area curve relationships. As many of the indicators are typically cover-weighted, the way in which cover of individual species is estimated across the schemes was also investigated with different scaling categories used across CS and NPMS and ECN and LTMN using a cell-count system. Finally, two components of error, plot relocation error and surveyor error, were investigated because of the potential impact on sensitivity to detect change.

To simulate vegetation patches effectively, we needed information on:

- How many species in total we might expect
- Which species are more common than others
- Which species are most likely occur with others
- How clustered or regularly distributed species are

As each of these aspects vary significantly across landscapes and habitats, it was deemed necessary to identify an example "target" community that the simulated patch was to represent. Having selected this target community, the necessary information on species composition could be extracted from existing sources to parameterise the simulation and obtain realistic vegetation patches.

Based on this information, we established an overarching approach to investigate the effect of the protocols in use across the schemes on resulting habitat condition indicators. For each aspect being investigated (e.g. plot size, cover estimation, surveyor error or re-location error), 1000 independently simulated patches of vegetation were generated. This provided a large number of replicates with which to formally test and evaluate the impact of different protocols. Each of the 1000 patches were sampled according to the protocols of the schemes for the aspect(s) under investigation. All other differences in protocols were held constant so that the specific effect could be isolated. Finally, indicator metrics were derived from these subsets of data, all taken from the same simulated patch, and formal comparisons were made between them.

5.3 The simulation routine

To simulate patches of vegetation we constructed a hypothetical grid of 5000 by 5000 cells to represent a 50 m by 50 m patch of vegetation divided into 10 cm cells. The premise of the simulation was to fill each of these cells with particular species, including bare ground, such that the total number of species, the composition of species, the cover of species and the clustering of species were all realistic. The Countryside Survey (CS) were used to parameterise the simulations, provide the information required as detailed in the previous section, and provide a reality check against which the resulting patches could be compared. The detailed steps used to simulate vegetation patches are described below.

5.3.1 Step 1: Defining the target population

The CS data used to parameterise the simulations were subsetted according to the type of community to simulate. Initially, focus was on a broad subset of heath and bog plots, but more refined subsets were also selected using the ordination described in Section 2.5 to establish a transitional gradient of different populations. In this case, a region of the ordination space was selected (as shown in Figure 5.1) and a minimum of 20 plots extracted that are close to this point to reflect the community. It was these selected subsets of CS plots that then defined the target population of interest and hence the species pool, and their relative cover, with which the simulated grid was populated.

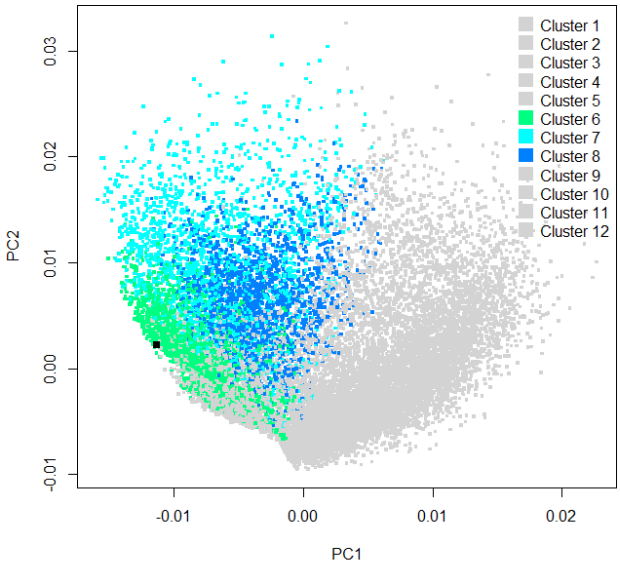


Figure 5.1: Ordination plot of vegetation composition from all scheme data with identified health and bog clusters (6, 7 and 8) coloured in. Added to this is a hypothetical point to define a target population around which to base simulated vegetation patches.

5.3.2 Step 2: Estimating the total species number

The X plots used within the Countryside Survey record species composition within a nested set of five plot sizes (4 m^2, 25 m^2, 50 m^2, 100 m^2 and 200 m^2). This provides information on the total number of species within each of these areas. This information was used to establish a relationship between the number of species and the area by fitting a linear model to the log-log relationship of the two variables. This simple linear model was then extrapolated to estimate how many species should be present in a 50 m x 50 m plot.

5.3.3 Step 3: Determining the species pool

The subset of CS data obtained in Step 1 was then used to establish the overall species pool within our hypothetical 50 m x 50 m vegetation patch. Species were randomly sampled with probability proportional to their average cover so that species with consistently high cover values in plots indicative of the target community were more likely to be sampled than rarer species. Once a given species was selected, the subset of plots from which to sample was reduced to only those in which that species occurred to ensure that the composition of species was appropriate. Species were iteratively sampled in this way – selecting one at a time, followed by subsetting – until the required number of unique species, as determined by the fitted species-area curve in Step 2, was reached.

5.3.4 Step 4: Assigning cover values

For each species within the species pool established in Step 3, a cover value was randomly sampled from the observed set of all plots comprising the subset target population established in Step 1. This then gave a list of an appropriate number of unique species, with the appropriate composition of co-occurring species, and an assigned total cover value for each.

5.3.5 Step 5: Populating the grid

All the cover values were then rescaled to sum to 100% before the corresponding number of 10 cm cells was selected across the grid to assign to each particular species. Cells were selected at random, but with neighbouring cells preferentially selected according to the species present. This maintained an appropriate level of clustering for each species. An example of a simulated vegetation patch is shown in Figure 5.2 with each colour representing a hypothetical species. This demonstrates the concept of defining a grid of cells that are then populated by different species.

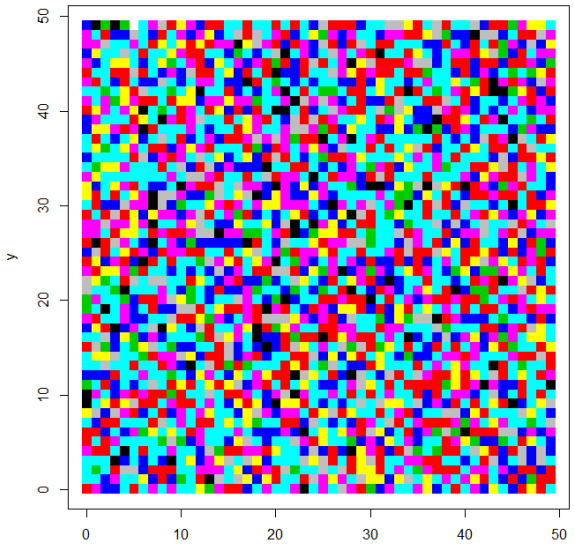


Figure 5.2: An example of a simulated vegetation patch sized 50 m by 50 m. Colours represent hypothetical species. Note that this is shown at a 1 m cell scale rather than 10 cm to enable visual inspection.

5.3.6 Step 6: Reality checking

Finally, having filled the grid cells with species and generated a hypothetical patch of vegetation, a couple of checks were undertaken to ensure that the simulated data provided a realistic representation of observed data. To do so, we plotted the number of species against area for a number of nested areas within the simulated 50 m by 50 m area. The fitted species area relationship estimated in Step 2 was overlaid in order to assess whether the simulated data was adequately capturing this. In addition, an ordination of the subset of vegetation plots was compiled and the simulated plot added passively to test whether it fell appropriately within the multivariate space. An example of these plots is shown in Figure 5.3, where the extracted species-area values agree approximately with the fitted relationship, and the multivariate representation of the species composition falls within the space defined by the target population.

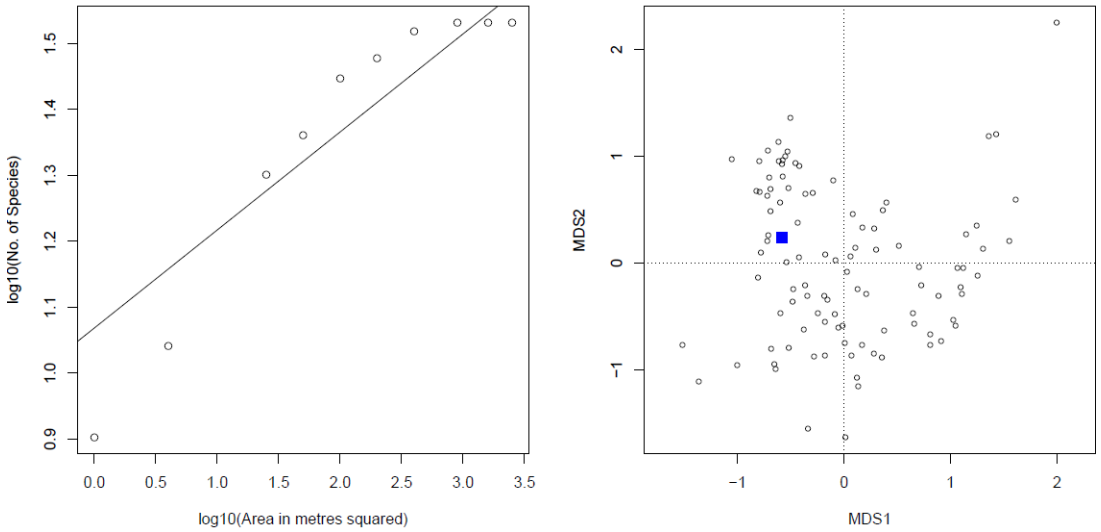


Figure 5.3: Left hand plot - species area relationship (on log-log scale) estimated from the Countryside Survey nested X plot data (shown as solid black line) with values added from simulated vegetation patch (open circles). Right hand plot – ordination of vegetation species composition from target population plots in the Countryside Survey data (open circles) with the simulated patch passively added in (blue square) to enable comparison.

5.4 Sample according to scheme protocols

5.4.1 Sampling

Having simulated a hypothetical patch of vegetation across the 50 m by 50 m grid, we then proceeded to sample it according to the various scheme protocols. To represent the plot sizes of each scheme, the corresponding area was simply extracted from the 50 m by 50 m simulated patch, with all pseudo plots sampled in this way, centred on the central point of the simulated grid. Cover estimates representing CS and NPMS were taken to be the proportional cover across the extracted pseudo plot for each species, whereas for ECN and LTMN the presence of each species within 40 cm by 40 cm cell blocks was established and then up-scaled to provide cover estimates. This reflects the protocols adopted by these schemes.

Surveyor accuracy was included across the schemes by adding in some random noise to the cover estimates assigned to species. This noise was typically small, but varied across the schemes with greater uncertainty assigned to the NPMS data – in acknowledgement that this is an entirely volunteer-based survey, and lower uncertainty to CS where there is greater standardisation of expertise in surveyors and a high degree of quality control procedures in place.

Re-location error was investigated by randomly changing the point in the 50 m by 50 m patch on which the pseudo plot was centred.

5.4.2 Comparison

We compared the effect of plot size, cover estimation, surveyor accuracy and relocation error. To do so, 1000 patches of vegetation within a 50 m x 50 m area were simulated. Each was subsetted according to the protocols of the respective schemes and aspects under investigation. Indicator metrics of Ellenberg N, R and W and the number of CSM negative and CSM positive species were calculated from the generated pseudo vegetation plots.

Boxplots of the indicators across the 1000 simulated samples were then produced and compared, and formal statistical tests (specifically, multiple sample Anderson–Darling tests) were applied. Formal statistical tests comparing the distributions across 5 samples can, however, be overly sensitive, so only significance levels below 0.01 were noted.

5.5 Results

In the first comparison, we were interested in the effects of plot size across the schemes. Sampling of the 1000 sets of simulated vegetation patches to represent each scheme, according to the routine in Section 5.3, differed only in terms of area. All other potential differences between the schemes were held constant to allow comparison of the effects on indicators resulting from the differences of plot size alone. The resulting boxplots are shown in Figure 5.4. It should be noted that since only plot size differed, the schemes with the same plot size, namely CS, LTMN and ECN coarse scale (VC) are based on exactly the same data and therefore have identical boxplots.

Few other features were immediately apparent from the boxplots in Figure 5.4. The Ellenberg scores all showed reasonable consistency across the schemes in terms of the mean value, the only difference appearing to be the smaller variation in the NPMS data – a consequence of this being a larger plot that is less sensitive to the differences across the 1000 simulated sets. The CSM indicators showed considerable differences in mean values using ECN fine scale (VF) and the NPMS methodology - far less than and far greater than the other schemes respectively. This is because CSM indicators are a species count measure and therefore the area surveyed has a direct impact on the number observed, according to species-area logic. The ECN fine scale plots cover a smaller area, 160 of the 10cm cells, compared to 400 for CS, LTMN and ECN coarse scale, whilst the NPMS covers 2500 of the 10cm cells. Formal statistical comparisons across these results showed that only the two CSM indicators provided significant evidence to reject the null hypothesis that the distribution of values across the different schemes was the same ($p<0.0001$ in both cases).

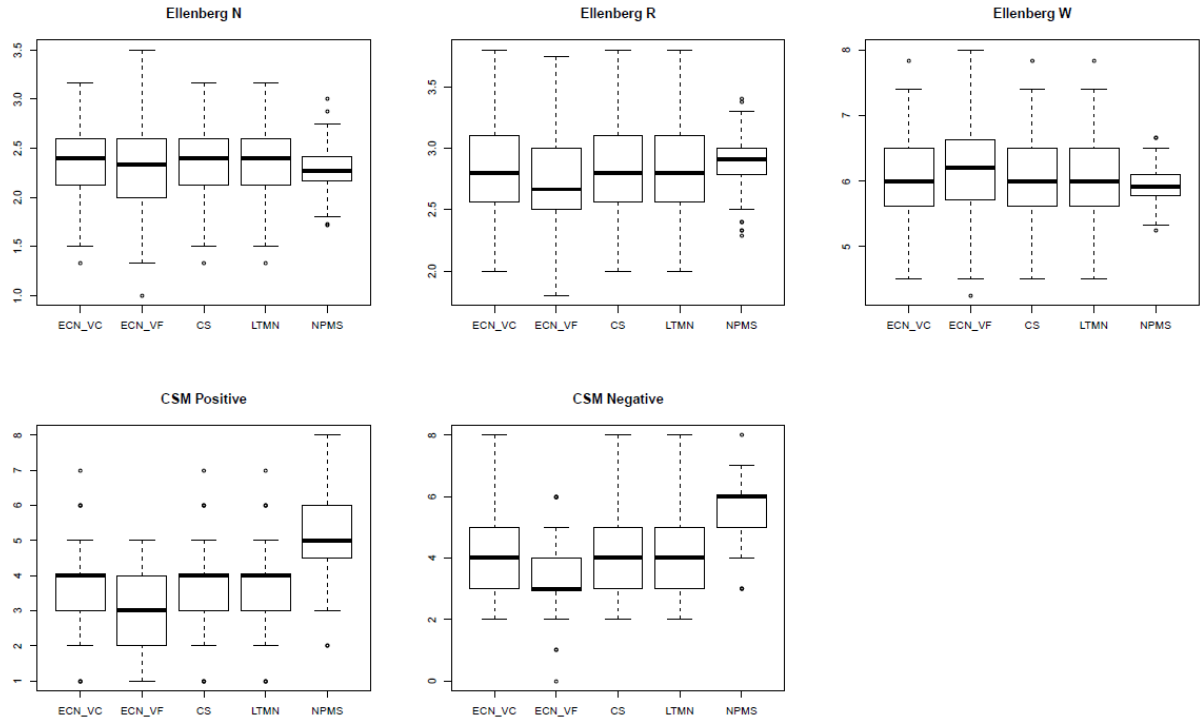


Figure 5.4: Boxplots of different indicator metrics derived from 1000 simulated patches of vegetation sampled in a consistent manner except for the difference in plot sizes across the different schemes under consideration.

We then extended this comparison to include the differences in plot size and the ways in which the schemes estimate cover. Once again, all other differences were held constant. In this case, because both the plot size and the mechanism by which cover is estimated is the same across the ECN coarse scale and LTMN plots, these were essentially using the same data. Boxplots of indicators are shown in Figure 5.5.

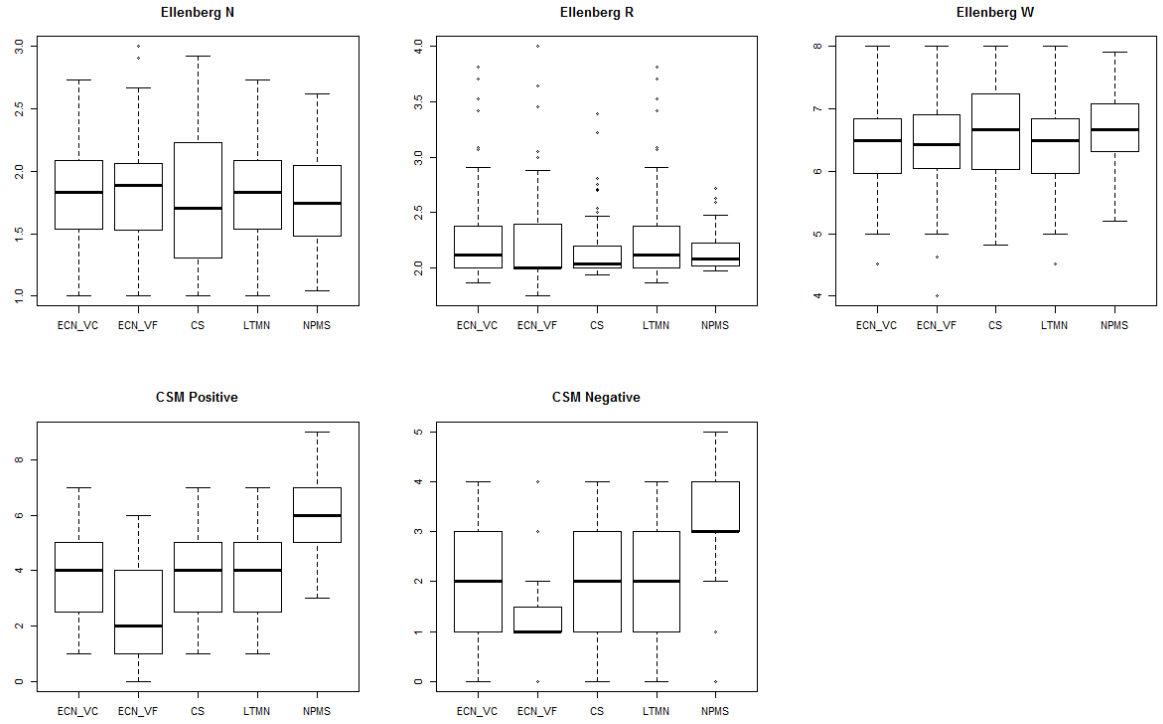


Figure 5.5: Boxplots of different indicator metrics derived from 1000 simulated patches of vegetation sampled in a consistent manner except for the difference in plot sizes and cover estimation across the different schemes under consideration.

Once again, the mean values in the Ellenberg scores appear to be reasonably consistent across plot sizes and cover estimation methods of each of the schemes. There were some differences in variability, particularly for Ellenberg R where the CS and NPMS showed less variability than the other schemes. These schemes both adopt an overall surveyor assessment of cover rather than using the cell count method of the other schemes. The effect of plot size on the CSM indicators seen in Figure 5.4 is still apparent and, as these are not cover weighted, there is no pattern beyond that already identified. The Anderson-Darling test confirms that the Ellenberg indicators can be assumed to be independent of scheme, with the exception of Ellenberg R, where the difference in variability provided some evidence for a significant difference ($p=0.011$).

Finally, the effect of plot relocation and surveyor accuracy were added in addition to the effect of plot size and cover estimation. Both of these aspects affect the variability of the derived data. The resulting box plot is shown in Figure 5.6, where no two schemes are based on the same data due to the random variability now introduced. Once again, it appears that mean Ellenberg scores can be considered roughly equivalent ($p>0.1$ in all cases), perhaps with some small differences in overall variation. However, it is the CSM indicators that again show the biggest difference, with the effect of plot size still dominating.

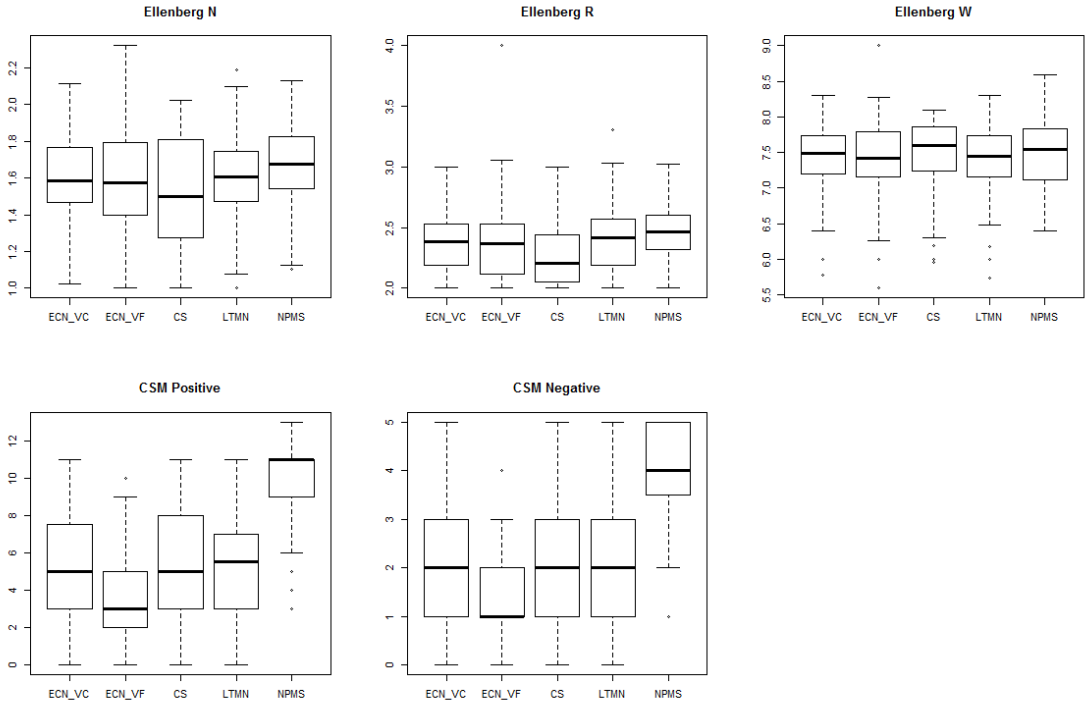


Figure 5.6: Boxplots of indicator metrics derived from 1000 simulated patches of vegetation sampled in a consistent manner except for the difference in plot sizes, cover estimation, relocation error and surveyor accuracy between the schemes under consideration.

Whilst the results shown in Figures 5.4, 5.5 and 5.6 provide consistent and understandable effects of the different protocols on indicator responses, these are all based on simulating vegetation patches of the same type. We therefore also investigated the impact of simulating different patches of vegetation typical of communities across a gradient in order to assess potential differences in the way the schemes would quantify change. Six target populations were considered that traversed the major axis of the species ordination previously compiled. This provided a gradient of community type with which to re-run the simulation and test the robustness of any conclusions drawn regarding key between-scheme differences. In Figure 5.7, the six target communities are superimposed onto the species level ordination previously ran in Section 5.3,1.

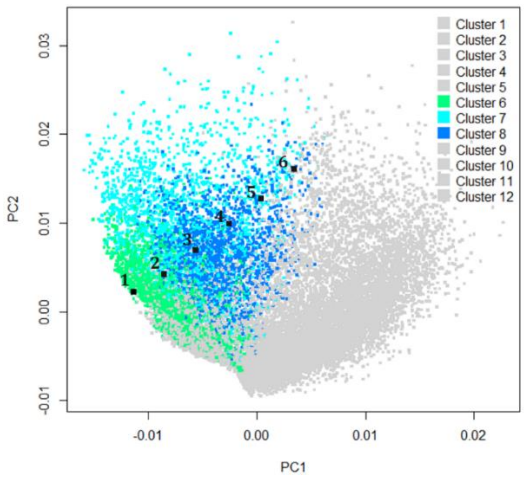


Figure 5.7: Species level ordination across all scheme data with six target population points labelled, providing a gradient of community types.

The resulting boxplots from this series of simulations are shown in Figure 5.8 with each row representing one of the different target communities 1:6 respectively from top to bottom. Whilst there are some differences across these plots, the main features identified in Figure 5.4-5.6 are still apparent. Slightly different patterns are evident for some indicators, depending on the community. One example is the Ellenberg R indicator, for which the CS scheme appears differs slightly from the others on the bottom row (the extreme of the community types), in comparison with the top rows, where there is consistency across all schemes.

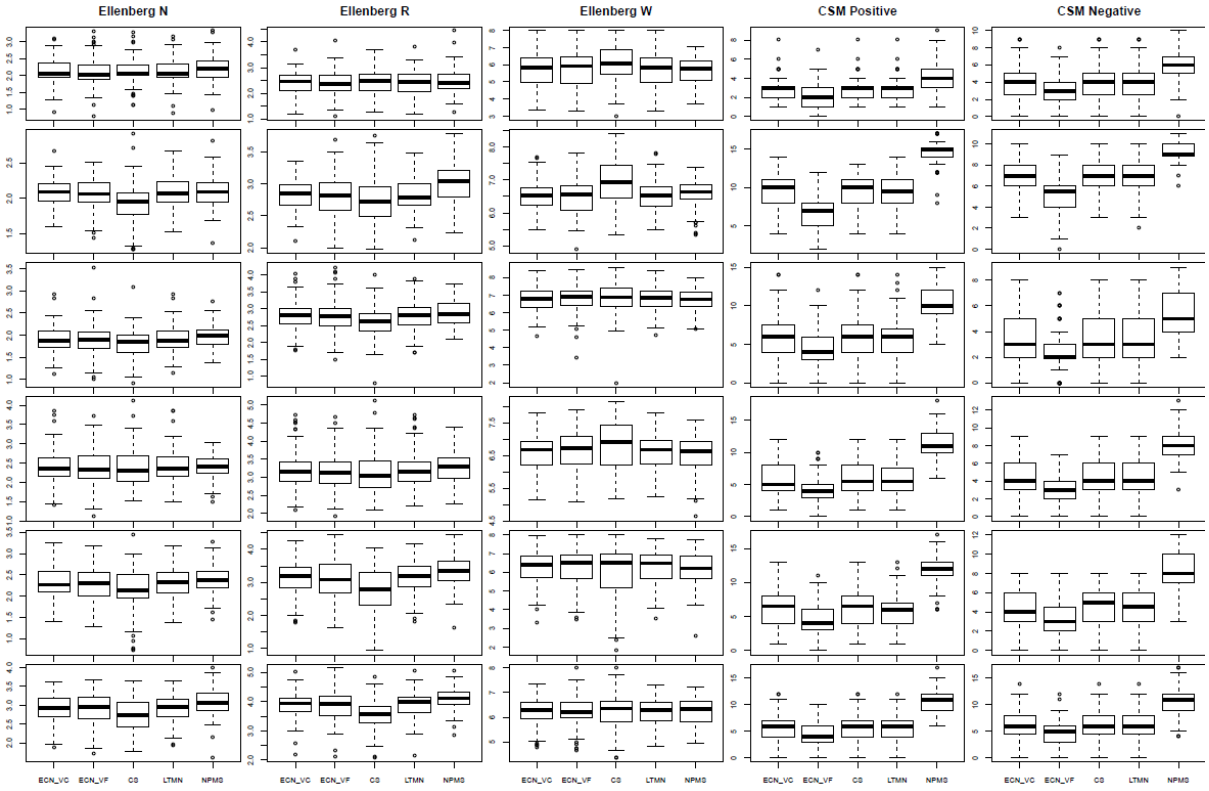


Figure 5.8: Boxplots of indicator metrics derived from 1000 simulated patches of vegetation sampled in a consistent manner except for the difference in plot sizes, cover estimation, relocation error and surveyor accuracy across the different schemes under consideration. Each row represents a different target community identified according to the ordination in Figure 5.7. Top to bottom rows represents target regions 1:6 respectively.

5.6 Summary

On the basis of the results from the simulations conducted, it seems reasonable to assume that differences in plot size between schemes do not affect mean Ellenberg scores, whether they are cover weighted or not. The variability in the Ellenberg indicators may differ slightly across schemes, especially when surveyor error or relocation error are considered, but this can be captured within a scheme specific random effect in any joint model, which should be a primary consideration for future work.

The CSM scores do vary significantly across schemes, with plot size having the biggest effect. The effect overwhelms any differences in variance or other differences. This will have to be accounted for explicitly in any joint model.

Out of all the indicators, Ellenberg N seems the most robust to differences across the schemes and this is consistent across gradients. It is therefore likely that this indicator will provide the best basis for integrated modelling and for producing robust results consistent across all schemes.

6 The integrated vegetation model: temporal change

6.1 Introduction to modelling temporal change

Having completed the assessment of the individual schemes (Chapter 4) and concluded there were no major impediments to combining data from the four schemes within a single analysis (Chapter 5), it was then possible to consider how best to integrate the datasets into a single model.

At this stage if was necessary to determine:

- Which elements of protocol differences between schemes e.g. quadrat size needed to be accounted for. This was informed by the simulation study results in Section 5
- How to construct the models to account for these differences
- Which elements of the models could be shared between datasets. This would include the effects of time and other covariates

In some cases the process was straightforward e.g. for Ellenberg N. The simulation study results suggested it was not necessary to account for quadrat size in modelling Ellenberg N, and there was good evidence from Chapter 4 that trends in Ellenberg N are comparable between schemes.

It was clear that there was a need to allow for separate intercepts as, for example, the LTMN plots tend to record lower Ellenberg N on average. It would also be sensible to allow different slopes via a random slope model. For Ellenberg N it may not be necessary to fit a random slope model as all trends are so similar, but a random slope model will be more transferable to other indicators where trends may be less comparable.

Some indicators appear more difficult to model, such as the richness of CSM positive indicators. The simulation work presented in Section 5 showed that it is important to account for differences in quadrat size while modelling CSM responses. In addition, results from Chapter 4 indicated that trends in CSM positive richness differed between monitoring schemes, varying between strongly negative (CS) to strongly positive (NPMS). This suggests that fitting integrated models to these data may not be sensible without a greater understanding of the differences between schemes. Accounting for protocol differences such as quadrat size may account for differences in intercepts e.g. higher richness on average in NPMS plots, but do not explain differences in trends.

To develop an appropriate model structure, we initially fitted an integrated temporal model only i.e. excluding any covariates related to drivers of change. We fitted a basic model first, which could then be added to, for example, by adding covariates. To do so, we expanded on the individual trend models presented in Section 4 to include an additional random effect level of scheme (Eq 2).

$$\text{Indicator} \sim \text{Year} + (1|\text{scheme/site/PlotID}) \quad \text{(Eq 2)}$$

This is the simplest way to include all schemes in a single model but enforces a shared trend over time. To allow trends over time to vary between schemes, whilst also estimating a shared trend, we can extend Eq 2 to fit a random slope model (Eq 3).

$$\text{Indicator} \sim \text{Year} + (1|\text{scheme/site/plotID}) + (\text{Year}|\text{scheme}) \quad \text{(Eq 3)}$$

In this model we allowed the effect of year to vary between schemes i.e. we fitted random slopes per scheme. Importantly, we did not allow sites within schemes, or plots within sites to have different slopes. Each plot within a scheme was assumed to have the same trend over time, but allowed to take a different baseline value (intercept).

6.2 Integrated temporal model results

To test the benefits of integrating datasets within a single model, we first assessed the potential of extending the models presented in Chapter 4 to model all schemes jointly. These models included effects of time but not covariates of potential drivers of change (Eq 3). We also included an offset for quadrat size in CSM models as indicated by the simulation work in Chapter 5.

For all three Ellenberg indicators we found that the integrated models produced trend estimates that had high precision (small standard errors; Table 6.1, Figure 6.1). However, we found that integrated models for the CSM indicators invariably had low precision.

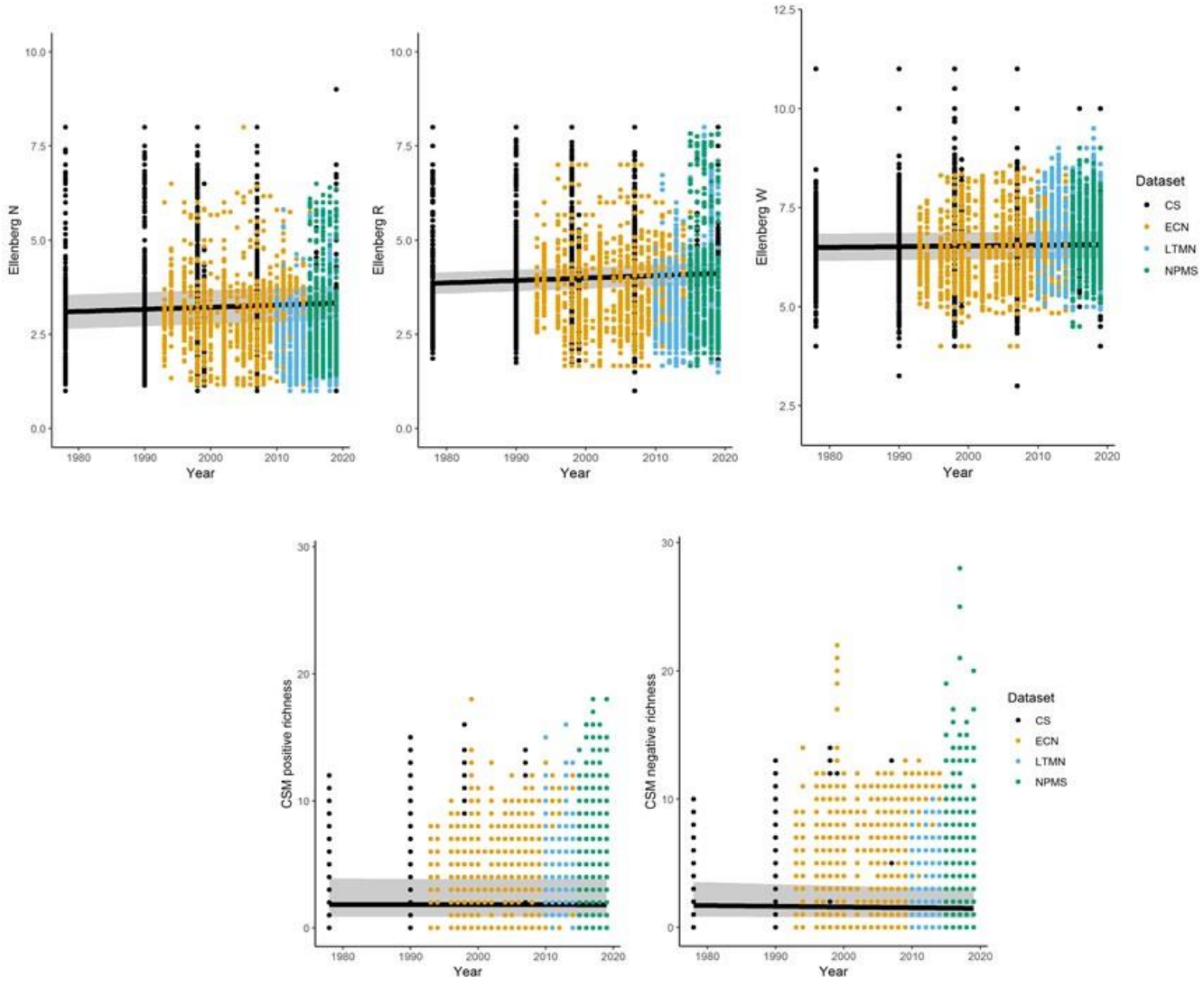


Figure 6.1 Estimated trends over time in five selected indicators produced via data integration of four vegetation monitoring schemes. Confidence intervals are shown around the estimated trend.

Table 6.1 Estimated trend slopes, confidence intervals, and p values in the five selected indicators produced via data integration of the four vegetation monitoring schemes

Indicator	Integrated model
Ellenberg N	0.0057 (0.0007) $P < 0.001$
Ellenberg R	0.00654 (0.0007) $P < 0.001$
Ellenberg W	0.0017 (0.0006) $P < 0.001$
CSM positive indicators	-0.0009 (0.003) $P = 0.762$
CSM negative indicators	-0.003 (0.002) $P = 0.125$

Overall, we can conclude that the integrated approach works well for the Ellenberg indicators, as a consequence of the similarities in trends across the individual datasets. We were able to obtain consistent and precise estimates of change over time in all three Ellenberg indicators by integrating CS, ECN, LTMN and NPMS data.

There is less evidence that integrating data from multiple schemes provides any benefits for characterisation of trends in heath and bog CSM indicators. The integrated models show high uncertainty in the direction of trends in the CSM indicators as a result of conflicting trends in the individual datasets. The reasons for this have not been fully explored but may reflect greater variability in counts of CSM indicators, potentially as a consequence of apparency, observer skill, phenology etc., compared to community averaged Ellenberg scores. Community average Ellenberg scores are thought to be fairly robust to differences in surveyor effort, skill and other observation-related processes e.g. weather, whereas counts of indicators may be more sensitive to these observation processes. It is also possible that CSM scores are more sensitive to local management effects. This could result in greater spatial heterogeneity and less linearity in long-term trends than might be expected for the Ellenberg indicators. The latter could be particularly important in influencing differences between the four schemes that cover markedly different time periods.

7 The integrated vegetation model: trend attribution

7.1 Introduction to the trend attribution modelling

The gradual increase in community Ellenberg R and Ellenberg N scores demonstrated from the analysis of the integrated datasets presented in Chapter 6, suggest widespread shifts in the assemblages of UK heaths and bogs that are broadly indicative of both ecosystem recovery from acidification and progressive atmospheric eutrophication from reactive nitrogen (N). While changes in the former are consistent with other recent evidence of the positive impacts of reductions in the emissions of transboundary air pollutants, and as such can be considered a positive environmental development, changes in the latter give cause for concern and are of significant relevance to national and international air quality and biodiversity policy. In order to confidently attribute these drivers to our observations of change, however, it is first necessary to determine the extent to which the changes in these indicators can be linked directly to changes in the deposition of sulphur (the predominant contributor to acid deposition historically) and N (the main atmospheric source of eutrophication in semi-natural systems).

Reductions in S deposition have been substantial across the UK in recent decades. Sulphur tends to behave relatively conservatively within catchments, i.e. inputs and outputs tend to be closely associated over relatively short time scales – at least in better drained soil types. The reductions in S deposition are therefore relatively easy to link to recent reductions in soil and water acidity. In contrast, N deposition, although in decline, has not fallen as fast as S deposition, and there remains considerable uncertainty around the extent to which N may or may not be continuing to accumulate within soils. This is partly due to the more complicated biogeochemical cycling of deposited reactive N and the challenges in accurately quantifying denitrication rates (i.e. loss of N back to the atmosphere).

Consequently, there is also considerable uncertainty around how bog and heath vegetation is expected to respond to elevated, although declining, rates of N deposition. Perhaps the most obvious explanation for the overall rise in Ellenberg N in heath and bog vegetation reported in Chapters 4 and 6 is that deposited N continues to accumulate and thus enrich or “eutrophy” these environments. Alternatively, it is feasible that recent shifts to assemblages with a higher nutrient preference might result from the increased presence of acid-sensitive species (as a consequence of reduced soil acidity), and/or species with different climatic preferences, that also tend to have a higher nutrient demand (Rose et al., 2016).

In this section, therefore, we describe the introduction of covariates into the integrated modelling framework, in order to test a range of hypotheses designed to improve our understanding of the relative influences of recovery from acidification, continued eutrophication from nitrogen deposition and changes in climate, on terrestrial vegetation indicators at ECN, CS, NPMS and LTMN sites. Our main focus of interest was the response of community mean Ellenberg R (soil acidity) and N (soil fertility) indicators, calculated for each plot in each scheme in each year of survey, although we also investigated potential drivers of change in Ellenberg W (soil moisture). The covariates applied, i.e. the explanatory variables, are based on metrics of sulphur and nitrogen deposition, and climate. Because we wished to focus on explaining differences in indicators between survey years and over time rather than explaining spatial variation, we standardised our response and covariate data relative to site means, as explained in the following sections.

7.2 Covariates

7.2.1 Climate covariates

Although our principal focus was on explaining change in indicators of air pollution impacts it was also important to consider the potential influence of climate on vegetation change. In part this is because warmer temperatures and a wetter growing season can promote growth of more competitive species, particularly perennial grasses. Hence weather can drive a similar response to that resulting from changes in macronutrient availability (see for example Dunnet et al 1998 and Silvertown et al 1994).

Climate attribution studies relevant to both heath and bog with a primary focus on temporal, as opposed to spatial, variation were used to guide selection of climate covariates. Such studies are scarce and their relevant features are summarised below:

- Mauquoy & Yeloff (2008) - Response of raised bog to climate change – summer temperature and summer rainfall were identified as key drivers of a shift from *Sphagnum* to vascular plants.
- Britton et al. (2017) – Changes in dwarf shrub heath and moorland were studied over a 35 year interval in Scotland. Significant predictors included rainfall (annual, spring, winter, autumn, summer) and maximum winter and summer temperature. As is typical, a broad selection of climate variables were applied so as to maximise signal attribution but without necessarily framing a hypothesis around any specific covariate.
- Kirk et al. (2010) – This study focused on soil pH change (Eng & Wales) between 1978 and 2003. Covariates included mean annual precipitation from 1978 to 2000 as the sole climate predictor, alongside sulphur (S) deposition and other soil variables.
- Stevens et al. (2016) - Vegetation change between 1965 and 2012-13 was analysed using mean Ellenberg scores and other indices to summarise vegetation quadrats in acid and calcareous upland grasslands. The temperature variables selected were maximum July and minimum January temperature and annual rainfall.

In light of our review of the relevant attribution literature, the final covariates were selected as follows.

- Annual precipitation (mm yr^{-1})
- Mean Maximum July temp (˚C)
- Mean Minimum Jan temp (˚ C)

We recognise that this selection is rather limited, but we were restricted by the time available to carry out this element of the work. There would clearly be merit in exploring additional variables including a separation of rainfall into winter and summer, a measure of growing season length and estimated evaporative loss and soil moisture.

7.2.2 Atmospheric S and N deposition covariates

Our choice of covariates to represent current and past deposition loads reflected the need to capture the historical peak, and then consistent decline, in sulphur deposition (Figure 7.1) and the cumulative loading of total N to terrestrial ecosystems in Britain (Payne et al 2013; RoTap 2012). We therefore used the following sources of pollutant estimates to derive our explanatory variables:

- Total N deposition averaged for the period 2003-2005. This was derived from CBED 5 x 5 km estimates as recommended and used in RoTAP (2012). The rationale here was that we expected an approximately linear spatial relationship between long term catchment

accumulation of deposited reactive N (which is not possible to estimate directly) and the N deposition load prior to recent reductions.

- Annual total S deposition at 5 x 5 km, based on a combination of FRAME + CBED deposition estimates, that included the 1970s peak in deposition and the subsequent widespread and marked reduction (e.g. Figure 7.1).

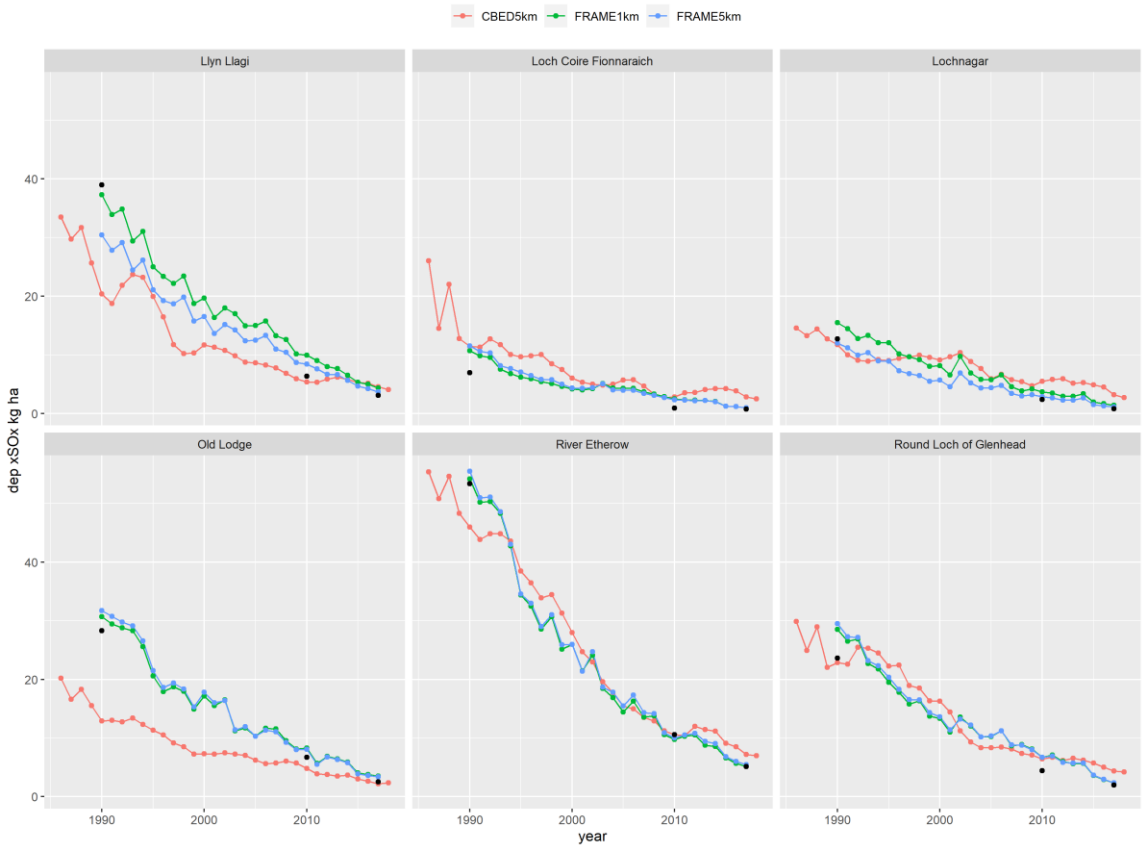


Figure 7.1: Sulphur deposition trajectories estimated by both FRAME and CBED models for UK Upland Acid Waters Monitoring sites (to represent a broad UK distribution) covering the majority of the period of observations by the terrestrial schemes and illustrating the overall decline in deposition and consistency between datasets and resolutions. The sites include Llyn Llagi (North Wales), Loch Coire Fionnaraich (Northwest Scotland), Lochnagar (Northeast Scotland), Old Lodge (Southeast England), River Etherow (southern Pennines) and the Round Loch of Glenhead (Southwest Scotland).

Covariates were transformed in a number of ways in order to optimise hypothesis testing. Specifically, we sought to remove spatial gradients in the response and covariate data so that analysis focussed on explaining differences in mean Ellenberg values between years (temporal differences) rather than between locations (spatial differences). In this respect our study is a departure from the many spatial gradient studies that have been used to infer cumulative impacts of deposition on the basis of spatial variation in ecological response.

For each plot in each scheme in each year, we calculated the difference between S deposition for that year and mean S long-term (1970 to 2018) deposition for the site. Hence, larger positive values for a plot indicated higher S deposition for that year relative to the mean for that location. Panel 1 in Figure 7.2, illustrates how this transformation removes spatial variation in mean S deposition between sampled locations (evident in panels 2 and 3) while maintaining information on the longer deposition trend at each location.

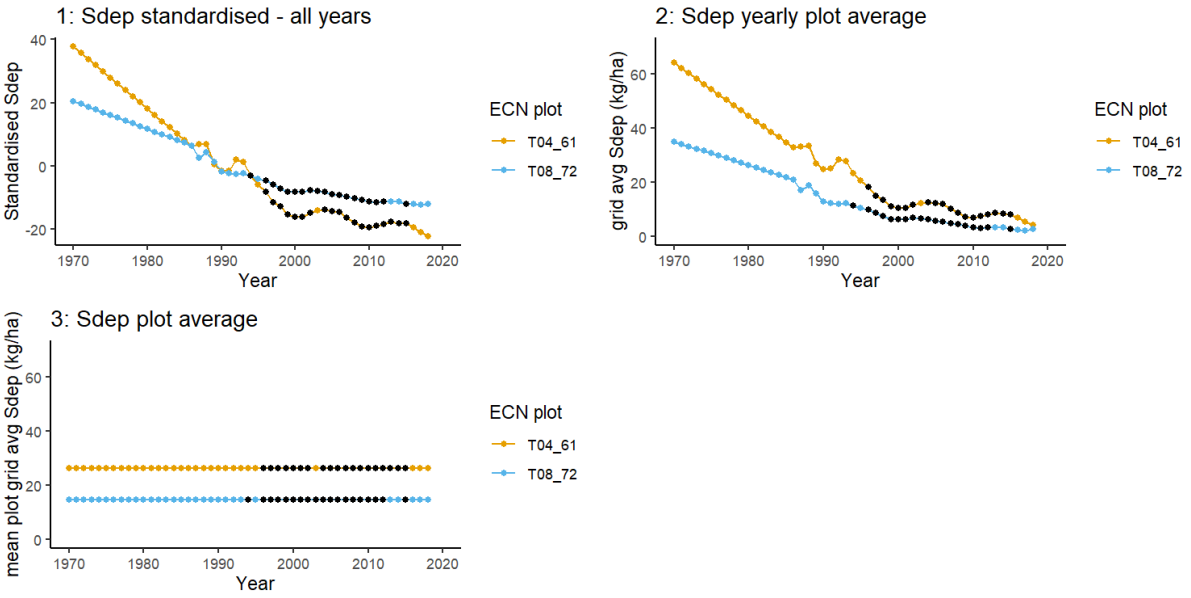


Figure 7.2: Options tested for construction of the sulphur deposition covariate. Panel 1represents the selected best option. Examples are provided for two randomly selected vegetation plots from ECN Moor House (TO4-61) and ECN Glensaugh (TO8_72).

7.2.3 Centring the response variables

The mean Ellenberg R and N values in each sampled location in each scheme in each year were also centred on the mean Ellenberg R and N values across years, again to remove spatial differences. This in turn meant that the mean Ellenberg score for a plot across all years could be used as an independent explanatory variable, for example to test whether temporal differences in Ellenberg scores between plots across the sampling period in response to deposition were additionally dependent on the average acidity status of the vegetation (Figure 7.3).

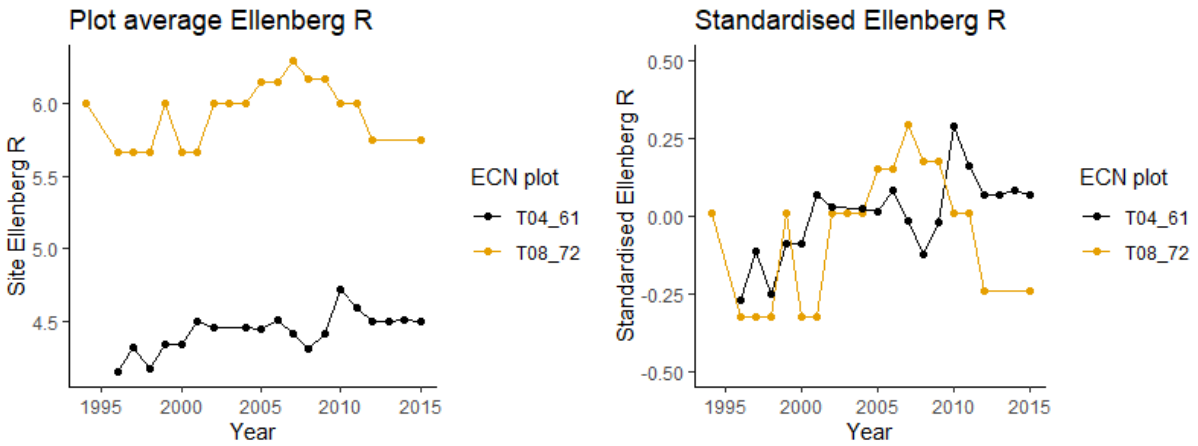


Figure 7.3: Options tested for construction of the sulphur deposition covariate. Panel 2 represents the selected best option. Examples are provided for two randomly selected vegetation plots from ECN Moor House (TO4-61) and ECN Glensaugh (TO8_72).

7.3 Hypothesis testing

We then developed a set of statistical models to test a range of hypotheses, focussing on those raised in the introductory paragraphs of this chapter. These are as follows and their outcomes are considered in detail in the results section:

Hypothesis 1: N deposition has driven a long term increase in Ellenberg N

Hypothesis 2: Reductions in acid (S) deposition have driven a long term increase, hence recovery, in vegetation as indicated by mean Ellenberg R.

Hypothesis 3: The eutrophying impact of cumulative N deposition is most evident where recovery from acidification has been greatest, i.e. the increase in Ellenberg N is dependent not only high N deposition but has increased to a greater extent where high N deposition and declining S deposition coincide.

Hypothesis 4: A climate favourable to nitrogen-loving plants is associated with a greater increase in Ellenberg N.

Hypothesis 5: An increase in precipitation has increased the occurrence of plants of wetter conditions or/and decreased plants of drier conditions.

Hypothesis 6: An increase in precipitation (and consequent effect on soil moisture) also explains increases in Ellenberg R in addition to a separate effect of decreased S deposition.

Hypothesis 7: The increase in Ellenberg N is correlated with an increase in precipitation in addition to a separate effect of N deposition.

In the following sections we examine the hypotheses listed above sequentially. We refer to the Ellenberg R metric as EbR, and the Elleberg N metric as EbN. In each case tabulated model results are followed by an interpretation. Models are numbered according to the hypothesis numbers, and in some cases model variants are identified with a suffix, e.g. Model 2.2.

7.3.1 Hypothesis 1: N deposition has driven long term increase in Ellenberg N

MODEL 1 STRUCTURE: Standardised EbN (response variable) at each quadrat in each year = intercept + (plot-specific 2003-2005 average N deposition) + (year of survey) + interaction between the two.

Table 7.1: Model 1. Results table for the random effects model. N dep = plot-specific 2003-2005 average N deposition; Year = year of survey.

	Value	Std.Error	DF	t-value	p-value
Year	0.002028	0.0009442	4934	2.148211	0.0317
N dep	-0.316579	0.1146111	4934	-2.762202	0.0058
Year * N dep	0.000158	0.0000572	4934	2.760216	0.0058

Model 1 interpretation.

Although the coefficient for the separate N deposition variable is negative, the positive coefficient for the interaction term in Table 7.1 indicates that Ellenberg N has been increasing most at sites where N deposition was highest in 2003-05 – and hence those sites that have received the highest reactive N loads historically.

What is the size of the modelled N deposition effect?

An illustration of the ecological, rather than purely statistical, significance of the impact of N deposition, as inferred by Model 1, can be achieved by adding the change in mean Ellenberg N over 49 years (1978 to 2019) predicted using Model 1, onto the starting values of each quadrat in 1978 and plotting the new distribution of predicted values (Figure 7.4). The significant interaction term in Model 1 tells us that the rate of change over time depends on the history of N deposition at a site (as indicated by the 2003-05 deposition estimate). We therefore added a predicted increase over time at the 5 percentile value of total N deposition in the dataset (5.2 kg N ha^{-1} yr^{-1}) and at the 95%tile (27 kg N ha^{-1} yr^{-1}). The density of mean Ellenberg N scores for Acid Grassland in CS2007 is also shown, representing an 'undesirable' and more fertile reference distribution. Model 1 estimates an average increase of 0.23 of an Ellenberg N unit over the 49 years at the 95%tile deposition. The full range of Ellenberg N scores is 1 to 9. Hence at high N deposition, 49 years of loading is estimated to result in a 2.5% change along this vegetation fertility gradient from the most infertile to the most fertile vegetation in Britain.

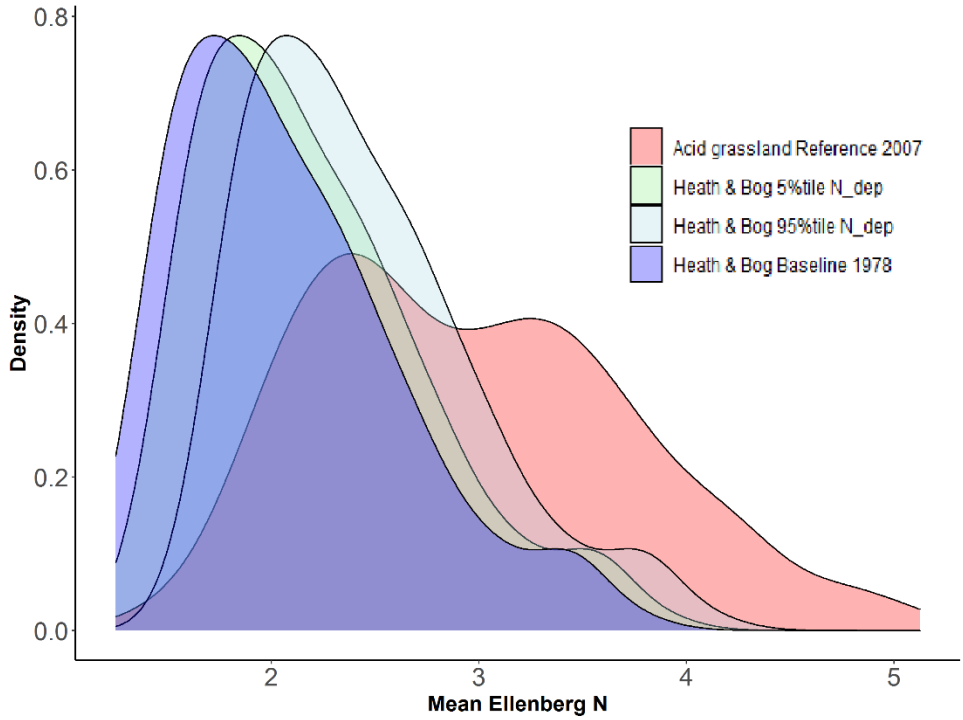


Figure 7.4: Application of Model 1, to predict how the distribution of mean Ellenberg N values in 2019 are expected to have increased from their 1978 values. Predictions were added to the baseline values assuming a lower (5%tile) or upper (95%tile) N deposition value. A reference distribution of mean Ellenberg N values for plots in Acid grassland is shown to aid assessment of the magnitude of the estimated change.

It is important to note that nitrogen deposition has been falling in recent years, although less sharply than sulphur. So the relationship described by Model 1 is consistent with either a lagged response to N deposition and/or the effect of a continued gradual increase in the availability of plant available N in these plots, possibly as a consequence of continued accumulation of deposited N or other changes in soil chemistry.

7.3.2 Hypothesis 2: Reductions in acid (S) deposition have driven a long term increase, hence recovery, in vegetation as indicated by mean Ellenberg R.

A series of models with varying numbers of covariates were fitted to test this hypothesis, in which variation in the Ellenberg R score of each plot was modelled as a function of S deposition and time. In the more complex models we also included mean plot Ellenberg R score to reflect the long-term acidity status of the plot. The Akaike information criterion (AIC) values for each model are provided in Table

7.2 as a guide to the best-fitting model when taking the number of covariates into account. Models with the lowest AIC are normally considered to have the best fit, although AIC values do not communicate the level of statistical significance of a model and do not necessarily convey a high ability to explain the variation in the response data..

MODEL 2 GENERAL STRUCTURE: Standardised EbR at each quadrat in each year = intercept + (time standardised S deposition) + (year of survey) + (interaction between the two)

Table 7.2: Models tested and their AIC values. S dep = standardised S deposition, Year = year of survey.

Effects in model	**AIC**
Random effects only (Model 2.1)	6938
S dep (Model 2.2)	6894
Year (Model 2.3)	6863
S dep + mean Ellenberg R across years (Model 2.4)	6870
S dep + mean Ellenberg R across years + interaction term (Model 2.5)	**6849**
Year + mean Ellenberg R across years + interaction term (Model 2.6)	**6836**

Table 7.3: Model 2.2. Results table for the random effects model. Response variable = Standardised EbR at each quadrat in each year. S dep = standardised S deposition.

	Value	Std.Error	DF	t-value	p-value
S deposition	-0.004	0.0005	4888	-7.1	0.0000

The results for the simple model, involving S deposition as the sole predictor (Model 2.2), indicate a statistically significant negative relationship, such that plots with higher standardised mean Ellenberg R values are associated with lower S deposition. The standardised effect size (-0.09) indicates a very small effect relative to the wider variation in Ellenberg R in the dataset. This was expected given the multiple factors and relationships likely to be responsible for this variation in addition to S deposition (e.g. Van den Berg et al 2010; Van der Wal et al 2003). Moreover the 5 x 5 km grid cell estimates may not fully reflect changes that may have occurred at a sub-grid scale. A plot of the modelled slope against the observations indeed shows how much residual variation there is in the response along the S deposition gradient (Figure 7.5). Importantly however, the variable Year (as sole predictor) provided a better fit than S deposition, leaving open the possibility that another monotonically changing driver provides the dominant mechanism (see comment within the following interpretation section).

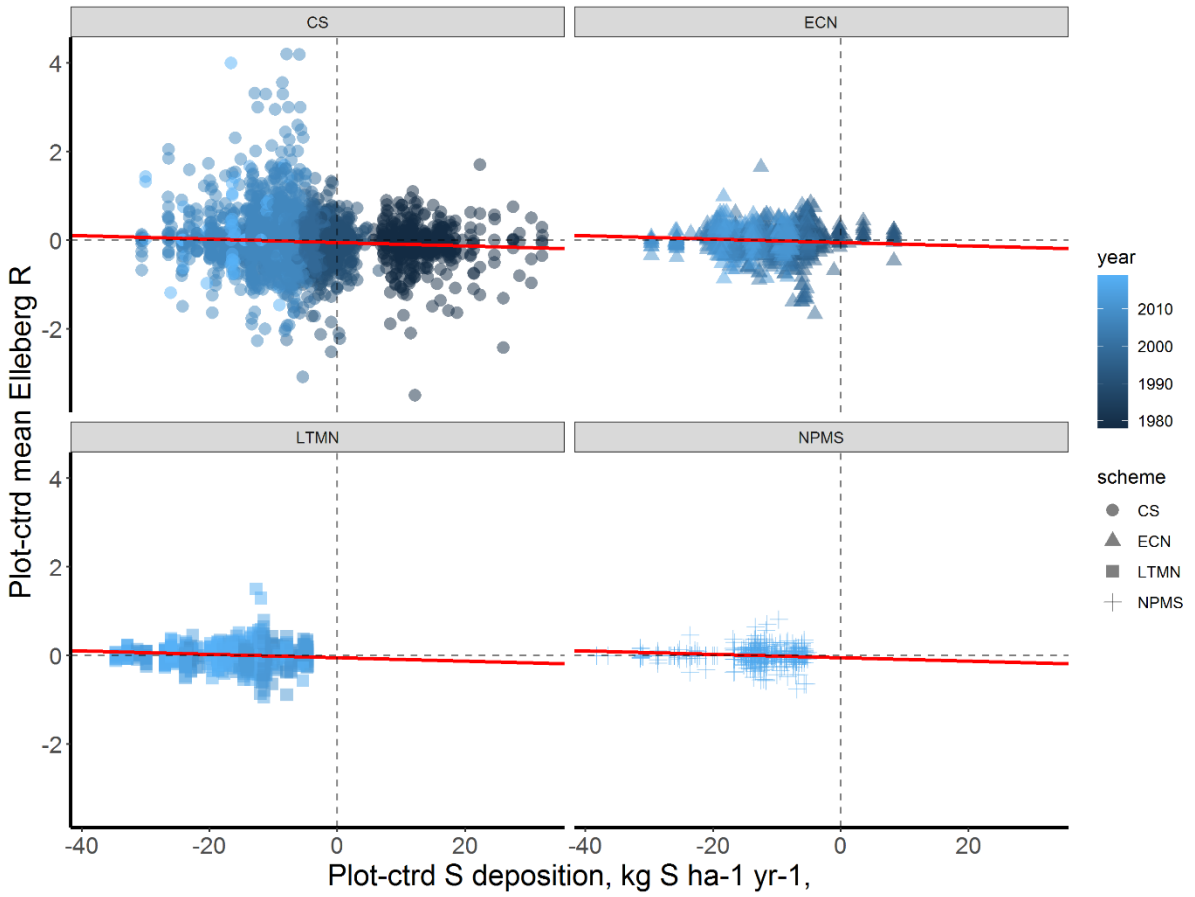


Figure 7.5. Raw data points and fitted line from Model 2.2 (S deposition as the sole predictor of temporal differences in mean Ellenberg R). The fitted regression line is the same in each panel and reflects the fitting of the model to all scheme datasets combined. Contributing plots are shown in separate panels to indicate scheme-specific coverage of the change in S deposition gradient and years sampled.

We also tested the hypothesis that the response of mean Ellenberg R over time is not only driven by S deposition but also conditioned by the long term acidity status of the impacted plot, as inferred from the long-term mean Ellenberg R value for each plot (i.e. averaged across years). This is an important test, since evidence suggests that more organic, lower pH soils differ from mineral soils in their response to acid deposition. The interaction term in Model 2.5 (Table 7.4) was highly significant. On the starting assumption that S deposition is indeed the dominant driver, the coefficients in Model 2.5 indicate that vegetation plots indicative of less acidic conditions have been more responsive to the reduction in S deposition than those indicative of more acidic environments.

Table 7.4: Model 2.5 Response variable = Standardised EbR at each quadrat in each year. S dep = standardised S deposition. Mean EbR = mean long-term Ellenberg R score for the plot.

	Value	Std.Error	DF	t-value	p-value
S deposition	0.0026	0.00146	4887	1.7775	0.0755
Mean EbR	0.0026	0.00460	3090	0.5665	0.5711
Interaction	-0.0020	0.00041	4887	-4.7994	0.0000

The importance of the interaction in Model 2.5 is illustrated in Figure 7.6. Vegetation plots with a higher long term mean Ellenberg R score, i.e. those associated with less acid soils overall (pink line),

showed stronger responses to changing S deposition than those associated with more acidic conditions (darker green line in Figure 7.6).

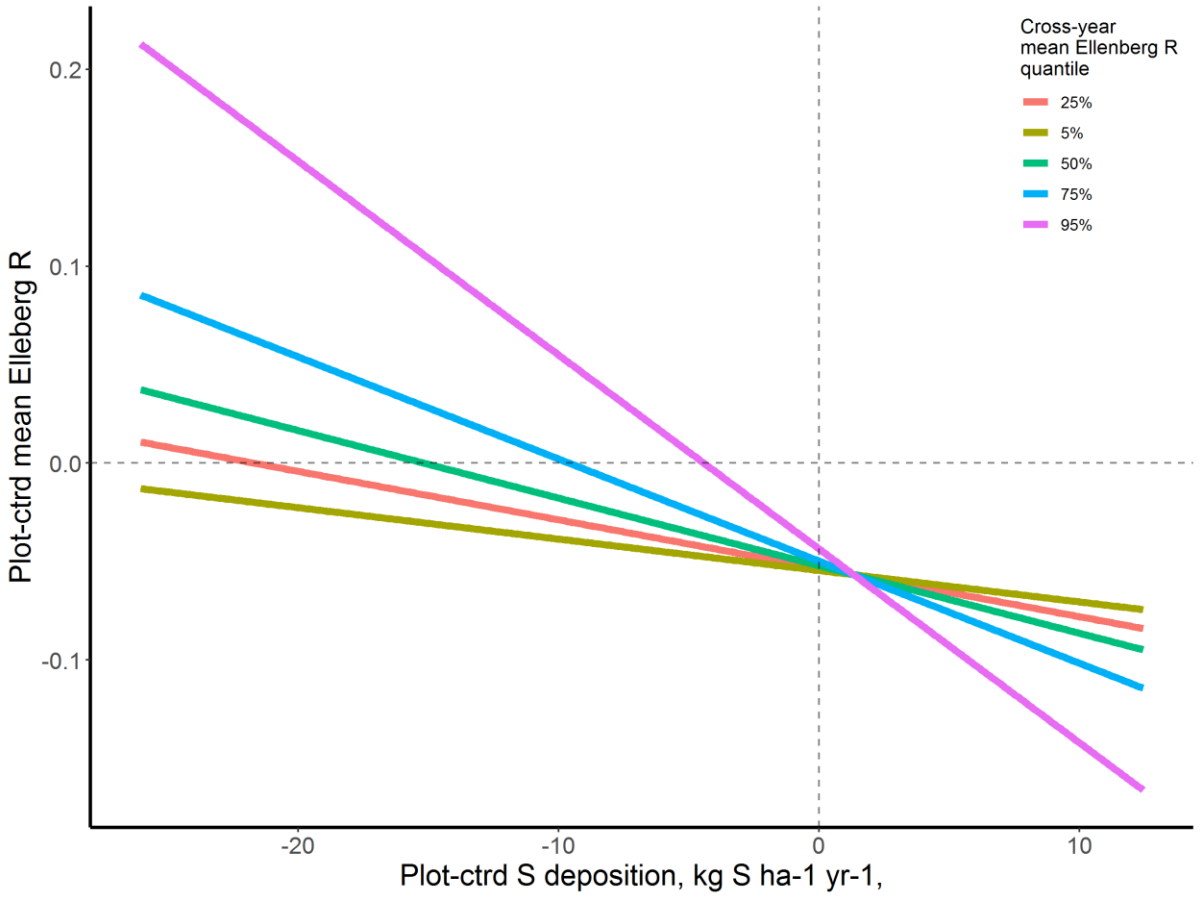


Figure 7.6: Each line shows predictions from Model 2.5 at varying percentile values of the cross-year mean Ellenberg R value to illustrate the effect of the significant interaction term (see Table 4). The interaction terms estimates how the response of plots to the S deposition gradient varies with mean Ellenberg R where the mean is calculated across all sampling years for that location and is therefore an indicator of the average pH regime of that location.

Table 7.5: Model 2.6. Results table for the random effects model with lowest AIC. Response variable = Standardised EbR at each quadrat in each year. Covariates are Year of survey and mean EbR = mean long-term Ellenberg R score for the plot.

	Value	Std.Error	DF	t-value	p-value
(Intercept)	**-3.608**	**2.33669**	**4887**	**-1.5441**	**0.1226**
year	**0.0018**	**0.00117**	**4887**	**1.5068**	**0.1319**
EbR_mean	**-1.5945**	**0.60491**	**3090**	**-2.6358**	**0.0084**
year:EbR_mean	**0.0008**	**0.00030**	**4887**	**2.6630**	**0.0078**

Interpretation

When S deposition was applied as the sole covariate, we found a strong negative relationship between time-centred S deposition and time-centred EbR, i.e. consistent with an S deposition control on

Ellenberg R (Fig 5). It is important to note, however, that the variable Year (as sole predictor) provided a better fit than S deposition, and when Year and S deposition were both included in a two predictor model, S deposition was not statistically significant. Hence, the pattern of widespread increases in Ellenberg R is better explained by a pure linear variable than the more complex, but still monotonic, change in S deposition.

This, therefore, leaves open the possibility that a single factor, other than S deposition, dominates the Ellenberg R signal (but see interpretation for Model 6.2 below). A more plausible explanation is that the centred S deposition variable is overly sensitive to the estimated S deposition referenced to each plot-centred mean Ellenberg R value in each specific year. Lag effects in the vegetation response would lead us not to expect an instantaneous coupling between S dep and mean Ellenberg R. Moreover, it is possible that the plot and year-specific deposition estimate is also overly sensitive to yearly increases and decreases in estimated deposition (see Fig 7.2). Further work should explore a smoothed non-linear covariate that perhaps better conveys the longer term reduction in S dep and is less sensitive to annual variability. The last model tested was one with year, mean Ellenberg R and their interaction term as covariates (Table 7.5). This actually had the lowest AIC (Table 7.2) which suggest that the S dep variable indeed has shortcomings as an informative measure of long-term influential changes in S deposition compared to time alone. Compared to the spatially invariant time term we would expect spatial variation from plot to plot in the plot-centred S dep variable to contribute useful additional explanatory power but this is clearly not the case. We would also expect a degree of mismatch between the covariate and the plant community response because of their different resolutions; the former is a 5x5km estimate while the latter is a 2x2m measurement. However, it is unexpected that an index on year of survey should prove more explanatory. Further work is clearly needed to understand these results.

Bearing in mind the reservations above, while working on the assumption that S deposition is actually the dominant driver, the best fitting model (Model 2.5), included a significant interaction term between the long-term mean Ellenberg R score of the plots and S deposition. The model parameters (Table 7.4) indicate that the larger acidity-driven changes in vegetation (i.e. larger increases in Ellenberg R), in response to reductions in S deposition, have occurred in less acidic environments.

The standardized effect size for the interaction term is very small (-0.05). However the main effect parameter estimating the impact of S deposition (Table 3) indicates that for every 10 kg S ha^{-1} yr^{-1} reduction in deposition, mean Ellenberg R increases by 0.3. This equates to 3.8% of the range of mean Ellenberg R values in the heath & bog sample and 3% of the entire range of Ellenberg values assigned to British plant species (1 to 9). The range of mean Ellenberg R is high in the heath & bog plots because of the liberal selection criteria applied so as to optimise attribution of vegetation change – see Chapter 2.

Our findings in terms of the vegetation index Ellenberg R are consistent with patterns seen for soil pH change in terrestrial vegetation summarised in RoTAP (2012). Analysis of National Soil Inventory (NSI) data (Kirk et al., 2010) and the repeat of the GB Woodland Survey (Smart et al., 2014) have shown a dependence of soil pH over time on the mean pH at sampled locations. The NSI analysis also showed a negative correlation between pH change and S deposition change but only when historical deposition was included in the model. Analysis of Countryside Survey data also clearly show that the size of pH change over time is larger the higher the average pH of the habitat sampled (see Table 5.1 in RoTAP 2012). While we have focussed on a species compositional indicator of soil pH, the consistency of the pattern is compelling and suggests that changes above-ground are to some extent coupled with changes below-ground and both are responding to legacy effects of high but declining S deposition. The effect size is small however.

7.3.3 Hypothesis 3: The eutrophying impact of cumulative N deposition is most evident where recovery from acidification has been greatest i.e. the increase in Ellenberg N is dependent not only high N deposition but has increased to a greater extent where high N deposition and declining S deposition coincide.

Here we focussed on the hypothesis that the reduction in soil acidity, as soils recover from acidification, is allowing plants to exploit previously inaccessible N, where N had either accumulated or was still being deposited at high rates. This should mean that plants typical of more productive and less acidic conditions are likely to have responded more than others in areas with larger reductions in S deposition that have also been subject to higher levels of N deposition.

The key test of this hypothesis is the importance of the interaction term in the Model 3.3 in Table 7.6 below and reported fully in Table 7.7.

MODEL STRUCTURE: Model 3. Standardised EbN at each quadrat in each year = intercept + (time standardised S deposition) + (plot-specific 2003-2005 average N deposition) + interaction between the two.

Table 7.6: Models tested and their AIC values.

Effects in model	**AIC**
S deposition (Model 3.1)	6273
S deposition + N deposition load (2003-'05) (Model 3.2)	6266
S deposition + N deposition load (2003-'05) + interaction (Model 3.3)	6266

Table 7.7: Results table for Model 3.3 (see above) including the main effects of S and N deposition and their interaction. Response variable = Standardised mean Ellenberg N at each quadrat in each year.

	Value	Std.Error	DF	t-value	p-value
S deposition	-0.0059	0.0014	4886	-4.2766	0.0000
N deposition	-0.0013	0.0008	4886	-1.6398	0.1011
Interaction	0.0000	0.0000	4886	1.2686	0.2046

Interpretation

The AIC value for Model 3.3 is no lower than the equivalent model without an interaction (Model 3.2), which indicates that the interaction is not important, and therefore that this hypothesis does not hold. This is not likely to be an artefact of the spatial correlation between N and S deposition estimates which are only weakly correlated (see Figure 7.7). Our conclusion here is that there is no evidence that the relationship between change in Ellenberg N and change in S deposition is additionally dependent on N deposition load.

Climate-related hypotheses

With all the hypotheses that follow we sought to test whether trends in climate variables can explain a fraction of the observed trend in mean Ellenberg values across the schemes and 49 years of observations. The mechanisms envisaged here differ fundamentally from those where climate may act as a spatial correlate of spatial or temporal change in vegetation. Using three climate variables, we tested whether any trend across the sampled locations toward climate conditions more amenable to the growth of competitive versus broadly stress-tolerant species might also explain change in Ellenberg

R and N. We also tested explicitly whether long-term increases in rainfall could explain change in Ellenberg Wetness, that is a shift toward favouring wetter conditions.

7.3.4 Hypothesis 4: A climate favourable to nitrogen-loving plants is associated with a greater increase in Ellenberg N.

Models 4.1 and 4.2 (below) were structured to test whether change in Ellenberg N could be linked to change in minimum January temperature.

MODEL STRUCTURE: Model 4.1. Standardised EbN at each quadrat in each year = intercept + (time standardised minimum January temperature) + (year of survey) + interaction between the two

Table 7.8: Model 4.1. Results table. Year and minimum January temperature treated separately (no interaction). Response variable = Standardised mean Ellenberg N at each quadrat in each year.

	Value	Std.Error	DF	t-value	p-value
Year	0.0041	0.0005	4887	7.9101	0.0000
Min Jan temp	-0.0004	0.0039	4887	-0.1120	0.9108

Table 7.9: Model 4.2. Results table. As for Model 4.1 but with an added interaction between Year and min Jan temperature. Response variable = Standardised mean Ellenberg N at each quadrat in each year.

	Value	Std.Error	DF	t-value	p-value
Year	0.0041	0.0005	4886	7.9013	0.0000
Min Jan temp	0.0628	0.5715	4886	0.1099	0.9125
Interaction	-0.0000	0.0002	4886	-0.1107	0.9118

Interpretation

The results do not indicate any effect of change in minimum January temperature on the change in mean Ellenberg N between plots over time.

7.3.5 Hypothesis 5: An increase in total rainfall has increased the occurrence of plants of wetter conditions or/and decreased plants of drier conditions.

MODEL STRUCTURE: Model 5. Standardised Ellenberg Wetness (EbW) at each quadrat in each year = intercept + (time standardised total annual precipitation) + (year of survey).

Table 7.10: Model 5. Results table with no interaction between year and rainfall. Response variable = Standardised mean Ellenberg Wetness at each quadrat in each year.

	Value	Std.Error	DF	t-value	p-value
Year	0.0004	0.0003	4887	1.2671	0.2052

Annual rainfall	-0.0000	0.0000	4887	-0.5167	0.6054

Interpretation

The results provide no indication of a relationship between changes in Ellenberg Wetness values and precipitation over time. It is possible that annual rainfall is a somewhat blunt instrument as an explanatory variable and future work should test for relationships with other possible covariates such as summer rainfall and estimates of evaporative loss. Other response variables could also be tested where feasible, for example total bryophyte cover or *Sphagnum* cover, where focussing just on heath and bog.

7.3.6 Hypothesis 6: An increase in precipitation, through its influence on soil moisture, also explains increases in Ellenberg R in addition to a separate effect of decreased S deposition.

MODEL STRUCTURE: Model 6. Standardised EbR at each quadrat in each year = intercept + (time standardised annual precipitation) + (time standardised S deposition) + (year of survey) + interaction between the three

Table 7.11: Model 6.1 Results table. Response variable = Standardised mean Ellenberg R at each quadrat in each year.

	Value	Std.Error	DF	t-value	p-value
Year	0.0041	0.0008	4886	5.4750	0.0000
Annual precip	-0.0000	0.0000	4886	-2.8965	0.0038
S deposition	-0.0007	0.0008	4886	-0.7938	0.4274

Table 7.12: Model 6.2 Results table with added interactions between year, precipitation and S deposition. Response variable = Standardised mean Ellenberg R at each quadrat in each year.

	Value	Std.Error	DF	t-value	p-value
Year	0.0041	0.0008	4882	5.2846	0.0000
Annual precip	0.0057	0.0062	4882	0.9109	0.3624
S deposition	-0.1741	0.0834	4882	-2.0886	0.0368
Year * Annual precip	-0.0000	0.0000	4882	-0.9242	0.3554
Year * S deposition	0.0001	0.0000	4882	2.0845	0.0372
Annual precip * S deposition	0.0003	0.0003	4882	1.1680	0.2429
3-way interactions	0.0000	0.0000	4882	-1.1837	0.2366

Interpretation

During the testing of Hypothesis 2 we found a strong year ~ S deposition correlation and so the lack of significance of S deposition in Model 6.1 is not surprising given the inclusion of Year in this model. The model, however, suggests that changes in mean Ellenberg R between plots and over time tend to be smaller in areas where precipitation has increased most. Further work is required to map these patterns to determine any geographical patterning in the apparent relationships between change in precipitation and species that vary in their association with higher or lower pH conditions.

Tests of interaction terms indicate no evidence for any conditional relationships between mean Ellenberg R in response to time and rainfall and S deposition (Table 7.12). Interestingly though, and in contrast to modelling under Hypothesis 2, S deposition does feature as a significant predictor in Model 6.2, while the Year * S deposition interaction is also significant. This provides further support for the idea that S deposition is an important driver of Ellenberg R, but only when considered in combination with a more complex model formulation. It remains perfectly possible that the change in S deposition is a contributor, if not the dominant contributor, to the Ellenberg R shift. At this stage it seems reasonable to conclude that our modelling implicates S deposition in changing Ellenberg R, but limitations in the number of covariates we could investigate and the complexity of models we could fit mean that further work may be necessary to explain changes with greater confidence.

7.3.7 Hypothesis 7: An increase in precipitation in addition to a separate effect of N deposition is correlated with an increase in Ellenberg N.

MODEL STRUCTURE: Model 7. Standardised EbN at each quadrat in each year = intercept + (time standardised annual precipitation) + (plot-specific 2003-2005 average N deposition) + (year of survey) + interaction between the two

Table 7.13: Model 7.1 Results table with no interaction between year, rainfall and N deposition. Response variable = Standardised mean Ellenberg N at each quadrat in each year.

	Value	Std.Error	DF	t-value	p-value
Year	0.0041	0.0005	4886	8.5607	0.0000
Annual rainfall	-0.0000	0.0000	4886	-1.9621	0.0498
N deposition	-0.0001	0.0006	4886	-0.2088	0.8346

Table 7.14: Model 7.2. Results table with added interaction between year, rainfall and N deposition. Response variable = Standardised mean Ellenberg N at each quadrat in each year.

	Value	Std.Error	DF	t-value	p-value
Year	0.0021	0.0010	4882	2.1482	0.0317
Annual rainfall	0.0042	0.0075	4882	0.5599	0.5756
N deposition	-0.2923	0.1172	4882	-2.4930	0.0127
Year * Annual rainfall	-0.0000	0.0000	4882	-0.5635	0.5731
Year * N deposition	0.0001	0.0001	4882	2.4922	0.0127
Annual rainfall * N deposition	-0.0003	0.0004	4882	-0.7803	0.4352
3-way interaction	0.0000	0.0000	4882	0.7802	0.4353

Interpretation

Here we tested a hypothesis generated by evidence for the positive effect of seasonal rainfall on species able to readily exploit higher macronutrient availability. Analysis of a 35 year trend in weather and vegetation indicated that wetter summers favoured nutrient-loving, perennial grasses, a signal likely to translate into an increase in Ellenberg N (Dunnett et al 1998). The ability of changing rainfall to drive increased grass biomass with competitive impacts on other species, hence a similar signal to that expected from eutrophication, was also evidenced by Silvertown et al (1994) based on analysis of the long running Rothamsted Park Grass experiment. Here the key test was therefore the interaction between plot centred annual rainfall (a covariate that conveys differences in rainfall at each plot location over time) and N deposition. Results indicated that this interaction was not significant (Table 7.14). Hence based on analysis of patterns across the four schemes we did not find evidence of a rainfall related enhancement of the positive N deposition effect on mean Ellenberg N detected in hypothesis 1 (see Table 7.1).

7.4 Attribution modelling discussion

Our cross-scheme analysis provides new evidence of relationships between S and N deposition and mean Ellenberg values. These patterns are consistent with the impact of eutrophication and recovery from acidification on heath and bog plant assemblages over 49 years. While these are correlative relationships only, our application of appropriate standardisation of response and covariates allowed us to focus solely on changes through time across the multiple schemes without the confounding effects of spatial variation. Being able to directly quantify change over time avoids some of the pitfalls that beset spatial gradient analysis as a source of indirect evidence of the impacts of global change drivers on ecosystems over time (e.g. Damgaard 2019).

The effect sizes we detected are small, yet this is to be expected outside the realm of a controlled experiment or heavily constrained sampling domain. Here we attribute a response to pollutant deposition as a fraction of a larger amount of variation most of which as yet remains unexplained. The benefit is that we have analysed a realistic and representative sample of heath and bog across the British landscape.

Our case study focussed on heath and bog did not recover any signals attributable to linear trends in climate variables. More sensitive covariates such as summer rainfall, would be worth trying in future work as well as in any extension of the attribution modelling to other vegetation types.

8 Discussion and conclusions

The primary aim of this project was to determine the potential to combine the vegetation data of some of the most established national monitoring and survey schemes within a single analysis in order to maximise our understanding of long-term vegetation change across the UK. It was decided from the outset that we would focus on heath and bog vegetation as a proof of concept. While the ultimate focus was on quantifying, and understanding the causes of, long-term change, the process required a sequence of steps, several of which we consider valuable in their own right.

First, computer code was produced to enable efficient extraction of all vegetation data (i.e. not only from heath and bog) from the various scheme databases or other repositories. This can be re-applied whenever these databases are updated. In the course of this step we encountered some problems associated with the consistency of raw data formats. As a consequence of working through these issues with the data providers, we were able to provide advice on best practice which they have subsequently taken up. In the course of the data extraction process we also gathered information on the technical specifics of the various sampling methodologies applied by the different schemes, that then fed into later steps.

Second, species identifiers used by the different schemes were harmonised through the production of a common species dictionary. Code was produced as a package "*vegtaxon*", in the statistical programming language R, which matches Latin names of UK vascular plant species to the current accepted name. Considerable interest has already been expressed in the potential of the package to make integration and harmonisation of vegetation data from different sources simpler and more efficient – including with regard to Defra's developing UK APIENS project that reports on air quality impacts on ecosystems. Species-specific values for a range of indicators were then linked to the harmonised species names.

Third, a method was developed to identify all plots from the full integrated vegetation dataset that, at any point in their records, showed the necessary characteristics of either heaths or bogs. This approach has already been re-applied under a separate 25YEP indicator within the current UKCEH-Defra MoA, focussed on unimproved grassland. A range of the most appropriate vegetation indicators were subsequently selected to characterise spatial and temporal variation in the heath and bog assemblages.

Fourth, we then carried out a first round of modelling of time trends in the Ellenberg and CSM metrics of the four schemes separately. This demonstrated remarkably tight agreement between schemes in temporal patterns and rates of change in the selected Ellenberg metrics, particularly with respect to Ellenberg N and Ellenberg R, while also highlighting significant differences in the average levels between schemes. This gave us some confidence in the potential to bring these data together in a single analysis, providing the difference in levels was accounted for. In contrast to the comparison of Ellenberg trends, we found considerable disagreement between schemes in the temporal patterns of the CSM metrics, which provided our first warning that integration of these data in a single analysis was likely to be challenging.

Fifth, we developed an original approach to simulating vegetation swards, in order to test the effect that differences between sampling protocols may exert on the calculation of the range of metrics of interest. The simulated plots were sufficiently realistic to show the expected species area relationship, and fall within the ordination range for heath and bogs determined for this project. On the basis of the simulations we concluded that plot size did not affect mean Ellenberg score, whether the vegetation data were cover weighted or not. The computation of Ellenberg N seemed particularly robust to methodological differences between schemes, suggesting that modelling of this metric derived from integrated schemes should be similarly robust.

The simulations provided further evidence that the calculation of CSM scores is much more sensitive to the specific scheme sampling methodologies. By simulating variation in the plot assemblage across an environmental gradient, as a surrogate for temporal change, we were also able to show that estimates of change in Ellenberg metrics were consistent across schemes. While we are confident in the general simulation approach, it is still relatively simplistic and does not yet take into account the tendency of some species to cluster spatially or form strong associations with other species. We therefore intend to continue developing the approach, and envisage widespread application of the method in developing better models for assessing vegetation survey data, and potential future survey design.

As a consequence of the sequence of steps mapped out above we were able to design the most appropriate structure for our integrated models to describe change in the vegetation metrics. These included random slopes and random intercepts to allow for differences between schemes. Unsurprisingly, given the comparisons of trends between schemes, we found the integrated models for the Ellenberg metrics worked extremely well, and provided consistent and precise estimates of change over time for all three of them. The NPMS time series is too short to date for trends to be clearly discernible.

In contrast, and again consistent with observations made in the earlier steps, we found little evidence for the value of combining data from multiple schemes for the assessment of change in the CSM metrics. The direction of trends in the integrated CSM models was highly uncertain as a result of conflicting trends in the individual datasets. The extent to which these differences result from the sensitivity of these metrics to plot size and methods, as opposed to the potentially more spatially and temporally heterogeneous variation in the metrics themselves, still requires further investigation.

The introduction of covariates to attempt to explain variation and change in the integrated Ellenberg signals sheds new light on the factors that appear to be dominating the modelled increases in Ellenberg N and Ellenberg R. Unusually for assessments of long term vegetation change, our approach of standardising both response and explanatory variables removes the potential influence of spatial effects and therefore allows a more robust test of hypotheses concerning the drivers of temporal change.

Our analysis provides the most robust quantification to date of national-scale increases in Ellenberg N, Ellenberg R and Ellenberg W metrics. It also adds considerable strength to a developing evidence base that suggests that terrestrial vegetation in the UK is changing progressively, albeit very gradually, in response to long-term shifts in in the deposition and accumulation of air pollutants.

The clearest signal in our heath and bog data, a progressive increase in Ellenberg N, occurred in CS, ECN and LTMN data at remarkably similar rates. This is particularly striking, given that LTMN commenced only in 2010. While we only tested for linear trends, the similar rates of change quantified across the very different time scales of the three schemes suggests this trend towards more-nutrient loving species has been occurring over many years and appears to have continued until quite recently at least.

Our attribution modelling upheld the hypothesis that the change in Ellenberg N in these communities is linked directly to the amount of reactive N deposited historically, although it was not possible to discern whether the continuing rise in the metric represents a lagged response to the historical load, or whether there is sufficient contemporary deposition for these communities to be continuing to respond dynamically to continued soil N accumulation. This distinction is clearly important from an air quality policy perspective, but will require further investigation, for example through the development of non-linear modelling approaches capable of capturing changes in the rate of change in Ellenberg N over time, and by exploring relationships between N deposition, soil chemistry and vegetation where these are measured together, i.e. at ECN sites. Currently there is a trade-off between the advantages of including multiple datasets in an integrated model and the flexibility to include e.g. non-linear terms

or site-level covariates when modelling datasets individually. Developing methods to address this would allow more sophisticated integrated models to be constructed.

The integrated model of Ellenberg R also demonstrated a clear cross-scheme shift towards species with a preference for less acidic conditions, albeit with slightly more variation between schemes in the response relative to the change in Ellenberg N. We found tentative evidence to link the response directly to an atmospheric deposition driver – in this case the large reduction in sulphur deposition that has been occurring progressively since the 1980s to the present and is driving a reduction in soil acidity. Interestingly, our more detailed model demonstrated that the response in Ellenberg R to the reduction in S deposition was stronger in habitats with assemblages indicative of less acid conditions. This again needs to be explored in further detail, but it seems likely that the primary explanation for this is that more peaty environments are rich in (natural) organic acids that provide significant buffering against changes in mineral (pollutant) acidity, and have therefore responded less to changes in acid deposition regimes over time. A second potentially contributory factor is that the soil chemistry of some chronically acidified soils may not yet have recovered to a point at which acid sensitive species are able to thrive. For example, levels of inorganic aluminium could still present toxicity barriers to some taxa. Our hypothesis that responses of Ellenberg N to nitrogen fertilisation were conditional on the extent of recovery from acidification was not upheld but deserves more thorough investigation.

Our integrated model of temporal change in Ellenberg W again provided evidence of a very gradual but highly significant cross-network upward trend, this time indicating a progressive shift towards species with a preference for wetter conditions. In contrast to Ellenberg N and R though, we were unable to establish a significant link with climate variables, the most obvious hypothesis being that there would be a significant effect of annual precipitation, which has been increasing across most of the UK over the full period covered by the schemes. It is quite feasible, however, that annual precipitation is too coarse a measure for explaining the response. With more time we would have explored other hydrological explanations such as changes in precipitation at a seasonal scale – there has also been a very strong increase in summer precipitation over the survey period for example, while seasonal estimates of water balance, or soil moisture might also be appropriate although relatively uncertain. Further work should explore these other facets in more detail.

Overall therefore we have demonstrated that there is considerable potential for integrating data for these surveys, and possibly others, in order to shed new light on the nature and causes of vegetation change. The value of the integration is brought home by the significant reduction in standard errors of the temporal models for the Ellenberg metrics. The fact that there is a clear regional shift in the Ellenberg metrics over recent decades, that can be linked directly to regional changes in drivers (particularly with respect to air pollutants), clearly needs to be taken into account when assessing the potential impact of other more local drivers, such as Environmental Land Management (ELM) on these heath and bog habitats at least, and possibly others.

The work also highlights that signals of temporal change in ECN plots, that are much fewer in number and geographical spread that the other surveys, are still broadly consistent with the changes observed more widely. Since the ECN plots are monitored more frequently, and surveys are co-located with a range of other environmental measurements, including air and soil chemistry and meteorology, there would seem to be clear wider value in exploring cause-effect relationships at these intensively monitored sites. Ultimately, all these datasets have an important role to play in tracking and understanding long-term environmental change. Now we have demonstrated that the information from them can be usefully combined, the differences in characteristics such as extent, frequency and co-location should be seen as a strength of the UK's diverse set of long-term environmental monitoring and observation assets.

The project represents a major step forward in our ability to exploit the interoperability of what until now have been considered rather disparate sources of data. There is clearly potential to both explore

the signals we have begun to quantify and decipher within heath and bog habitats in much greater detail, and also to extend the approach to other habitats.

Acknowledgements

This work was funded under the UKCEH-Defra Memorandum of Agreement project 07111, and supported by the Natural Environment Research Council award number NE/R016429/1 as part of the UK-SCAPE programme delivering National Capability. The UKCEH project team are extremely grateful for the support and advice provided by a number of Natural England staff over the past two years. We particularly wish to thank Mike Morecroft, Andy Nisbet, Ruth Oatway, Sarah Grinsted, Alice Noble, David Glaves, Isabel Alonso, Iain Diack and Alistair Crowle. We also thank Oli Pescott and colleagues (UKCEH) for their help in providing access to the NPMS datasets, and Sue Rennie with respect to access to the ECN data.

References

Bartelheimer, M. and Poschod, P. (2016), Functional characterizations of Ellenberg indicator values – a review on ecophysiological determinants. Functional Ecology, 30: 506-516. https://doi.org/10.1111/1365-2435.12531

Britton, A.J., Hester, A.J., Hewison, R.L., Potts, J.M. and Ross, L.C. (2017), Climate, pollution and grazing drive long-term change in moorland habitats. Appl Veg Sci, 20: 194-203. https://doi.org/10.1111/avsc.12260

Carey, P.D., Wallis, S., Chamberlain, P.M., Cooper, A., Emmett, B.A., Maskell, L.C., McCann, T., Murphy, J., Norton, L.R., Reynolds, B., Scott, W.A., Simpson, I.C., Smart, S.M. and Ullyett, J.M. (2008), Countryside Survey: UK Results from 2007. NERC/Centre for Ecology & Hydrology (CEH Project Number: C03259).

Damgaard, C. (2019), A Critique of the Space-for-Time Substitution Practice in Community Ecology. Trends in Ecology & Evolution, 34(5): 416-421. https://doi.org/10.1016/j.tree.2019.01.013.

Diekmann, M. and Falkengren-Grerup, U. (2002), Prediction of species response to atmospheric nitrogen deposition by means of ecological measures and life history traits. Journal of Ecology, 90: 108-120.

Emmett, B.A., Rowe, E.C., Stevens, C.J., Gowing, D.J., Henrys, P.A., Maskell, L.C. and Smart, S.M. (2011), Interpretation of evidence of nitrogen impacts on vegetation in relation to UK biodiversity objectives. JNCC Report, No. 449

Kirk, G.J., Bellamy, P.H. and Lark, R.M. (2010), Changes in soil pH across England and Wales in response to decreased acid deposition. Global Change Biology, 16: 3111-3119. https://doi.org/10.1111/j.1365-2486.2009.02135.x

Mauquoy, D., Yeloff, D., Van Geel, B., Charman, D.J. and Blundell, A. (2008), Two decadally resolved records from north-west European peat bogs show rapid climate changes associated with solar variability during the mid–late Holocene. J. Quaternary Sci., 23: 745-763. https://doi.org/10.1002/jqs.1158Maskell, L.C., Scholefield, P., Rowland, C., Smart, S. and Norton, L. (2020), Progress report Year 1 on 'Habitat quality' as part of the Defra 25 YEP indicator 'Habitat quantity, quality and connectivity of habitats'.

Maskell, L., Risser, H., Rowlands, C., Scholefield, P. and Norton, L. (2021), Development of the Defra 25 YEP indicator D1 Habitat quality (YR 2 report). Report to Defra

Maskell, L.C., Botham, M., Henrys, P., Jarvis, S., Maxwell, D., Robinson, D.A., Rowland, C.S., Siriwardena, G., Smart, S., Skates, J., Tebbs, E.J., Tordoff, G.M. and Emmett, B.A. (2019), Exploring relationships between land use intensity, habitat heterogeneity and biodiversity to identify and monitor areas of High Nature Value farming. Biological Conservation, 231: 30–38.

Met Office; Hollis, D., McCarthy, M., Kendon, M., Legg, T. and Simpson, I. (2019), HadUK-Grid Gridded Climate Observations on a 1km grid over the UK, v1.0.0.0 (1862-2017). Centre for Environmental Data Analysis, 14 November 2019. Accessed: June 2020.

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

Payne, R.J., Dise, N.B., Stevens, C.J., Gowing, D.J. and BEGIN Partners (2013), Impact of nitrogen deposition at the species level. Proceedings of the National Academy of Sciences, 110(3): 984-987. DOI: 10.1073/pnas.1214299109Pescott, O.L., Walker, K.J., Harris, F., New, H., Cheffings, C.M., Newton, N., Jitlal, M., Redhead, J., Smart, S.M. and Roy, D.B. (2019), The design, launch and

assessment of a new volunteer-based plant monitoring scheme for the United Kingdom. PLOS ONE, 14(4). https://doi.org/10.1371/journal.pone.0215891

Pitcairn, C.E.R., Leith, I.D., Sheppard, L.J., Sutton, M.A., Fowler, D., Munro, R.C., Tang, S. and Wilson, D. (1998), The relationship between nitrogen deposition, species composition and foliar nitrogen concentrations in woodland flora in the vicinity of livestock farms. Environmental Pollution, 102: 41-48.

Rose, R., Monteith, D.T., Henrys, P., Smart, S., Wood, C., Morecroft, M., Andrews, C., Beaumont, D., Benham, S., Bowmaker, V., Corbett, S., Dick, J., Dodd, B., Dodd, N., Flexen, M., McKenna, C., McMillan, S., Pallett, D., Rennie, S., Schafer, S., Scott, T., Sherrin, L., Turner, A. and Watson, H. (2016), Evidence for increases in vegetation species richness across UK Environmental Change Network sites linked to changes in air pollution and weather patterns. Ecological Indicators, 68: 52-62.

RoTAP (2012) Review of Transboundary Air Pollution: Acidification, Eutrophication, Ground Level Ozone and Heavy Metals in the UK. Contract Report to the Department for Environment, Food and Rural Affairs. Centre for Ecology & Hydrology.

Rowe, E.C., Ford, A.E.S., Smart, S.M., Henrys, P.A. and Ashmore, M.R. (2016), Using Qualitative and Quantitative Methods to Choose a Habitat Quality Metric for Air Pollution Policy Evaluation. PLoS ONE, 11(8): e0161085. doi:10.1371/journal.pone.0161085.

Smart, S.M., Ellison, A.M., Bunce, R.G.H., Marrs, R.H., Kirby, K.J., Kimberley, A., Scott, A.W. and Foster, D.R. (2014), Quantifying the impact of an extreme climate event on species diversity in fragmented temperate forests: the effect of the October 1987 storm on British broadleaved woodlands. J Ecol, 102: 1273-1287. https://doi.org/10.1111/1365-2745.12291

Smart, S. et al. (2016). Modular Analysis of Vegetation Information System (MAVIS) Plot Analyser (Ver 1.03).

Smart, S.M., Robertson, E.J., Shield, E.R. and van de Poll, H.M. (2003), Locating eutrophication effects across British vegetation between 1990 and 1998. Global Change Biology, 9: 1763-1774.

Stace, C. (2010). New Flora of the British Isles. 3rd ed. Cambridge: Cambridge University Press.

Stevens, C.J., Ceulemans, T., Hodgson, J.G., Jarvis, S., Grime, J.P. and Smart, S.M. (2016), Drivers of vegetation change in grasslands of the Sheffield region, northern England, between 1965 and 2012/13. Appl Veg Sci, 19: 187-195. https://doi.org/10.1111/avsc.12206

Sykes, J.M. and Lane, A.M.J. (1996). The United Kingdom Environmental Change Network: Protocols for standard measurements at terrestrial sites, The Stationery Office.

Appendix 1 Details of data exploration for each vegetation scheme

1.1 Countryside Survey data exploration

This document includes exploration of the CS data and the five calculated indicators (CSM + , CSM -, Ellenberg N, Ellenberg R and Ellenberg W)

First, load data.

This has already been filtered to include only heath and bog plots (as defined by the cluster analysis).

```
CS <- read.csv("CS_modelling_dataset_v2.csv")

summary(CS)
```

```
##             ID            plotID        repeat_plotID        year
##  CS_889X1_NEN:   41   835X1  :  100   835RPT20 :  100   Min.   :1978
##  CS_931U5_NEN:   38   1152X1 :   83   1152RPT8 :   83   1st Qu.:1990
##  CS_835X1_NEN:   37   931X1  :   82   931RPT19 :   82   Median :1998
##  CS_825U1_NEN:   35   1115X1 :   79   1115RPT12:   79   Mean   :1998
##  CS_1054X4_NI:   34   931X3  :   78   931RPT21 :   78   3rd Qu.:2007
##  CS_922X5_NEN:   33   987X1  :   73   987RPT16 :   73   Max.   :2019
##  (Other)     :49781   (Other):49504   (Other)  :49504
##      yearF          yearOS      scheme          site
##  Min.   :1978   Min.   : 1.00   CS:49999   Min.   :   6.0
##  1st Qu.:1990   1st Qu.:13.00              1st Qu.: 732.0
##  Median :1998   Median :21.00              Median : 933.0
##  Mean   :1998   Mean   :21.07              Mean   : 861.1
##  3rd Qu.:2007   3rd Qu.:30.00              3rd Qu.:1070.0
##  Max.   :2019   Max.   :42.00              Max.   :1272.0
##
##                species          EBERGR          EBERGN
##  Potentilla erecta: 2720   Min.   :1.000   Min.   :1.000
##  Calluna vulgaris : 2597   1st Qu.:2.000   1st Qu.:2.000
##  Molinia caerulea : 2169   Median :3.000   Median :2.000
##  Sphagnum         : 1941   Mean   :3.494   Mean   :2.634
##  Galium saxatile  : 1555   3rd Qu.:4.000   3rd Qu.:3.000
##  Nardus stricta   : 1522   Max.   :8.000   Max.   :9.000
##  (Other)          :37495   NA's   :3624    NA's   :2678
##      EBERGW          CSM_POS         CSM_NEG       EBERGR_site
##  Min.   : 1.000   Min.   :1       Min.   :1       Min.   :1.000
##  1st Qu.: 6.000   1st Qu.:1       1st Qu.:1       1st Qu.:2.714
##  Median : 7.000   Median :1       Median :1       Median :3.250
##  Mean   : 6.623   Mean   :1       Mean   :1       Mean   :3.480
##  3rd Qu.: 8.000   3rd Qu.:1       3rd Qu.:1       3rd Qu.:4.100
##  Max.   :12.000   Max.   :1       Max.   :1       Max.   :8.000
##  NA's   :3620     NA's   :24450   NA's   :30985   NA's   :58
##   EBERGN_site     EBERGW_site      CSM_POS_site     CSM_NEG_site
##  Min.   :1.000   Min.   : 3.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:1.846   1st Qu.: 6.000   1st Qu.: 4.000   1st Qu.: 2.000
##  Median :2.444   Median : 6.625   Median : 6.000   Median : 4.000
##  Mean   :2.616   Mean   : 6.635   Mean   : 5.966   Mean   : 4.557
```

```
##  3rd Qu.:3.100   3rd Qu.: 7.333   3rd Qu.: 8.000   3rd Qu.: 7.000
##  Max.   :9.000   Max.   :11.000   Max.   :16.000   Max.   :14.000
##  NA's   :49      NA's   :60
##     Easting          Northing
##  Min.   : 73145   Min.   :  37267
##  1st Qu.:192604   1st Qu.: 508348
##  Median :267144   Median : 740420
##  Mean   :264602   Mean   : 672916
##  3rd Qu.:327745   3rd Qu.: 832852
##  Max.   :642102   Max.   :1217914
##
```

Note that CS locations are in Easting and Northing unlike other datasets

```
#create LATITUDE and LONGITUDE columns


## libraries
require(rgdal) # for spTransform
require(stringr)

### shortcuts
ukgrid <- "+init=epsg:27700"
latlong <- "+init=epsg:4326"

### Create coordinates variable
coords <- cbind(Easting = as.numeric(as.character(CS$Easting)),
                Northing = as.numeric(as.character(CS$Northing)))

### Create the SpatialPointsDataFrame
dat_SP <- SpatialPointsDataFrame(coords,
                                 data = CS,
                                 proj4string = CRS("+init=epsg:27700"))

### Convert
dat_SP_LL <- spTransform(dat_SP, CRS(latlong))

## replace Lat, Long
CS$LONGITUDE <- coordinates(dat_SP_LL)[, 1]
CS$LATITUDE <- coordinates(dat_SP_LL)[, 2]
```

Currently the dataset retains the species-level information so that we can recalculate new indicators if we need to at a later date. However, Hannah has already calculated the square level average/sum indicator scores (EBERGR_site etc)

We can aggregate to plot level (by ID). Note this uses repeat_plotID to ensure that revisits where the plot was moved are not counted

```
CS_plot <- aggregate(cbind(EBERGR_site, EBERGN_site, EBERGW_site, CSM_POS_
site, CSM_NEG_site) ~ ID + repeat_plotID + scheme + site + year + yearF +
yearOS + LATITUDE + LONGITUDE, data = CS, FUN = mean)
```

Total of 5609 plot visits

```
summary(CS_plot)
```

```
##              ID         repeat_plotID  scheme         site
##  CS_1005U1_NEN:   1   1020RPT11:   5   CS:5609   Min.   :   6.0
##  CS_1005U2_NEN:   1   1020RPT12:   5             1st Qu.: 671.0
##  CS_1005U3_NEN:   1   1020RPT13:   5             Median : 910.0
##  CS_1005U4_NEN:   1   1020RPT14:   5             Mean   : 831.4
##  CS_1005U5_NEN:   1   1041RPT13:   5             3rd Qu.:1058.0
##  CS_1005X1_NEN:   1   1104RPT10:   5             Max.   :1272.0
##  (Other)      :5603   (Other)  :5579
##       year          yearF          yearOS         LATITUDE
##  Min.   :1978   Min.   :1978   Min.   : 1.00   Min.   :50.18
##  1st Qu.:1998   1st Qu.:1998   1st Qu.:21.00   1st Qu.:54.15
##  Median :1998   Median :1998   Median :21.00   Median :55.90
##  Mean   :1999   Mean   :1999   Mean   :21.94   Mean   :55.62
##  3rd Qu.:2007   3rd Qu.:2007   3rd Qu.:30.00   3rd Qu.:57.31
##  Max.   :2019   Max.   :2019   Max.   :42.00   Max.   :60.84
##
##    LONGITUDE       EBERGR_site     EBERGN_site     EBERGW_site
##  Min.   :-7.482   Min.   :1.000   Min.   :1.000   Min.   : 3.000
##  1st Qu.:-5.120   1st Qu.:2.571   1st Qu.:1.800   1st Qu.: 5.933
##  Median :-3.935   Median :3.083   Median :2.353   Median : 6.571
##  Mean   :-3.934   Mean   :3.444   Mean   :2.689   Mean   : 6.582
##  3rd Qu.:-2.853   3rd Qu.:4.000   3rd Qu.:3.125   3rd Qu.: 7.333
##  Max.   : 1.553   Max.   :8.000   Max.   :9.000   Max.   :11.000
##
##   CSM_POS_site     CSM_NEG_site
##  Min.   : 0.000   Min.   : 0.00
##  1st Qu.: 2.000   1st Qu.: 1.00
##  Median : 5.000   Median : 3.00
##  Mean   : 4.555   Mean   : 3.39
##  3rd Qu.: 7.000   3rd Qu.: 5.00
##  Max.   :16.000   Max.   :14.00
##
```

Histograms of each variables

Five variables of interest: count of CSM positive indicators, count of CSM negative indicators, mean Ellenberg N, mean Ellenberg R and mean Ellenberg W. Ellenberg values were not weighted by cover.

Distributional considerations:

1. CSM values are counts and so a Poisson or negative binomial distribution are likely to be most appropriate. We'll need to think about potential overdispersion (violating the assumption of equal mean and variance assumed by a Poisson distribution)

2. Ellenberg variables are continous. Although techically bound (see below) the mean values should lie sufficiently away from the bounds to be treated without modelling the bounds. This is something to check

Theoretical bounds of the Ellenberg values:

Each Ellenberg score associated with a plant species comes from a scale with defined limits. These vary between the different scores:

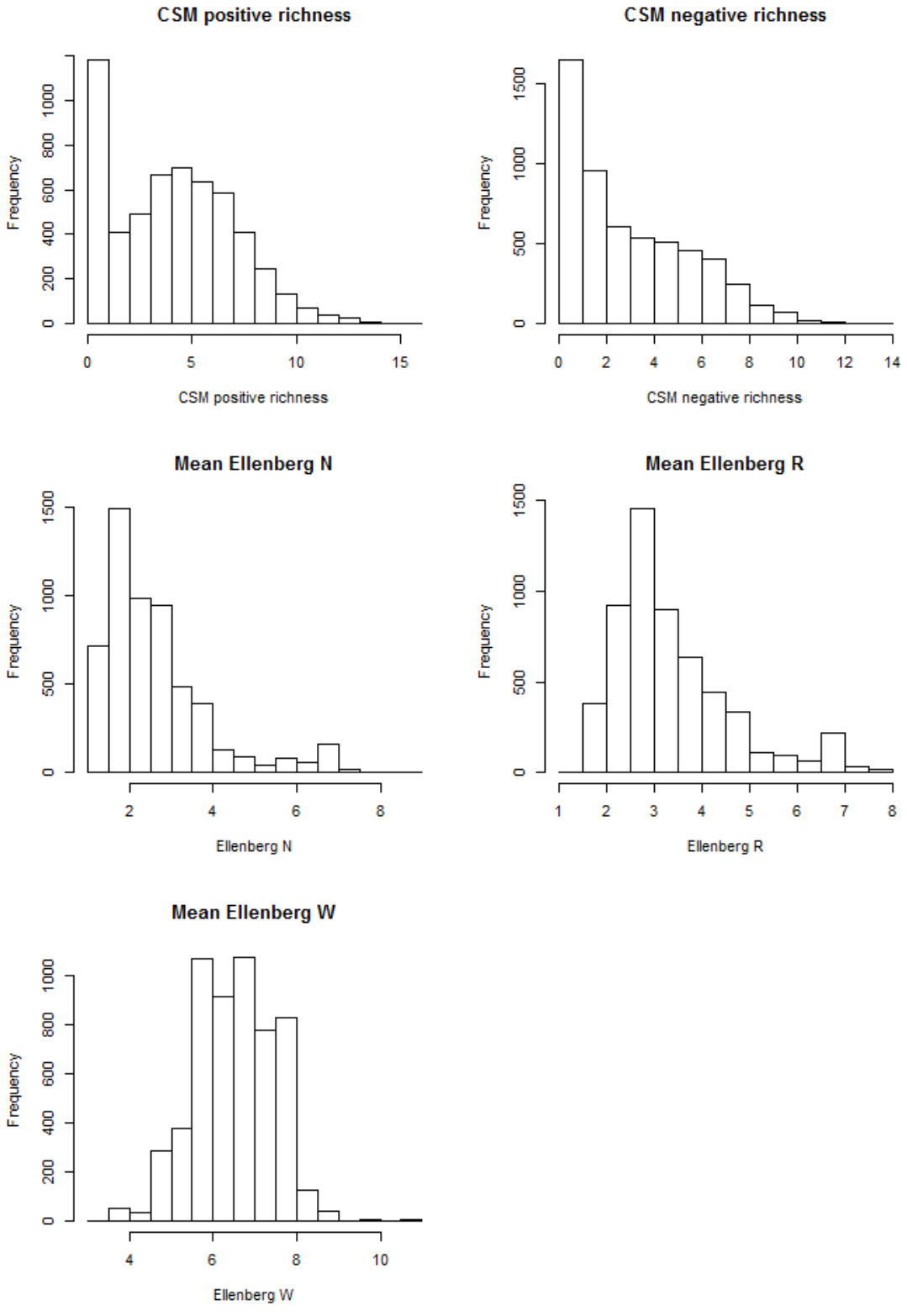
Ellenberg N: 1-9

Ellenberg R: 1-9

Ellenberg W: 1-12

To investigate the distributions of each response variable we can plot histograms:

```
par(mfrow=c(3,2))
hist(CS_plot$CSM_POS_site, main = "CSM positive richness", xlab = "CSM pos
itive richness")
hist(CS_plot$CSM_NEG_site, main = "CSM negative richness", xlab = "CSM neg
ative richness")
hist(CS_plot$EBERGN_site, main = "Mean Ellenberg N", xlab = "Ellenberg N")
hist(CS_plot$EBERGR_site, main = "Mean Ellenberg R", xlab = "Ellenberg R")
hist(CS_plot$EBERGW_site, main = "Mean Ellenberg W", xlab = "Ellenberg W")
```



The Ellenberg distributions look sort of ok for W but quite skewed for N and R. There is evidence for potential slight bimodality in Ellenberg N and R. The Ellenberg N distribution gets close to the lower bounds.

The CSM positive indicator looks potentially zero inflated but on further inspection this is just a function of poor histogram plotting. Evaluating the frequencies of CSM positives does not support excess zeroes:

```
table(CS$CSM_POS_site)
```

```
##
##    0    1    2    3    4    5    6    7    8    9   10   11   12   13
14
## 1964 1849 2348 3524 5723 6505 6498 6905 5469 3805 2229 1370  876  611
255
##   15   16
##   30   38
```

The CSM negative indicator distribution looks reasonable.

We can calculate the mean and variance of the CSMs:

```
mean(CS_plot$CSM_POS_site);var(CS_plot$CSM_POS_site)
```

```
## [1] 4.555001
```

```
## [1] 9.141455
```

```
mean(CS_plot$CSM_NEG_site);var(CS_plot$CSM_NEG_site)
```

```
## [1] 3.389909
```

```
## [1] 7.3039
```

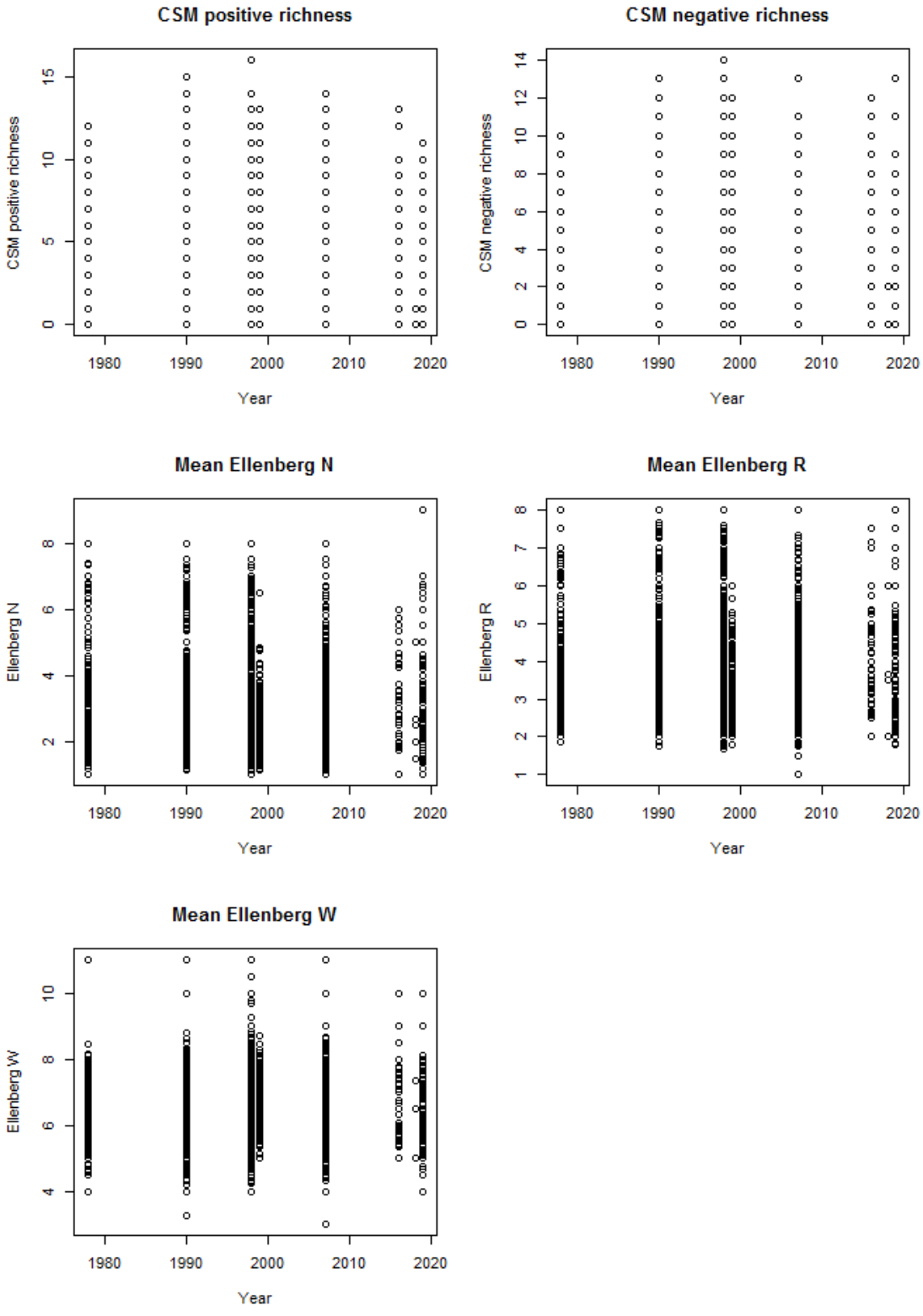
Neither show strong evidence of overdispersion

Plots of each variable against time

As the focus of this work is to look at trends in these variables over time (i.e. using year as a predictor) it is worth having a look at some exploratory scatterplots to identify any potentially non-linear relationships or datasets where there may not be sufficient temporal replication to calculate a trend.

Note that we are currently working without the 1978 data due to problems identified with this data.

```
par(mfrow=c(3,2))
plot(CS_plot$CSM_POS_site ~ CS_plot$year, main = "CSM positive richness",
ylab = "CSM positive richness", xlab = "Year")
plot(CS_plot$CSM_NEG_site ~ CS_plot$year, main = "CSM negative richness",
ylab = "CSM negative richness", xlab = "Year")
plot(CS_plot$EBERGN_site ~ CS_plot$year, main = "Mean Ellenberg N", ylab =
"Ellenberg N", xlab = "Year")
plot(CS_plot$EBERGR_site ~ CS_plot$year, main = "Mean Ellenberg R", ylab =
"Ellenberg R", xlab = "Year")
plot(CS_plot$EBERGW_site ~ CS_plot$year, main = "Mean Ellenberg W", ylab =
"Ellenberg W", xlab = "Year")
```

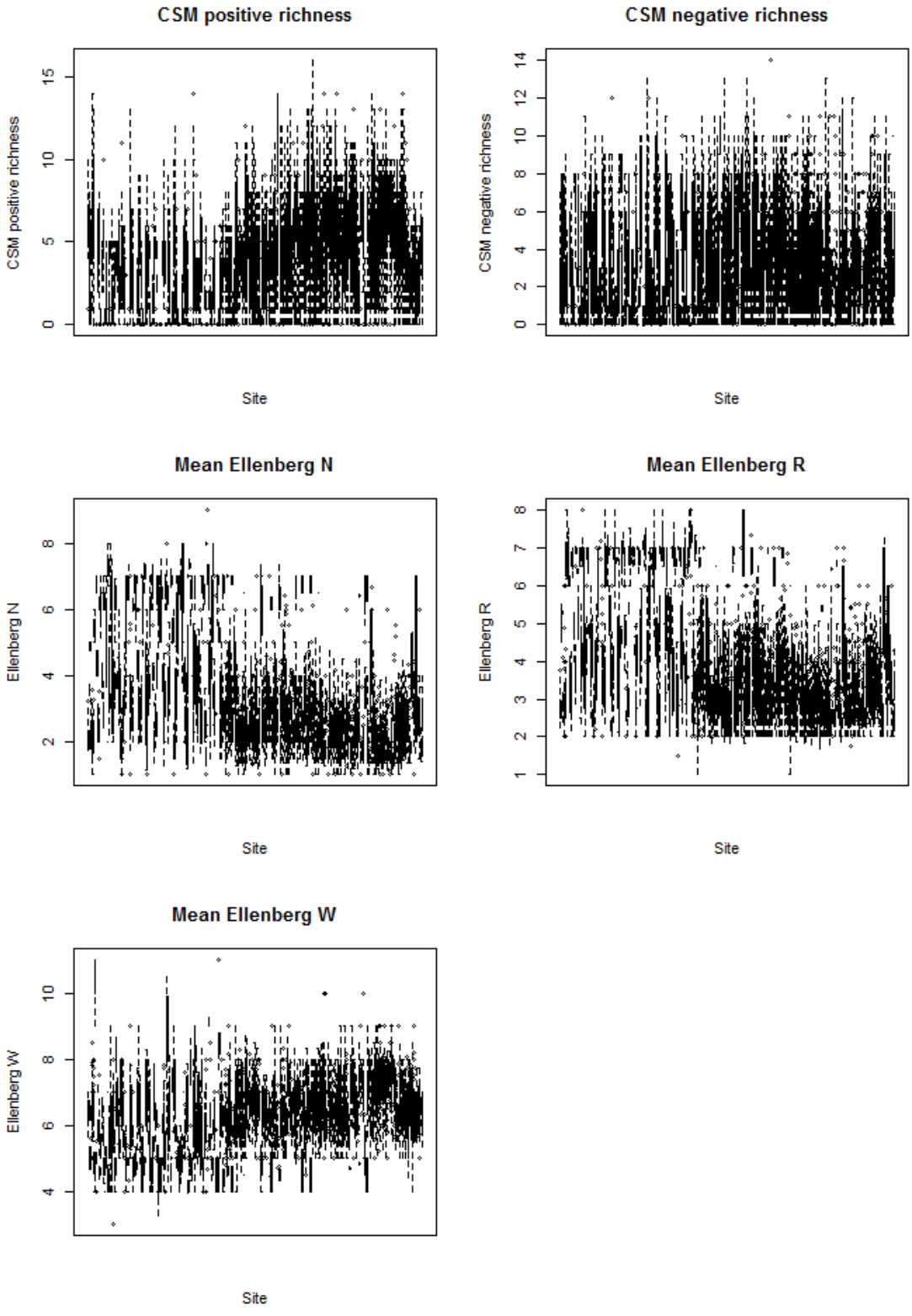


We can see here that the temporal distribution of plots varies across time. During the first 30 years of survey sampling occasions occurred every ~10 years. However, in the 98/99 survey plots were sampled across two years and in more recent times sampling occurs every 1 to 2 years. This might make calculation of temporal autocorrelation quite tricky as there are few neighbouring years of survey (and plots would never be visited two years in a row).

Plots of each variable against site

The model we plan to fit considers repeat visits to plots but not nesting of plots within sites or squares. Therefore it is useful to consider how much variation is due to site.

```
par(mfrow=c(3,2))
par(mgp= c(3,1,0))
boxplot(CS_plot$CSM_POS_site ~ CS_plot$site, main = "CSM positive richness
", ylab = "CSM positive richness", xaxt =  "n", xlab = "Site")
boxplot(CS_plot$CSM_NEG_site ~ CS_plot$site, main = "CSM negative richness
", ylab = "CSM negative richness", xaxt =  "n", xlab = "Site")
boxplot(CS_plot$EBERGN_site ~ CS_plot$site, main = "Mean Ellenberg N", yla
b = "Ellenberg N", xaxt =  "n", xlab = "Site")
boxplot(CS_plot$EBERGR_site ~ CS_plot$site, main = "Mean Ellenberg R", yla
b = "Ellenberg R", xaxt =  "n", xlab = "Site")
boxplot(CS_plot$EBERGW_site ~ CS_plot$site, main = "Mean Ellenberg W", yla
b = "Ellenberg W", xaxt =  "n", xlab = "Site")
```



It is quite difficult to tell from these plots as there are so many sites but it seems reasonable to assume that a lot of variation is due to site. However, we also know that plots are accurately revisited over time so it seems important to account for plot revisits as well as plots within a site being similar to each other.

Plots of each variable in space

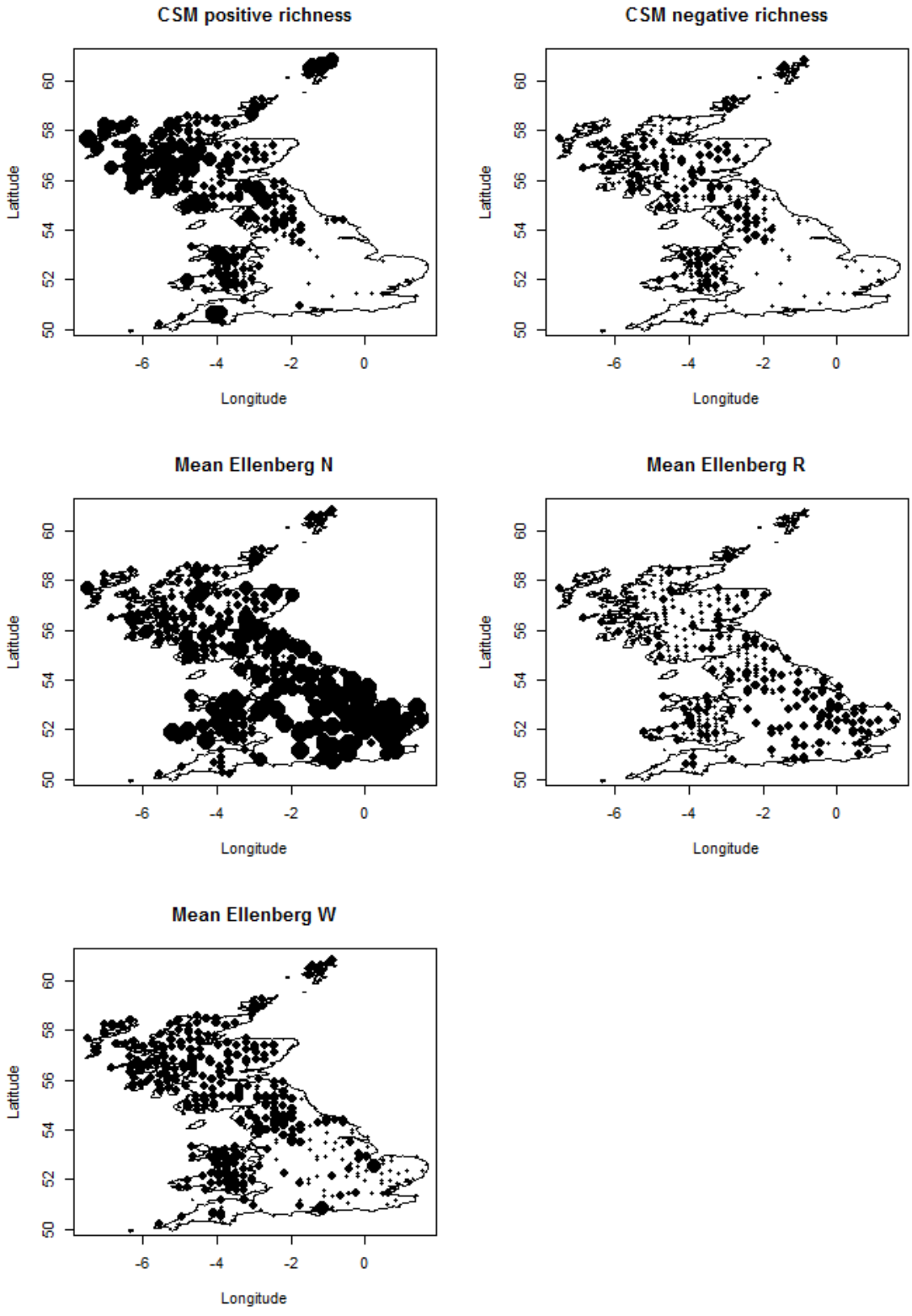
It is quite useful to think about plotting the variables in space, even in a very simplistic way. This will achieve two things:

3. Identify the spatial distribution of the data i.e. how well is the domain of interest covered?

4. Identify any potential spatial patterns in the response variables e.g. are CSM positive counts higher in the south for some reason? At this stage we've not included any covariates so spatial patterns may be due to climate, for example. We can investigate this later once we have the covariate data

The plots below show the locations of each measurement of the data, with the size of the point relative to the value of the response variable

```
#set up for figure by reading in file with GB outline

GB=read.table("GBoutline_latlong.txt",header=T)

par(mfrow=c(3,2))
par(mgp= c(3,1,0))
cex_ind <- round(CS_plot$CSM_POS_site/4)
plot(CS_plot$LATITUDE ~ CS_plot$LONGITUDE, pch = 20, cex = cex_ind, main =
"CSM positive richness", ylab = "Latitude", xlab = "Longitude")
lines(GB)
cex_ind <- round(CS_plot$CSM_NEG_site/6)
plot(CS_plot$LATITUDE ~ CS_plot$LONGITUDE, pch = 20, cex = cex_ind, main =
"CSM negative richness", ylab = "Latitude", xlab = "Longitude")
lines(GB)
cex_ind <- round(CS_plot$EBERGN/2)
plot(CS_plot$LATITUDE ~ CS_plot$LONGITUDE, pch = 20, cex = cex_ind, main =
"Mean Ellenberg N", ylab = "Latitude", xlab = "Longitude")
lines(GB)
cex_ind <- round(CS_plot$EBERGR/4)
plot(CS_plot$LATITUDE ~ CS_plot$LONGITUDE, pch = 20, cex = cex_ind, main =
"Mean Ellenberg R", ylab = "Latitude", xlab = "Longitude")
lines(GB)
cex_ind <- round(CS_plot$EBERGW/4)
plot(CS_plot$LATITUDE ~ CS_plot$LONGITUDE, pch = 20, cex = cex_ind, main =
"Mean Ellenberg W", ylab = "Latitude", xlab = "Longitude")
lines(GB)
```



There is quite strong evidence for geographical patterns in all indicators, in comparison to the other datasets investigated so far. Generally there seems to be a south-east to north-west gradient in indicators with sites in the northwest having higher CSM positive richness, higher CSM negative richness, lower Ellenberg N, lower Ellenberg R and higher Ellenberg W.

As expected there is very good coverage of heath and bog sites, covering both the upland areas in Scotland and Wales but also lowland heaths in England.

Summaries of data structure

There are a couple of other useful things we can extract about the data.

5. It would be useful to know how many times each plot has been revisited. The plot is going to be the unit over which we estimate the temporal autocorrelation so if a lot of plots have only been visited once then they won't contribute to this estimation

```
#calculate number of repeat visits

repvis <- tapply(CS_plot$year, CS_plot$repeat_plotID, function (x) length(
unique(x)))

#summarise
table(repvis)

## repvis
##    1    2    3    4    5
##  600 1284  423  243   40
```

Interestingly, despite the strength of CS being high repeats over time only 40 plots (out of a total 2590 unique repeat plot IDs) were visited on 5 occasions (the maximum possible) and most plots were revisited only twice.

If we looked at repeat visits to plot IDs…

```
repvis2 <- tapply(CS$year, CS$plotID, function (x) length(unique(x)))
table(repvis2)

## repvis2
##    1    2    3    4    5
##  533 1267  434  256   54
```

The same picture, most plots visited twice. This probably reflects the increase in survey size over time with more plots being added in later surveys. For the current surveys there may be plots that have not yet been revisited since 2007.

2. The number of plots per site might vary quite a lot between schemes so it would be good to extract some statistics about this

```
plotspersite <- tapply(CS_plot$repeat_plotID, CS_plot$site, function (x) l
ength(unique(x)))

table(plotspersite)

## plotspersite
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19
## 84 50 26 16 22 13 13 20 18 20 21 13 10 14 22 14  4  5  2
```

The range of plots per site is quite large, from 1 to 19. The maximum number of plots per site in any one year is 15 (5 X plots plus 10 U plots) but relocation may increase the number of unique repeat plot IDs. For a square with one of the highest number of unique plots (sq 804) 4 of the X plot locations moved in 1998/1999.

1.2 Environmental Change Network data exploration

This document includes exploration of the ECN data and the five calculated indicators (CSM + , CSM -, Ellenberg N, Ellenberg R and Ellenberg W)

First, load data.

This has already been filtered to include only heath and bog plots (as defined by the cluster analysis).

```
ECN <- read.csv("ECN_modelling_dataset.csv", stringsAsFactors = TRUE)

summary(ECN)
```

```
##                     ID            plotID      plot_type       year
##  ECN_VF_202_438_1999:   39   Min.   :  1.0   VC :5117   Min.   :1993
##  ECN_VF_275_59_2011 :   36   1st Qu.: 92.0   VF :6333   1st Qu.:1999
##  ECN_VF_324_61_2012 :   34   Median :319.0   VFA:5853   Median :2005
##  ECN_VC_310_308_2011:   33   Mean   :326.2              Mean   :2005
##  ECN_VF_324_61_2000 :   33   3rd Qu.:441.0              3rd Qu.:2011
##  ECN_VF_324_61_2009 :   33   Max.   :918.0              Max.   :2015
##  (Other)            :17095
##      yearF          yearOS      scheme           site
##  Min.   :1993   Min.   : 1.00   ECN:17303   T04    :6155
##  1st Qu.:1999   1st Qu.: 7.00               T07    :4196
##  Median :2005   Median :13.00               T02    :3230
##  Mean   :2005   Mean   :12.84               T11    :2408
##  3rd Qu.:2011   3rd Qu.:19.00               T12    : 806
##  Max.   :2015   Max.   :23.00               T05    : 331
##                                             (Other): 177
##                     species          EBERGR          EBERGN
##  Deschampsia flexuosa   :  756   Min.   :1.000   Min.   :1.000
##  Galium saxatile        :  728   1st Qu.:2.000   1st Qu.:2.000
##  Calluna vulgaris       :  632   Median :3.000   Median :3.000
##  Alchemilla xanthochlora:  618   Mean   :3.712   Mean   :2.971
##  Festuca ovina          :  615   3rd Qu.:5.000   3rd Qu.:4.000
##  Agrostis capillaris    :  605   Max.   :8.000   Max.   :9.000
##  (Other)                :13349   NA's   :2260    NA's   :2153
##      EBERGW          CSM_POS         CSM_NEG       EBERGR_site
##  Min.   : 2.000   Min.   :1       Min.   :1       Min.   :1.667
##  1st Qu.: 5.000   1st Qu.:1       1st Qu.:1       1st Qu.:3.071
##  Median : 6.000   Median :1       Median :1       Median :3.611
##  Mean   : 6.365   Mean   :1       Mean   :1       Mean   :3.708
##  3rd Qu.: 7.000   3rd Qu.:1       3rd Qu.:1       3rd Qu.:4.364
##  Max.   :10.000   Max.   :1       Max.   :1       Max.   :7.000
##  NA's   :2260     NA's   :10046   NA's   :10937   NA's   :1
##   EBERGN_site     EBERGW_site     CSM_POS_site     CSM_NEG_site
##  Min.   :1.143   Min.   :4.000   Min.   : 0.000   Min.   : 0.00
##  1st Qu.:2.455   1st Qu.:5.750   1st Qu.: 4.000   1st Qu.: 3.00
##  Median :3.000   Median :6.231   Median : 6.000   Median : 6.00
##  Mean   :2.965   Mean   :6.357   Mean   : 6.012   Mean   : 6.13
##  3rd Qu.:3.500   3rd Qu.:6.944   3rd Qu.: 8.000   3rd Qu.: 9.00
##  Max.   :8.000   Max.   :8.556   Max.   :18.000   Max.   :22.00
##  NA's   :1       NA's   :1
```

```
##     EASTING          NORTHING
##  Min.   :260902   Min.   : 98780
##  1st Qu.:365951   1st Qu.:528812
##  Median :371622   Median :533600
##  Mean   :356076   Mean   :575267
##  3rd Qu.:384943   3rd Qu.:623474
##  Max.   :445315   Max.   :804000
##
```

One or two missing values for EBERGN_site etc. Like this this is due to a plot with a single species with missing Ellenbergs - ECN_VF_294_11_2011 containing only Taraxacum Currently the dataset retains the species-level information so that we can recalculate new indicators if we need to at a later date. However, Hannah has already calculated the square level average/sum indicator scores (EBERGR_site etc)

Note that the data are currently in Eastings and Northings. We will leave the data as-is for now but this will need to be harmonised between datasets at a later date

We can aggregate to plot level (by ID)

```
ECN_plot <- aggregate(cbind(EBERGR_site, EBERGN_site, EBERGW_site, CSM_POS
_site, CSM_NEG_site) ~ ID + plotID + plot_type + scheme + site + year + ye
arF + yearOS + NORTHING + EASTING, data = ECN, FUN = mean)
```

Total of 1251 plot visits

```
summary(ECN_plot)

##                     ID           plotID      plot_type scheme
##  ECN_VC_309_100_1994:   1   Min.   :  1.0   VC :441   ECN:1409
##  ECN_VC_309_100_2002:   1   1st Qu.: 86.0   VF :507
##  ECN_VC_309_100_2011:   1   Median :278.0   VFA:461
##  ECN_VC_309_11_1994 :   1   Mean   :315.3
##  ECN_VC_309_11_2002 :   1   3rd Qu.:441.0
##  ECN_VC_309_11_2011 :   1   Max.   :918.0
##  (Other)            :1403
##       site          year          yearF          yearOS
##  T04    :549   Min.   :1993   Min.   :1993   Min.   : 1.00
##  T07    :275   1st Qu.:1999   1st Qu.:1999   1st Qu.: 7.00
##  T02    :243   Median :2005   Median :2005   Median :13.00
##  T11    :181   Mean   :2005   Mean   :2005   Mean   :12.55
##  T12    :101   3rd Qu.:2011   3rd Qu.:2011   3rd Qu.:19.00
##  T08    : 23   Max.   :2015   Max.   :2015   Max.   :23.00
##  (Other): 37
##     NORTHING         EASTING        EBERGR_site     EBERGN_site
##  Min.   : 98780   Min.   :260902   Min.   :1.667   Min.   :1.143
##  1st Qu.:528802   1st Qu.:365951   1st Qu.:2.875   1st Qu.:2.231
##  Median :533237   Median :373553   Median :3.364   Median :2.750
##  Mean   :572086   Mean   :356674   Mean   :3.519   Mean   :2.855
##  3rd Qu.:623698   3rd Qu.:384804   3rd Qu.:4.111   3rd Qu.:3.375
##  Max.   :804000   Max.   :445315   Max.   :7.000   Max.   :8.000
##
##   EBERGW_site     CSM_POS_site     CSM_NEG_site
##  Min.   :4.000   Min.   : 0.000   Min.   : 0.000
```

```
##  1st Qu.:5.750   1st Qu.: 4.000   1st Qu.: 1.000
##  Median :6.231   Median : 5.000   Median : 5.000
##  Mean   :6.347   Mean   : 5.311   Mean   : 4.704
##  3rd Qu.:7.000   3rd Qu.: 7.000   3rd Qu.: 7.000
##  Max.   :8.556   Max.   :18.000   Max.   :22.000
##
```

Histograms of each variables

Five variables of interest: count of CSM positive indicators, count of CSM negative indicators, mean Ellenberg N, mean Ellenberg R and mean Ellenberg W. Ellenberg values were not weighted by cover.

Distributional considerations:

1. CSM values are counts and so a Poisson or negative binomial distribution are likely to be most appropriate. We'll need to think about potential overdispersion (violating the assumption of equal mean and variance assumed by a Poisson distribution)

2. Ellenberg variables are continuous. Although techically bound (see below) the mean values should lie sufficiently away from the bounds to be treated without modelling the bounds. This is something to check

Theoretical bounds of the Ellenberg values:

Each Ellenberg score associated with a plant species comes from a scale with defined limits. These vary between the different scores:

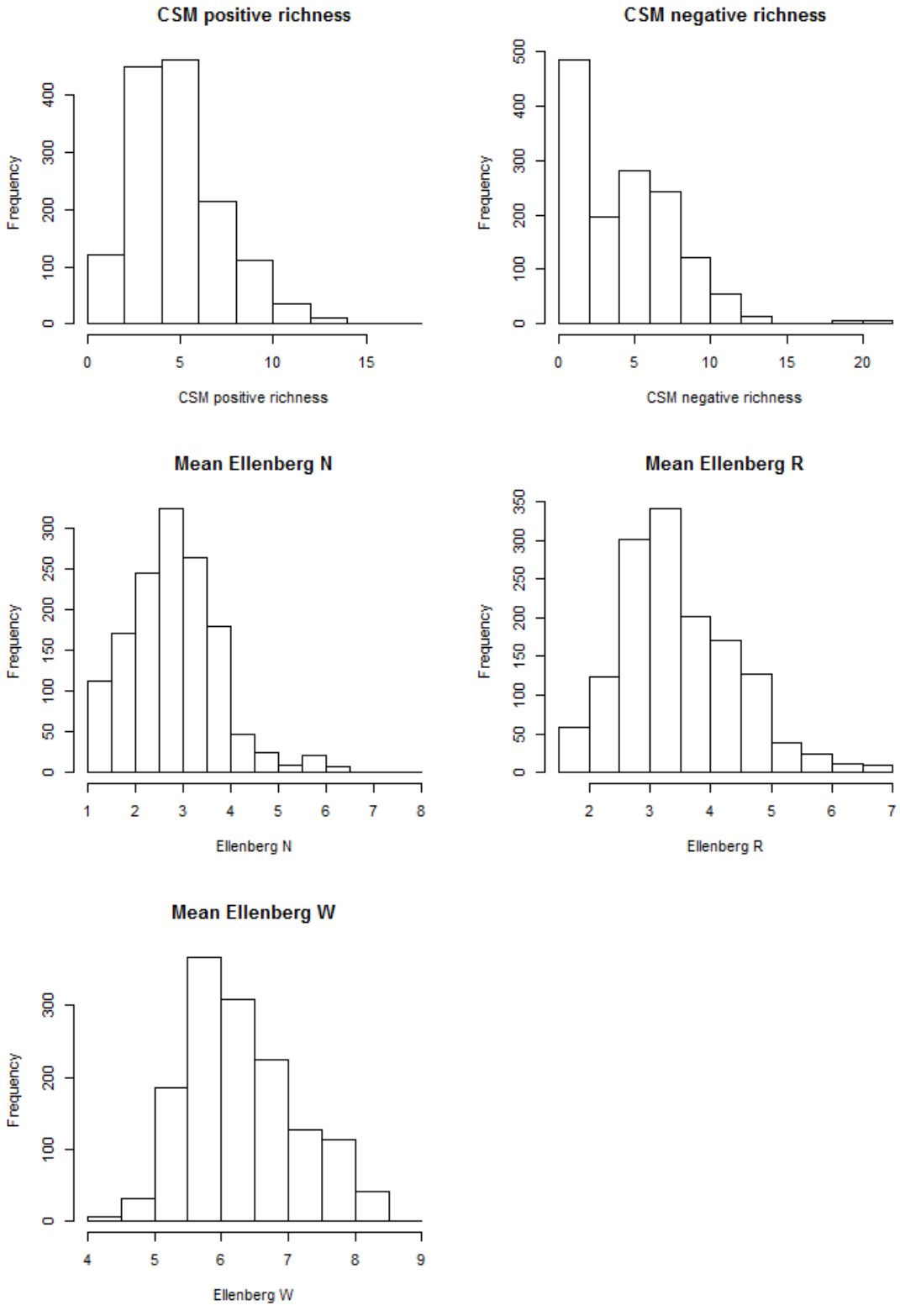
Ellenberg N: 1-9

Ellenberg R: 1-9

Ellenberg W: 1-12

To investigate the distributions of each response variable we can plot histograms:

```
par(mfrow=c(3,2))
hist(ECN_plot$CSM_POS_site, main = "CSM positive richness", xlab = "CSM po
sitive richness")
hist(ECN_plot$CSM_NEG_site, main = "CSM negative richness", xlab = "CSM ne
gative richness")
hist(ECN_plot$EBERGN_site, main = "Mean Ellenberg N", xlab = "Ellenberg N"
)
hist(ECN_plot$EBERGR_site, main = "Mean Ellenberg R", xlab = "Ellenberg R"
)
hist(ECN_plot$EBERGW_site, main = "Mean Ellenberg W", xlab = "Ellenberg W"
)
```



The distributions look ok in general. Ellenberg N and R distributions do hit the lower bound, Ellenberg N in particular is skewed. Possible indications of two populations in the CSM negative plot.

We can calculate the mean and variance of the CSMs:

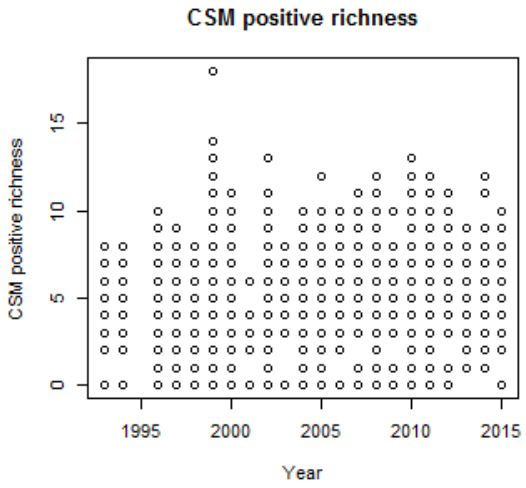
```
mean(ECN_plot$CSM_POS_site);var(ECN_plot$CSM_POS_site)

## [1] 5.310859

## [1] 6.478582

mean(ECN_plot$CSM_NEG_site);var(ECN_plot$CSM_NEG_site)

## [1] 4.704045

## [1] 14.25539
```

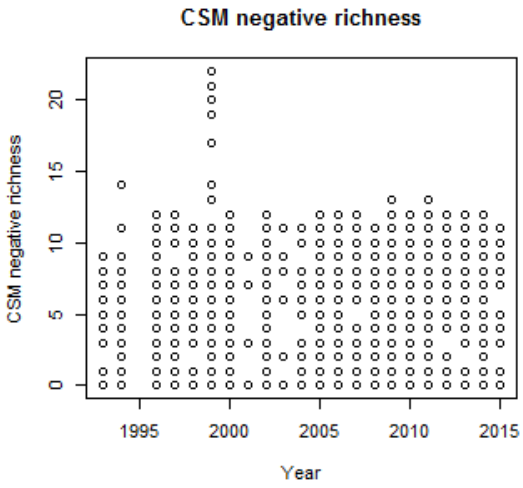
CSM negatives possibly overdispersed

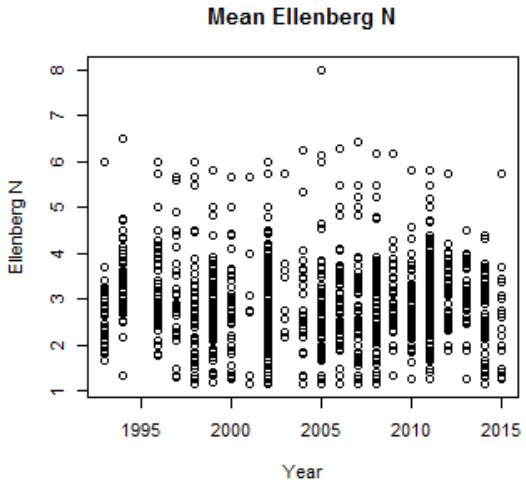
Plots of each variable against time

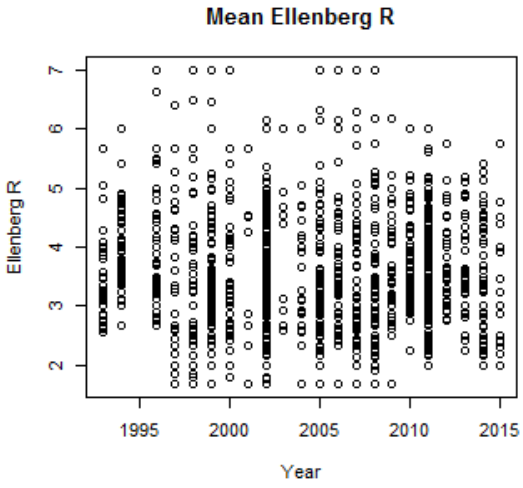
As the focus of this work is to look at trends in these variables over time (i.e. using year as a predictor) it is worth having a look at some exploratory scatterplots to identify any potentially non-linear relationships or datasets where there may not be sufficient temporal replication to calculate a trend.

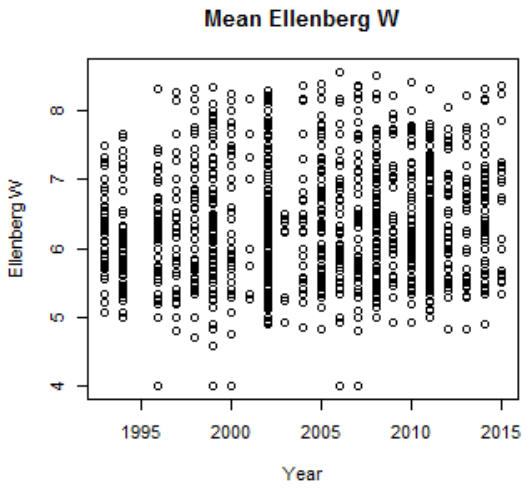
```
par(mfrow=c(3,2))
plot(ECN_plot$CSM_POS_site ~ ECN_plot$year, main = "CSM positive richness"
, ylab = "CSM positive richness", xlab = "Year")
plot(ECN_plot$CSM_NEG_site ~ ECN_plot$year, main = "CSM negative richness"
, ylab = "CSM negative richness", xlab = "Year")
plot(ECN_plot$EBERGN_site ~ ECN_plot$year, main = "Mean Ellenberg N", ylab
= "Ellenberg N", xlab = "Year")
plot(ECN_plot$EBERGR_site ~ ECN_plot$year, main = "Mean Ellenberg R", ylab
= "Ellenberg R", xlab = "Year")
plot(ECN_plot$EBERGW_site ~ ECN_plot$year, main = "Mean Ellenberg W", ylab
= "Ellenberg W", xlab = "Year")
```









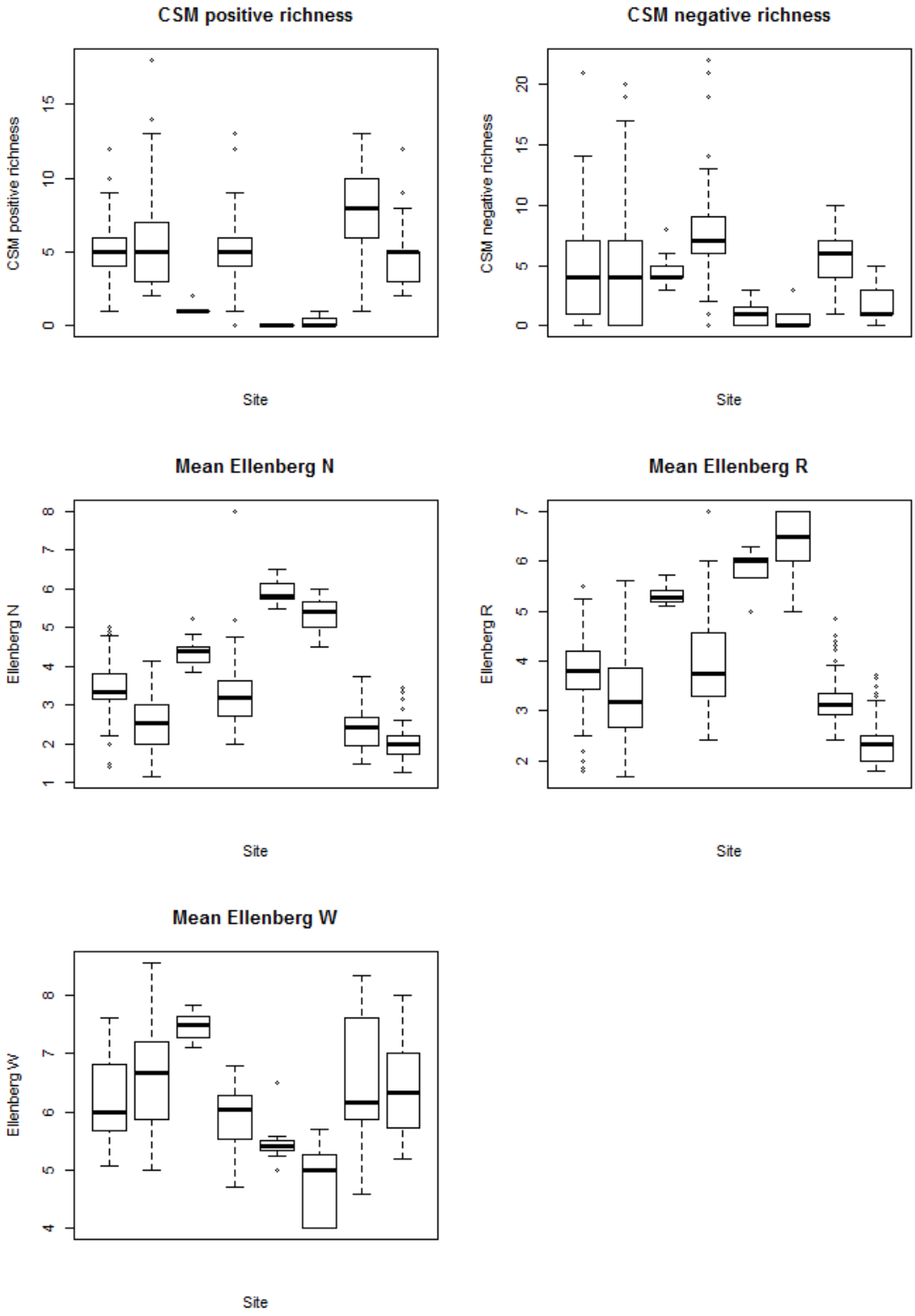


Similar to the LTMN data it might be that different cohorts are measured in different years? Some indications of non-linear patterns.

Plots of each variable against site

The model we plan to fit considers repeat visits to plots but not nesting of plots within sites or squares. Therefore it is useful to consider how much variation is due to site.

```
par(mfrow=c(3,2))
par(mgp= c(3,1,0))
boxplot(ECN_plot$CSM_POS_site ~ ECN_plot$site, main = "CSM positive richne
ss", ylab = "CSM positive richness", xaxt =  "n", xlab = "Site")
boxplot(ECN_plot$CSM_NEG_site ~ ECN_plot$site, main = "CSM negative richne
ss", ylab = "CSM negative richness", xaxt =  "n", xlab = "Site")
boxplot(ECN_plot$EBERGN_site ~ ECN_plot$site, main = "Mean Ellenberg N", y
lab = "Ellenberg N", xaxt =  "n", xlab = "Site")
boxplot(ECN_plot$EBERGR_site ~ ECN_plot$site, main = "Mean Ellenberg R", y
lab = "Ellenberg R", xaxt =  "n", xlab = "Site")
boxplot(ECN_plot$EBERGW_site ~ ECN_plot$site, main = "Mean Ellenberg W", y
lab = "Ellenberg W", xaxt =  "n", xlab = "Site")
```



Relatively few sites compared to other schemes and huge variation between them. Including a site level random effect might be important for this scheme.

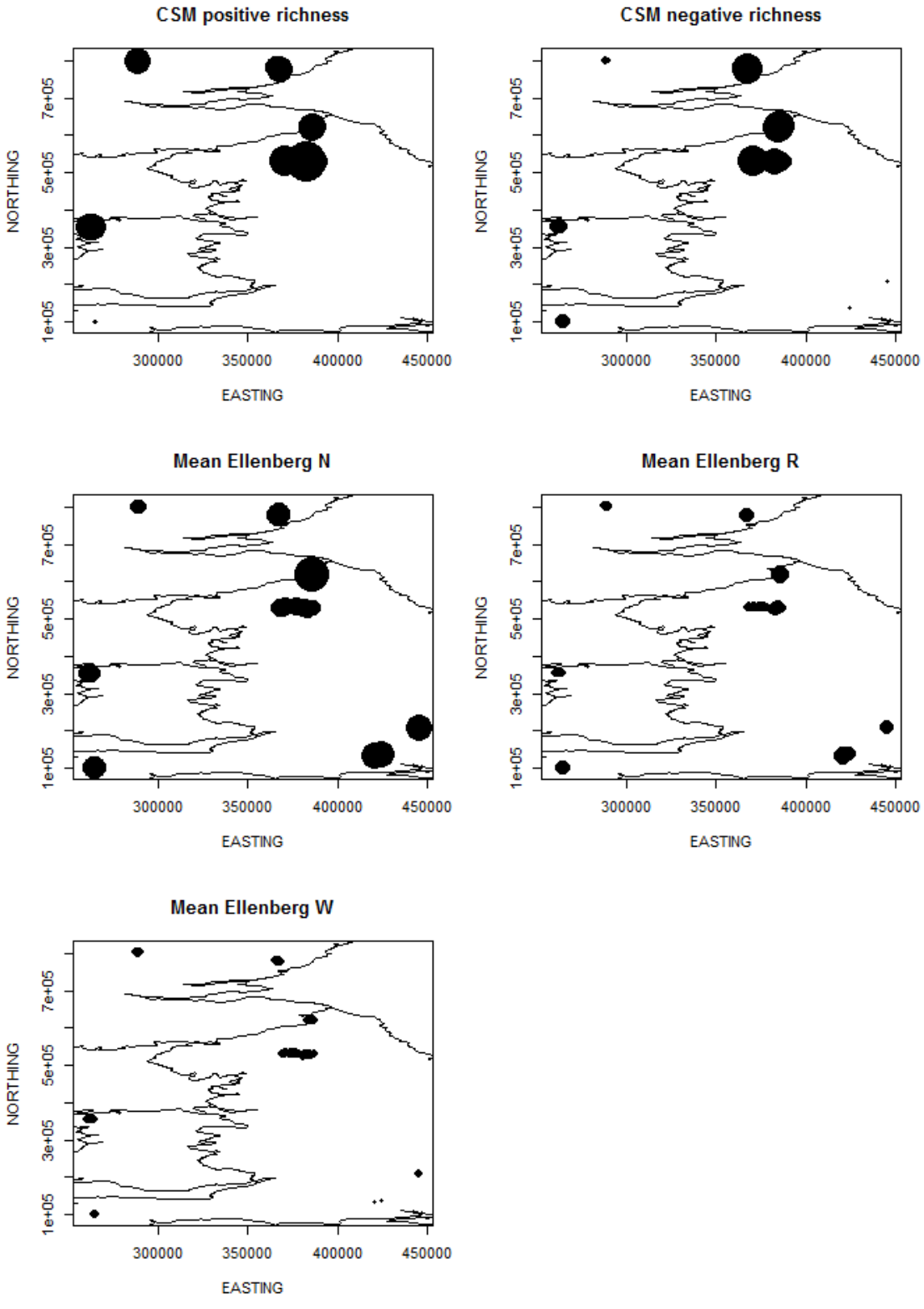
Plots of each variable in space

It is quite useful to think about plotting the variables in space, even in a very simplistic way. This will achieve two things:

1. Identify the spatial distribution of the data i.e. how well is the domain of interest covered?
2. Identify any potential spatial patterns in the response variables e.g. are CSM positive counts higher in the south for some reason? At this stage we've not included any covariates so spatial patterns may be due to climate, for example. We can investigate this later once we have the covariate data

The plots below show the locations of each measurement of the data, with the size of the point relative to the value of the response variable

```
#set up for figure by reading in file with GB outline

GB=read.table("GBoutline.txt",header=T)

par(mfrow=c(3,2))
par(mgp= c(3,1,0))
cex_ind <- round(ECN_plot$CSM_POS_site/2)
plot(ECN_plot$NORTHING ~ ECN_plot$EASTING, pch = 20, cex = cex_ind, main =
"CSM positive richness", ylab = "NORTHING", xlab = "EASTING")
lines(GB)
cex_ind <- round(ECN_plot$CSM_NEG_site/3)
plot(ECN_plot$NORTHING ~ ECN_plot$EASTING, pch = 20, cex = cex_ind, main =
"CSM negative richness", ylab = "NORTHING", xlab = "EASTING")
lines(GB)
cex_ind <- round(ECN_plot$EBERGN)
plot(ECN_plot$NORTHING ~ ECN_plot$EASTING, pch = 20, cex = cex_ind, main =
"Mean Ellenberg N", ylab = "NORTHING", xlab = "EASTING")
lines(GB)
cex_ind <- round(ECN_plot$EBERGR/2)
plot(ECN_plot$NORTHING ~ ECN_plot$EASTING, pch = 20, cex = cex_ind, main =
"Mean Ellenberg R", ylab = "NORTHING", xlab = "EASTING")
lines(GB)
cex_ind <- round(ECN_plot$EBERGW/4)
plot(ECN_plot$NORTHING ~ ECN_plot$EASTING, pch = 20, cex = cex_ind, main =
"Mean Ellenberg W", ylab = "NORTHING", xlab = "EASTING")
lines(GB)
```



Compared to other schemes ECN sites have limited geographic coverage meaning little can be seen from these plots in terms of spatial pattern. It may also mean that this scheme contributes little to any spatial understanding of patterns in indicator values.

Summaries of data structure

There are a couple of other useful things we can extract about the data.

1. It would be useful to know how many times each plot has been revisited. The plot is going to be the unit over which we estimate the temporal autocorrelation so if a lot of plots have only been visited once then they won't contribute to this estimation

```
#calculate number of repeat visits (note ECN plot IDs are not unique acros
s sites)

ECN_plot$plotID_new <- paste(ECN_plot$site, ECN_plot$plotID, sep = "_")

repvis <- tapply(ECN_plot$year, ECN_plot$plotID_new, function (x) length(u
nique(x)))

#summarise
table(repvis)

## repvis
##   1   2   3   4   5   6   7   8  10  14  15  16  17  18  19
##   4  37 113   7   1   7  31   1   4   1   1  11   3  12   8
```

This suggests a huge spread in revisit frequencies from 1 to 19! Most likely number of revisits is 3 but quite a few plots with more than that

It looks like the VFA plots occur between VF surveys for some sites, increasing the temporal frequency of revisits. Moor House has vegetation data for 19 years!

2. The number of plots per site might vary quite a lot between schemes so it would be good to extract some statistics about this

```
plotspersite <- tapply(ECN_plot$plotID, ECN_plot$site, function (x) length
(unique(x)))

table(plotspersite)

## plotspersite
##  2 23 43 44 51 74
##  3  1  1  1  1  1
```

The range of plots per site is also very wide, from 2 to 74. Only 8 of the ECN sites are included in this dataset

1.3. Long Term Monitoring Network data exploration

This document includes exploration of the LTMN data and the five calculated indicators (CSM + , CSM -, Ellenberg N, Ellenberg R and Ellenberg W)

First, load data.

This has already been filtered to include only heath and bog plots (as defined by the cluster analysis).

```
LTMN <- read.csv("LTMN_modelling_dataset_v3.csv")

summary(LTMN)
```

```
##                      ID            plotID            year
##  LTMN_VC_B38_25_2013 :   34   Min.   :  0.00   Min.   :2010
##  LTMN_VC_B38_25a_2018:   34   1st Qu.: 15.00   1st Qu.:2013
##  LTMN_VC_B12_26_2014 :   33   Median : 28.00   Median :2015
##  LTMN_VC_B38_26_2018 :   33   Mean   : 32.08   Mean   :2015
##  LTMN_VC_B38_24_2013 :   32   3rd Qu.: 43.00   3rd Qu.:2018
##  LTMN_VC_B03_44a_2017:   31   Max.   :128.00   Max.   :2019
##  (Other)             :10234
##      yearF          yearOS        scheme           site
##  Min.   :2010   Min.   : 1.000   LTMN:10431   B38    :1233
##  1st Qu.:2013   1st Qu.: 4.000                B49    : 926
##  Median :2015   Median : 6.000                B47    : 680
##  Mean   :2015   Mean   : 6.474                B10    : 663
##  3rd Qu.:2018   3rd Qu.: 9.000                B40    : 611
##  Max.   :2019   Max.   :10.000                B29    : 578
##                                               (Other):5740
##                      species         EBERGR          EBERGN
##  Calluna vulgaris        : 821   Min.   :1.000   Min.   :1.000
##  Erica tetralix          : 586   1st Qu.:2.000   1st Qu.:1.000
##  Molinia caerulea        : 558   Median :3.000   Median :2.000
##  Eriophorum vaginatum    : 524   Mean   :3.313   Mean   :2.311
##  Eriophorum angustifolium: 495   3rd Qu.:4.000   3rd Qu.:3.000
##  Deschampsia flexuosa    : 303   Max.   :9.000   Max.   :8.000
##  (Other)                 :7144   NA's   :963     NA's   :961
##      EBERGW          CSM_POS        CSM_NEG      EBERGR_site
##  Min.   : 2.000   Min.   :1      Min.   :1      Min.   :1.500
##  1st Qu.: 6.000   1st Qu.:1      1st Qu.:1      1st Qu.:2.400
##  Median : 7.000   Median :1      Median :1      Median :2.812
##  Mean   : 7.115   Mean   :1      Mean   :1      Mean   :3.309
##  3rd Qu.: 8.000   3rd Qu.:1      3rd Qu.:1      3rd Qu.:4.000
##  Max.   :10.000   Max.   :1      Max.   :1      Max.   :8.000
##  NA's   :963      NA's   :4192   NA's   :7575
##   EBERGN_site     EBERGW_site     CSM_POS_site    CSM_NEG_site
##  Min.   :1.000   Min.   :4.917   Min.   : 0.00   Min.   : 0.000
##  1st Qu.:1.625   1st Qu.:6.500   1st Qu.: 4.00   1st Qu.: 1.000
##  Median :2.125   Median :7.208   Median : 6.00   Median : 3.000
##  Mean   :2.301   Mean   :7.114   Mean   : 5.92   Mean   : 3.197
##  3rd Qu.:2.750   3rd Qu.:7.778   3rd Qu.: 8.00   3rd Qu.: 5.000
##  Max.   :6.400   Max.   :9.500   Max.   :16.00   Max.   :14.000
##
```

```
##     EASTINGS        NORTHINGS         country
##  Min.   :166826   Min.   : 13607   England:10431
##  1st Qu.:334033   1st Qu.:184769
##  Median :374500   Median :393102
##  Mean   :387407   Mean   :357367
##  3rd Qu.:487290   3rd Qu.:482799
##  Max.   :568388   Max.   :643250
##
```

All the LTMN data is from England unlike the other datasets which cover all of GB Currently the dataset retains the species-level information so that we can recalculate new indicators if we need to at a later date. However, Hannah has already calculated the square level average/sum indicator scores (EBERGR_site etc)

Note that the data are currently in Eastings and Northings. We will leave the data as-is for now but this will need to be harmonised between datasets at a later date

We can aggregate to plot level (by ID)

```
LTMN_plot <- aggregate(cbind(EBERGR_site, EBERGN_site, EBERGW_site, CSM_PO
S_site, CSM_NEG_site) ~ ID + plotID + scheme + site + year + yearF + yearO
S + NORTHINGS + EASTINGS + country, data = LTMN, FUN = mean)
```

Total of 1251 plot visits

```
summary(LTMN_plot)
```

```
##                     ID           plotID        scheme          site
##  LTMN_VC_B03_43_2012 :   1   Min.   :  0.00   LTMN:1270   B10    :100
##  LTMN_VC_B03_43_2017 :   1   1st Qu.: 14.00               B31    :100
##  LTMN_VC_B03_44_2012 :   1   Median : 28.00               B41    :100
##  LTMN_VC_B03_44a_2017:   1   Mean   : 32.03               B49    : 98
##  LTMN_VC_B03_45_2012 :   1   3rd Qu.: 43.00               B35    : 94
##  LTMN_VC_B03_45a_2017:   1   Max.   :128.00               B26    : 89
##  (Other)             :1264                                (Other):689
##       year          yearF          yearOS        NORTHINGS
##  Min.   :2010   Min.   :2010   Min.   : 1.00   Min.   : 13607
##  1st Qu.:2013   1st Qu.:2013   1st Qu.: 4.00   1st Qu.:166227
##  Median :2015   Median :2015   Median : 6.00   Median :391884
##  Mean   :2015   Mean   :2015   Mean   : 6.44   Mean   :355943
##  3rd Qu.:2018   3rd Qu.:2018   3rd Qu.: 9.00   3rd Qu.:495733
##  Max.   :2019   Max.   :2019   Max.   :10.00   Max.   :643250
##
##     EASTINGS         country      EBERGR_site     EBERGN_site
##  Min.   :166826   England:1270   Min.   :1.500   Min.   :1.000
##  1st Qu.:336801                  1st Qu.:2.286   1st Qu.:1.500
##  Median :377250                  Median :2.571   Median :2.000
##  Mean   :401619                  Mean   :2.980   Mean   :2.151
##  3rd Qu.:490188                  3rd Qu.:3.200   3rd Qu.:2.500
##  Max.   :568388                  Max.   :8.000   Max.   :6.400
##
##   EBERGW_site     CSM_POS_site     CSM_NEG_site
##  Min.   :4.917   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:6.359   1st Qu.: 3.000   1st Qu.: 0.000
```

```
##  Median :7.200   Median : 5.000   Median : 2.000
##  Mean   :7.088   Mean   : 4.913   Mean   : 2.249
##  3rd Qu.:7.778   3rd Qu.: 6.000   3rd Qu.: 3.000
##  Max.   :9.500   Max.   :16.000   Max.   :14.000
##
```

Histograms of each variables

Five variables of interest: count of CSM positive indicators, count of CSM negative indicators, mean Ellenberg N, mean Ellenberg R and mean Ellenberg W. Ellenberg values were not weighted by cover.

Distributional considerations:

1. CSM values are counts and so a Poisson or negative binomial distribution are likely to be most appropriate. We'll need to think about potential overdispersion (violating the assumption of equal mean and variance assumed by a Poisson distribution)

2. Ellenberg variables are continous. Although techically bound (see below) the mean values should lie sufficiently away from the bounds to be treated without modelling the bounds. This is something to check

Theoretical bounds of the Ellenberg values:

Each Ellenberg score associated with a plant species comes from a scale with defined limits. These vary between the different scores:

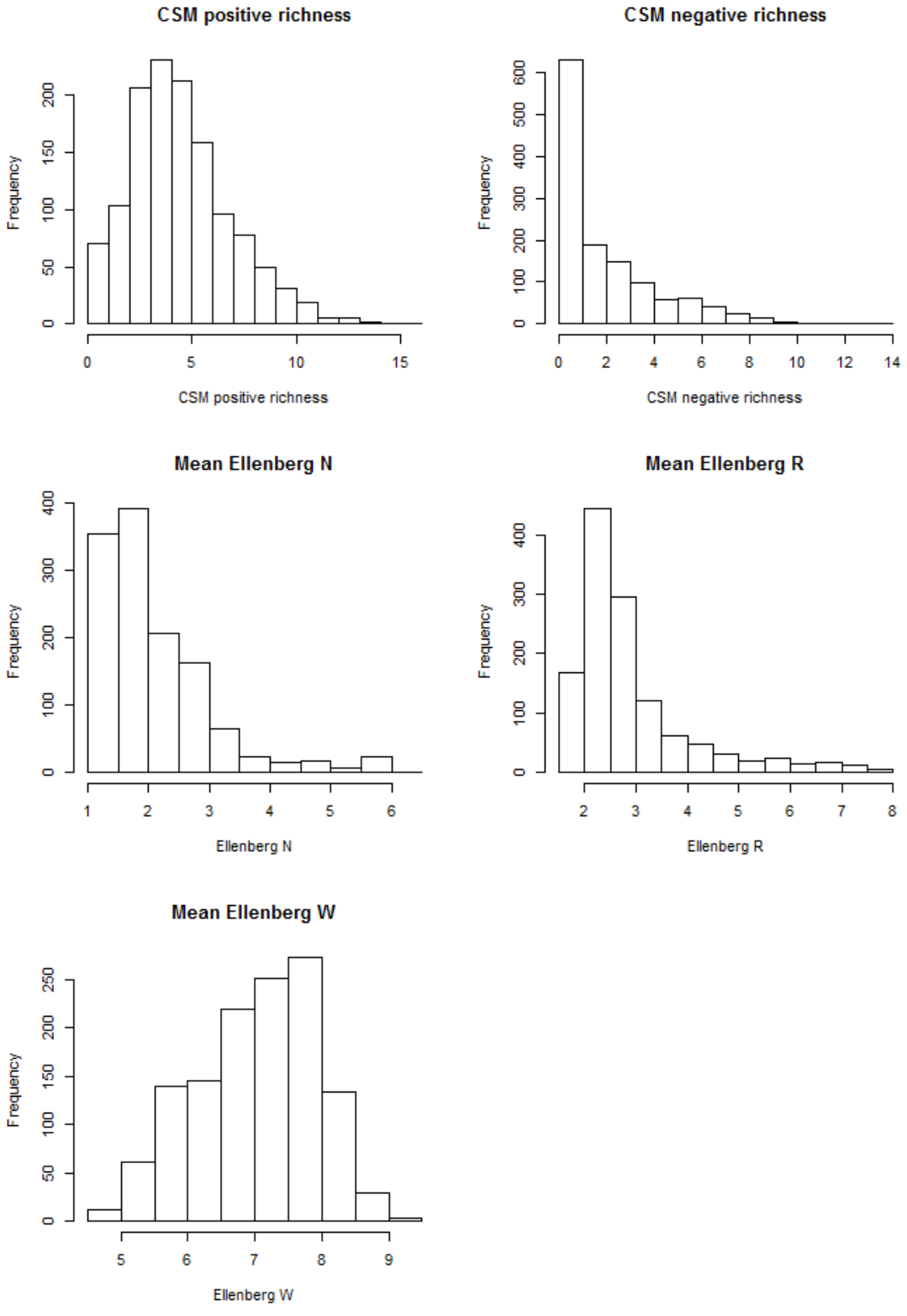
Ellenberg N: 1-9

Ellenberg R: 1-9

Ellenberg W: 1-12

To investigate the distributions of each response variable we can plot histograms:

```
par(mfrow=c(3,2))
hist(LTMN_plot$CSM_POS_site, main = "CSM positive richness", xlab = "CSM p
ositive richness")
hist(LTMN_plot$CSM_NEG_site, main = "CSM negative richness", xlab = "CSM n
egative richness")
hist(LTMN_plot$EBERGN_site, main = "Mean Ellenberg N", xlab = "Ellenberg N
")
hist(LTMN_plot$EBERGR_site, main = "Mean Ellenberg R", xlab = "Ellenberg R
")
hist(LTMN_plot$EBERGW_site, main = "Mean Ellenberg W", xlab = "Ellenberg W
")
```



Ellenberg N and R distributions are quite right skewed and hit the lower limits of the Ellenberg range which may be a problem. Ellenberg W is close enough to normal, although slightly left skewed. The CSM positive distribution looks a good fit to the Poisson, while the CSM negative distribution might be zero inflated?

We can calculate the mean and variance of the CSMs:

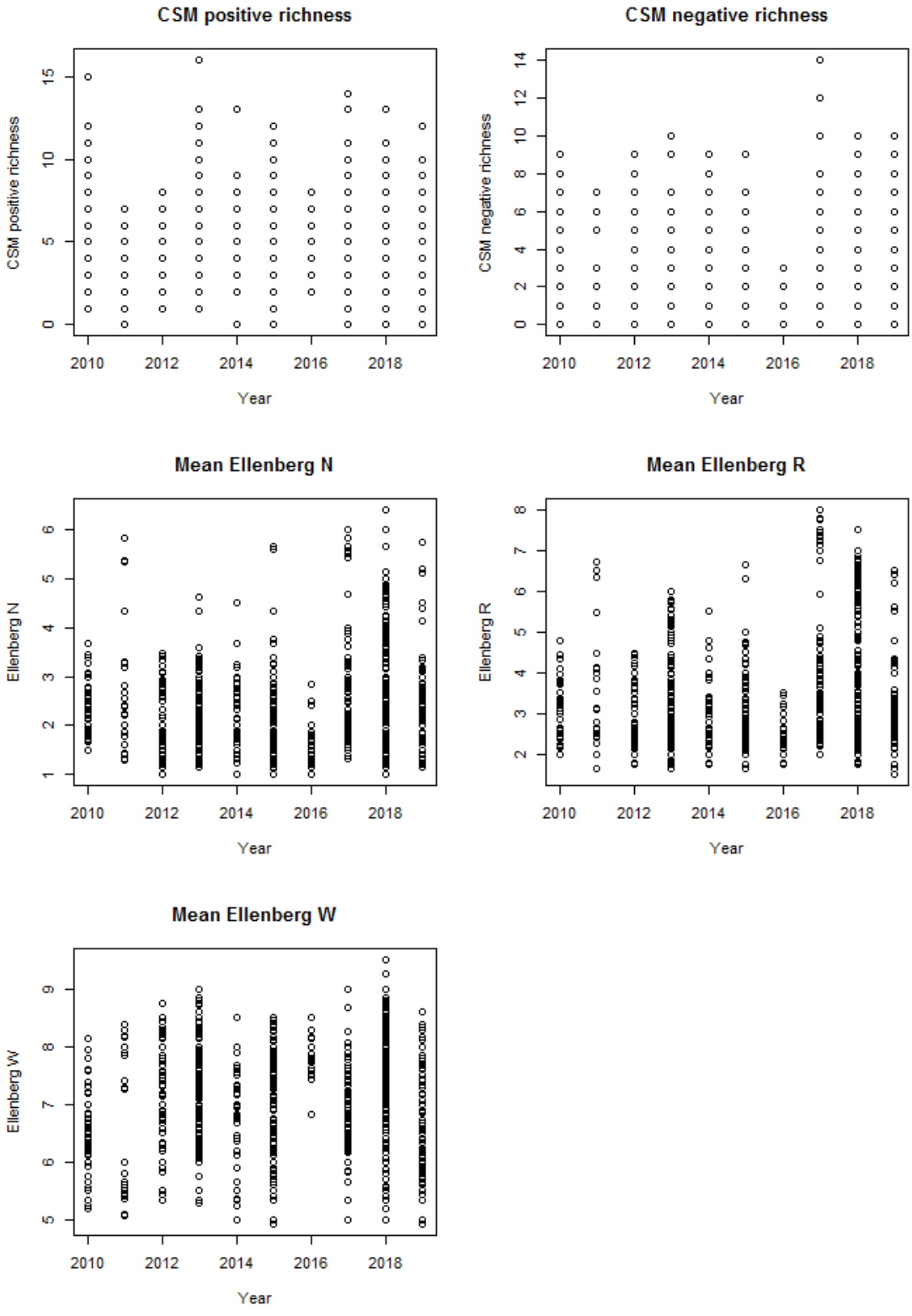
```
mean(LTMN_plot$CSM_POS_site);var(LTMN_plot$CSM_POS_site)

## [1] 4.912598

## [1] 6.267375

mean(LTMN_plot$CSM_NEG_site);var(LTMN_plot$CSM_NEG_site)

## [1] 2.248819

## [1] 5.27295
```

Neither show strong evidence of overdispersion

Plots of each variable against time

As the focus of this work is to look at trends in these variables over time (i.e. using year as a predictor) it is worth having a look at some exploratory scatterplots to identify any potentially non-linear relationships or datasets where there may not be sufficient temporal replication to calculate a trend.

```
par(mfrow=c(3,2))
plot(LTMN_plot$CSM_POS_site ~ LTMN_plot$year, main = "CSM positive richnes
s", ylab = "CSM positive richness", xlab = "Year")
plot(LTMN_plot$CSM_NEG_site ~ LTMN_plot$year, main = "CSM negative richnes
s", ylab = "CSM negative richness", xlab = "Year")
plot(LTMN_plot$EBERGN_site ~ LTMN_plot$year, main = "Mean Ellenberg N", yl
ab = "Ellenberg N", xlab = "Year")
plot(LTMN_plot$EBERGR_site ~ LTMN_plot$year, main = "Mean Ellenberg R", yl
ab = "Ellenberg R", xlab = "Year")
plot(LTMN_plot$EBERGW_site ~ LTMN_plot$year, main = "Mean Ellenberg W", yl
ab = "Ellenberg W", xlab = "Year")
```

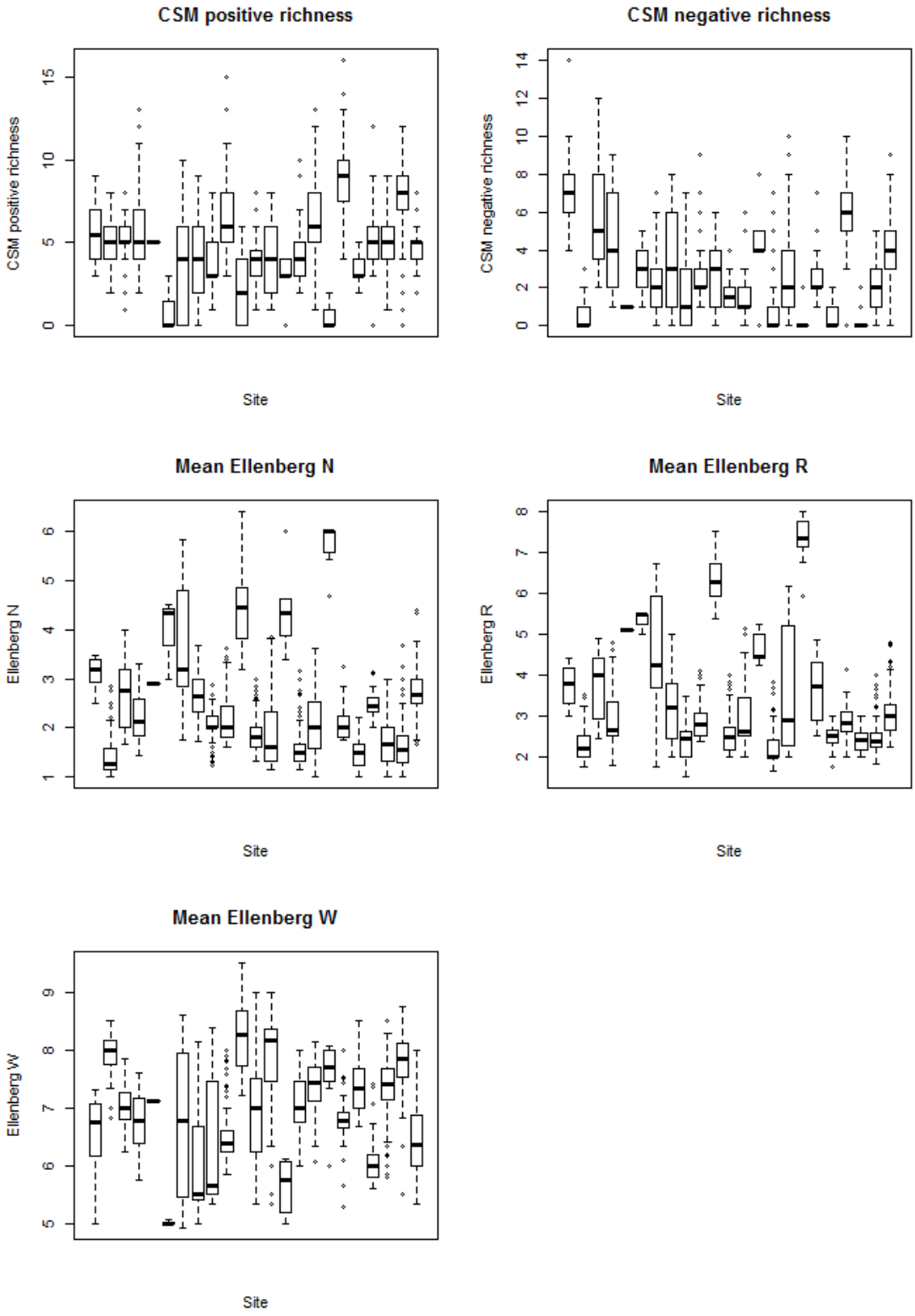


An issue here is that different cohorts of plots are measured in each year e.g. it appears that the plots measured in 2016 were particularly wet and acidic. This may make fitting a linear trend problematic.

Plots of each variable against site

The model we plan to fit considers repeat visits to plots but not nesting of plots within sites or squares. Therefore it is useful to consider how much variation is due to site.

```
par(mfrow=c(3,2))
par(mgp= c(3,1,0))
boxplot(LTMN_plot$CSM_POS_site ~ LTMN_plot$site, main = "CSM positive rich
ness", ylab = "CSM positive richness", xaxt =  "n", xlab = "Site")
boxplot(LTMN_plot$CSM_NEG_site ~ LTMN_plot$site, main = "CSM negative rich
ness", ylab = "CSM negative richness", xaxt =  "n", xlab = "Site")
boxplot(LTMN_plot$EBERGN_site ~ LTMN_plot$site, main = "Mean Ellenberg N",
ylab = "Ellenberg N", xaxt =  "n", xlab = "Site")
boxplot(LTMN_plot$EBERGR_site ~ LTMN_plot$site, main = "Mean Ellenberg R",
ylab = "Ellenberg R", xaxt =  "n", xlab = "Site")
boxplot(LTMN_plot$EBERGW_site ~ LTMN_plot$site, main = "Mean Ellenberg W",
ylab = "Ellenberg W", xaxt =  "n", xlab = "Site")
```



There is quite a lot of variation due to site i.e. plots within sites are generally more likely to be similar. This reflects all sorts of between-site differences we've not yet accounted for e.g. climate, N deposition. We will need to repeat this exercise once we've fit models accounting for these factors to assess whether there is still lots of variation due to site, or whether this has been explained by adding in the covariates.

Plots of each variable in space

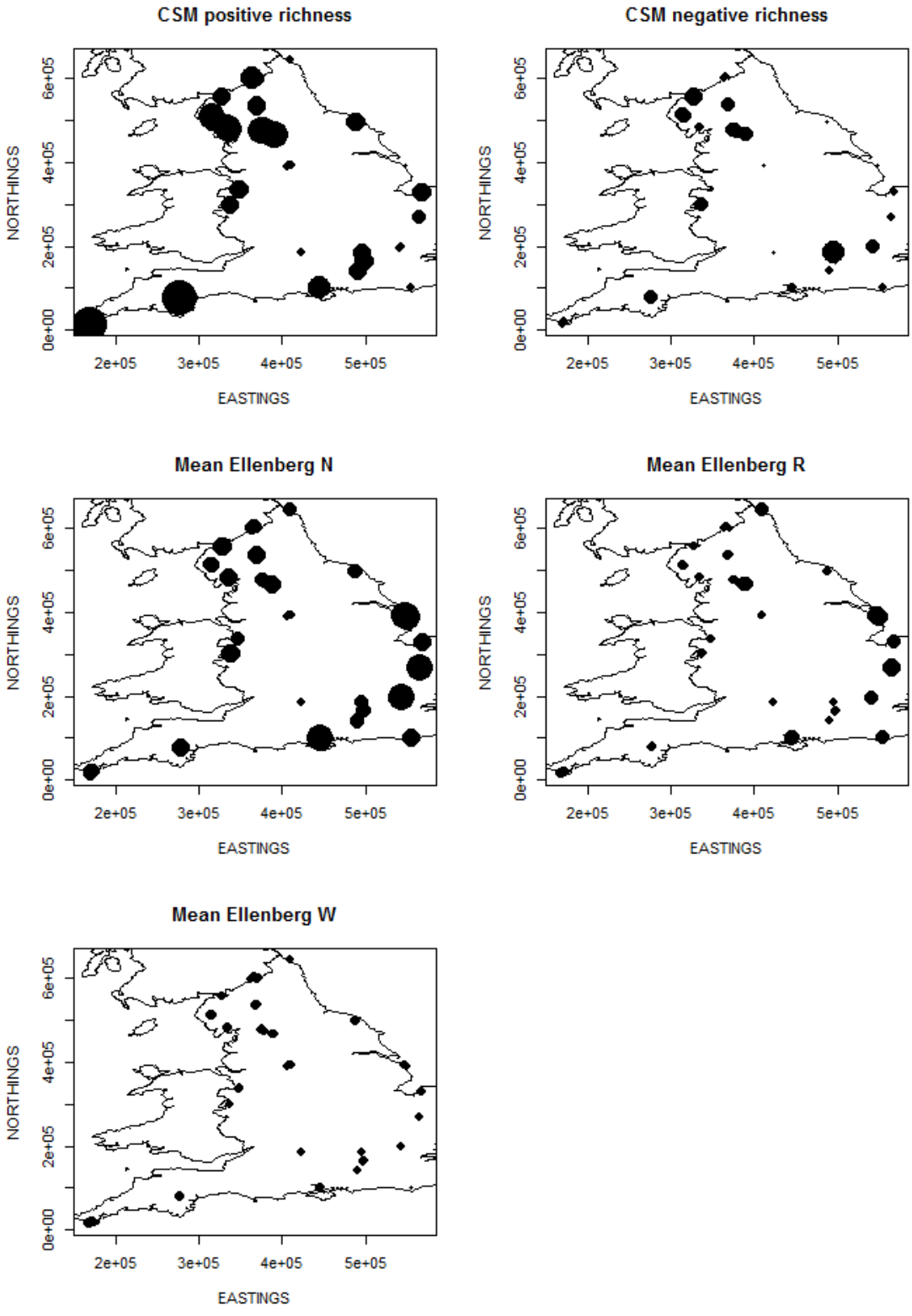
It is quite useful to think about plotting the variables in space, even in a very simplistic way. This will achieve two things:

1. Identify the spatial distribution of the data i.e. how well is the domain of interest covered?

2. Identify any potential spatial patterns in the response variables e.g. are CSM positive counts higher in the south for some reason? At this stage we've not included any covariates so spatial patterns may be due to climate, for example. We can investigate this later once we have the covariate data

The plots below show the locations of each measurement of the data, with the size of the point relative to the value of the response variable

```
#set up for figure by reading in file with GB outline

GB=read.table("GBoutline.txt",header=T)

par(mfrow=c(3,2))
par(mgp= c(3,1,0))
cex_ind <- round(LTMN_plot$CSM_POS_site/2)
plot(LTMN_plot$NORTHINGS ~ LTMN_plot$EASTINGS, pch = 20, cex = cex_ind, ma
in = "CSM positive richness", ylab = "NORTHINGS", xlab = "EASTINGS")
lines(GB)
cex_ind <- round(LTMN_plot$CSM_NEG_site/3)
plot(LTMN_plot$NORTHINGS ~ LTMN_plot$EASTINGS, pch = 20, cex = cex_ind, ma
in = "CSM negative richness", ylab = "NORTHINGS", xlab = "EASTINGS")
lines(GB)
cex_ind <- round(LTMN_plot$EBERGN)
plot(LTMN_plot$NORTHINGS ~ LTMN_plot$EASTINGS, pch = 20, cex = cex_ind, ma
in = "Mean Ellenberg N", ylab = "NORTHINGS", xlab = "EASTINGS")
lines(GB)
cex_ind <- round(LTMN_plot$EBERGR/2)
plot(LTMN_plot$NORTHINGS ~ LTMN_plot$EASTINGS, pch = 20, cex = cex_ind, ma
in = "Mean Ellenberg R", ylab = "NORTHINGS", xlab = "EASTINGS")
lines(GB)
cex_ind <- round(LTMN_plot$EBERGW/4)
plot(LTMN_plot$NORTHINGS ~ LTMN_plot$EASTINGS, pch = 20, cex = cex_ind, ma
in = "Mean Ellenberg W", ylab = "NORTHINGS", xlab = "EASTINGS")
lines(GB)
```



There are a couple of things to note:

1. Some of the plots seem to be in the sea off the coast of Cornwall… These look to be from site B40

2. The spread of sites is reasonable

3. There don't seem to be any strong geographic patterns in any variable

Summaries of data structure

There are a couple of other useful things we can extract about the data.

1. It would be useful to know how many times each plot has been revisited. The plot is going to be the unit over which we estimate the temporal autocorrelation so if a lot of plots have only been visited once then they won't contribute to this estimation

```
#calculate number of repeat visits (note LTMN plot IDs are not unique acro
ss sites)

LTMN_plot$plotID_new <- paste(LTMN_plot$site, LTMN_plot$plotID, sep = "_")

repvis <- tapply(LTMN_plot$year, LTMN_plot$plotID_new, function (x) length
(unique(x)))

#summarise
table(repvis)

## repvis
##   1   2   3
##  83 459  88
```

Most plots have been visited twice. With this dataset there is definitely no ability to look at autocorrelation within plots.

2. The number of plots per site might vary quite a lot between schemes so it would be good to extract some statistics about this

```
plotspersite <- tapply(LTMN_plot$plotID, LTMN_plot$site, function (x) leng
th(unique(x)))

table(plotspersite)

## plotspersite
##  1  3  5 11 12 15 18 20 25 28 33 34 38 40 42 47 49 50
##  2  1  1  1  1  1  1  2  1  1  1  1  2  1  1  1  1  3
```

This suggests that the range of plots per site is very large, from 1 to 57! Most sites have quite a lot of plots so we will probably need to think about a random site effect.

1.4. National Plant Monitoring Scheme data exploration

This document includes exploration of the NPMS data and the five calculated indicators (CSM + , CSM -, Ellenberg N, Ellenberg R and Ellenberg W)

First, load data.

This has already been filtered to include only heath and bog plots (as defined by the cluster analysis) and only inventory level surveys.

```
npms <- read.csv("NPMS_modelling_dataset_v3.csv")

summary(npms)
```

```
##                       ID           plotID            year
##  NPMS_154_145208_2017:  44   Min.   : 43894   Min.   :2015
##  NPMS_154_163653_2018:  43   1st Qu.:144793   1st Qu.:2016
##  NPMS_154_144927_2015:  41   Median :146622   Median :2017
##  NPMS_154_136287_2015:  37   Mean   :154221   Mean   :2017
##  NPMS_154_163653_2017:  37   3rd Qu.:165348   3rd Qu.:2018
##  NPMS_154_165425_2017:  37   Max.   :222145   Max.   :2019
##  (Other)             :6006
##      yearF          yearOS       scheme          site
##  Min.   :2015   Min.   :1.000   NPMS:6245   NN4282 : 311
##  1st Qu.:2016   1st Qu.:2.000               NC0429 : 297
##  Median :2017   Median :3.000               NC5762 : 291
##  Mean   :2017   Mean   :3.151               SU7253 : 277
##  3rd Qu.:2018   3rd Qu.:4.000               SN9143 : 276
##  Max.   :2019   Max.   :5.000               SN8834 : 252
##                                             (Other):4541
##                   species         EBERGR          EBERGN
##  Calluna vulgaris     : 272   Min.   :1.000   Min.   :1.000
##  Potentilla erecta    : 262   1st Qu.:3.000   1st Qu.:2.000
##  Erica tetralix       : 186   Median :4.000   Median :2.000
##  Molinia caerulea     : 185   Mean   :3.974   Mean   :2.999
##  Anthoxanthum odoratum: 144   3rd Qu.:5.000   3rd Qu.:4.000
##  Nardus stricta       : 133   Max.   :8.000   Max.   :9.000
##  (Other)              :5063   NA's   :654     NA's   :654
##      EBERGW          CSM_POS        CSM_NEG      EBERGR_site
##  Min.   : 2.000   Min.   :1      Min.   :1      Min.   :1.667
##  1st Qu.: 6.000   1st Qu.:1      1st Qu.:1      1st Qu.:3.000
##  Median : 7.000   Median :1      Median :1      Median :3.875
##  Mean   : 6.793   Mean   :1      Mean   :1      Mean   :3.974
##  3rd Qu.: 8.000   3rd Qu.:1      3rd Qu.:1      3rd Qu.:4.537
##  Max.   :11.000   Max.   :1      Max.   :1      Max.   :7.833
##  NA's   :654      NA's   :3690   NA's   :4221   NA's   :2
##   EBERGN_site     EBERGW_site     CSM_POS_site     CSM_NEG_site
##  Min.   :1.333   Min.   :4.500   Min.   : 0.000   Min.   : 0.00
##  1st Qu.:2.130   1st Qu.:6.154   1st Qu.: 5.000   1st Qu.: 3.00
##  Median :2.750   Median :6.800   Median : 7.000   Median : 6.00
##  Mean   :3.000   Mean   :6.795   Mean   : 7.581   Mean   : 6.68
##  3rd Qu.:3.558   3rd Qu.:7.421   3rd Qu.:10.000   3rd Qu.: 9.00
##  Max.   :6.500   Max.   :9.000   Max.   :18.000   Max.   :28.00
##  NA's   :2       NA's   :2
```

```
##     LATITUDE       LONGITUDE        CRS                    country
##  Min.   :50.16   Min.   :-7.483   4326:  35   Britain         :5866
##  1st Qu.:52.08   1st Qu.:-4.595   OSGB:5866   Northern_Ireland: 379
##  Median :54.25   Median :-3.618   OSIE: 344
##  Mean   :54.58   Mean   :-3.614
##  3rd Qu.:56.91   3rd Qu.:-2.856
##  Max.   :58.99   Max.   : 1.500
##
```

```
#still contains NI data which we'll remove for now

npms <- npms[npms$country != "Northern_Ireland",]
```

Currently the dataset retains the species-level information so that we can recalculate new indicators if we need to at a later date. However, Hannah has already calculated the square level average/sum indicator scores (EBERGR_site etc)

We can aggregate to plot level (by ID)

```
npms_plot <- aggregate(cbind(EBERGR_site, EBERGN_site, EBERGW_site, CSM_PO
S_site, CSM_NEG_site) ~ ID + plotID + scheme + site + year + yearF + yearO
S + LATITUDE + LONGITUDE + CRS + country, data = npms, FUN = mean)
```

Total of 433 plot visits (we removed 38 from NI)

```
summary(npms_plot)
```

```
##                      ID          plotID        scheme         site
##  NPMS_154_145208_2017:  2   Min.   : 43894   NPMS:431   NN4282 : 20
##  NPMS_154_182261_2017:  2   1st Qu.:144631              SU7253 : 20
##  NPMS_154_135109_2015:  1   Median :146657              NC5762 : 16
##  NPMS_154_135109_2016:  1   Mean   :155529              SE0173 : 15
##  NPMS_154_135109_2017:  1   3rd Qu.:165351              SN8834 : 15
##  NPMS_154_135109_2019:  1   Max.   :222145              NC0429 : 12
##  (Other)             :423                               (Other):333
##       year          yearF          yearOS         LATITUDE
##  Min.   :2015   Min.   :2015   Min.   :1.000   Min.   :50.16
##  1st Qu.:2016   1st Qu.:2016   1st Qu.:2.000   1st Qu.:51.81
##  Median :2017   Median :2017   Median :3.000   Median :54.15
##  Mean   :2017   Mean   :2017   Mean   :3.158   Mean   :54.27
##  3rd Qu.:2018   3rd Qu.:2018   3rd Qu.:4.000   3rd Qu.:56.80
##  Max.   :2019   Max.   :2019   Max.   :5.000   Max.   :58.99
##
##    LONGITUDE        CRS                   country     EBERGR_site
##  Min.   :-7.371   4326:  0   Britain         :431   Min.   :1.667
##  1st Qu.:-4.276   OSGB:431   Northern_Ireland:  0   1st Qu.:2.750
##  Median :-3.308   OSIE:  0                          Median :3.500
##  Mean   :-3.114                                     Mean   :3.914
##  3rd Qu.:-1.982                                     3rd Qu.:4.500
##  Max.   : 1.500                                     Max.   :7.833
##
##   EBERGN_site     EBERGW_site     CSM_POS_site     CSM_NEG_site
##  Min.   :1.333   Min.   :4.500   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:2.000   1st Qu.:6.185   1st Qu.: 3.000   1st Qu.: 2.000
```

```
##  Median :2.571   Median :6.750   Median : 6.000   Median : 4.000
##  Mean   :3.031   Mean   :6.790   Mean   : 6.316   Mean   : 5.151
##  3rd Qu.:3.598   3rd Qu.:7.400   3rd Qu.: 9.000   3rd Qu.: 7.000
##  Max.   :6.500   Max.   :9.000   Max.   :18.000   Max.   :28.000
##
```

Histograms of each variables

Five variables of interest: count of CSM positive indicators, count of CSM negative indicators, mean Ellernberg N, mean Ellenberg R and mean Ellenberg W. Ellenberg values were not weighted by cover.

Distributional considerations:

1. CSM values are counts and so a Poisson or negative binomial distribution are likely to be most appropriate. We'll need to think about potential overdispersion (violating the assumption of equal mean and variance assumed by a Poisson distribution)

2. Ellenberg variables are continous. Although techically bound (see below) the mean values should lie sufficiently away from the bounds to be treated without modelling the bounds. This is something to check

Theoretical bounds of the Ellenberg values:

Each Ellenberg score associated with a plant species comes from a scale with defined limits. These vary between the different scores:

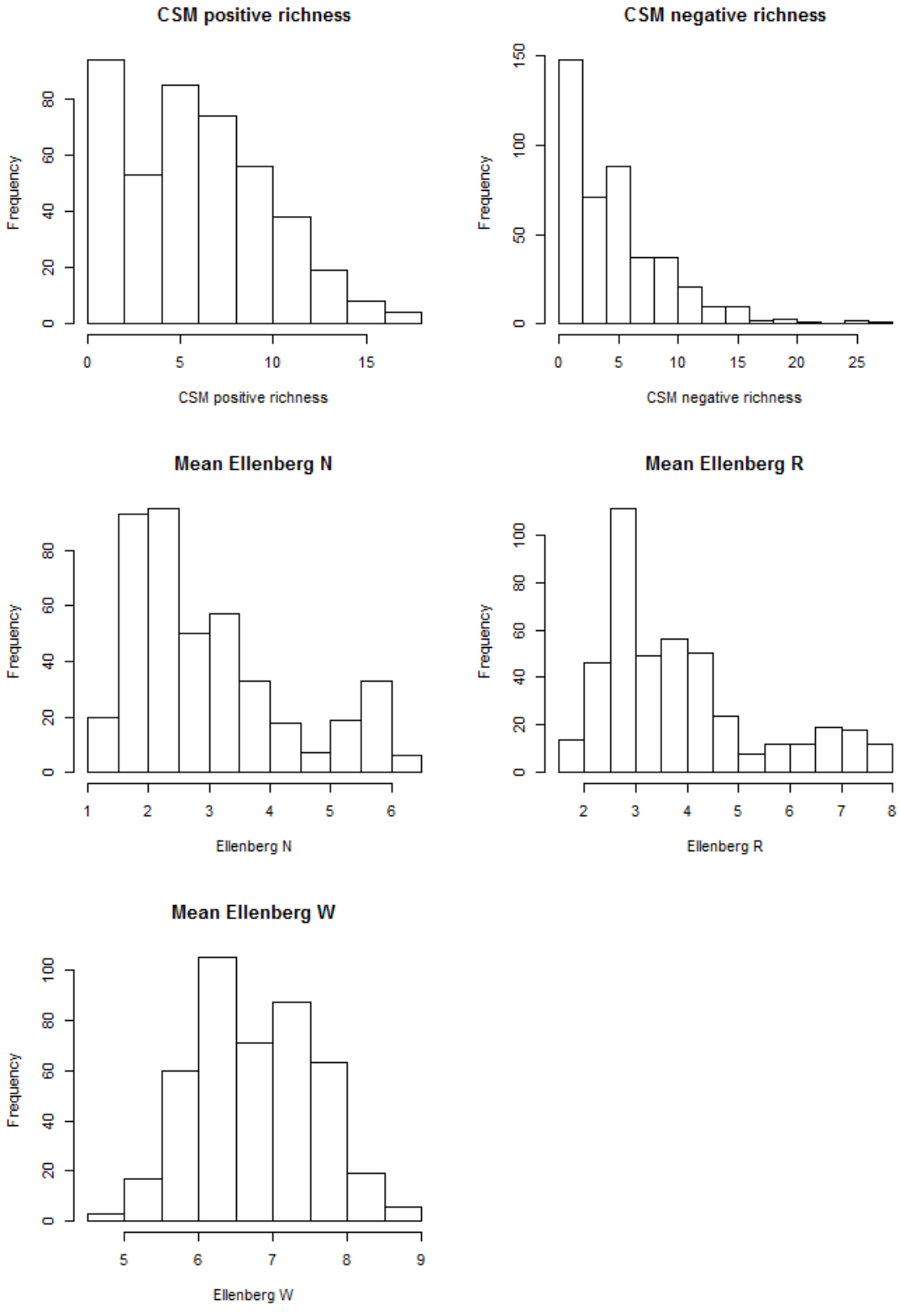
Ellenberg N: 1-9

Ellenberg R: 1-9

Ellenberg W: 1-12

To investigate the distributions of each response variable we can plot histograms:

```
par(mfrow=c(3,2))
hist(npms_plot$CSM_POS_site, main = "CSM positive richness", xlab = "CSM p
ositive richness")
hist(npms_plot$CSM_NEG_site, main = "CSM negative richness", xlab = "CSM n
egative richness")
hist(npms_plot$EBERGN_site, main = "Mean Ellenberg N", xlab = "Ellenberg N
")
hist(npms_plot$EBERGR_site, main = "Mean Ellenberg R", xlab = "Ellenberg R
")
hist(npms_plot$EBERGW_site, main = "Mean Ellenberg W", xlab = "Ellenberg W
")
```



Interestingly, it appears that whilst the range of Ellenberg W does not come close to the bounds, both Ellenberg N and Ellenberg R distributions have values close to the theoretical bounds. This is most evident for Ellenberg N which has communities with a mean Ellenberg N or 1.

Ellenberg W has a roughly normal distribution but both Ellenberg N and R show skewed and slightly bimodal distributions. We might need to consider appropriate distributions carefully, although we can use a normal initially and inspect residuals.

At first glance it appears that overdispersion may be more of a problem for CSM negative richness than CSM positive richness. We can look at this in more detail by calculating the mean and variance of each variable and comparing:

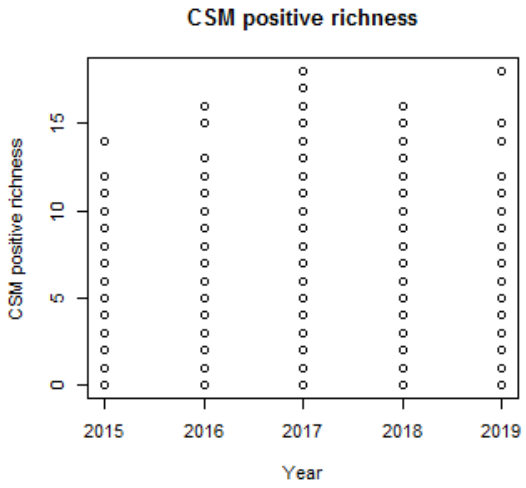
```
mean(npms_plot$CSM_POS_site);var(npms_plot$CSM_POS_site)

## [1] 6.315545

## [1] 16.4909

mean(npms_plot$CSM_NEG_site);var(npms_plot$CSM_NEG_site)

## [1] 5.150812

## [1] 20.75627
```

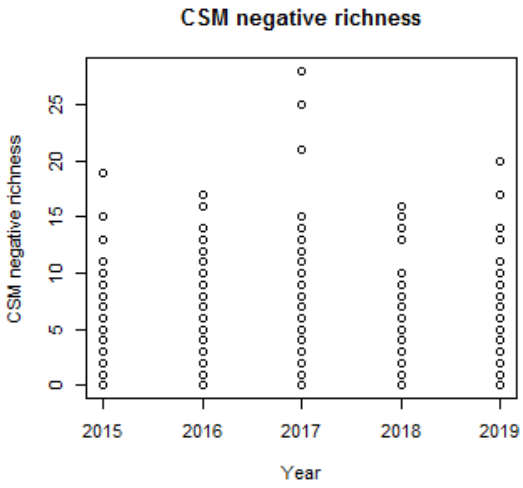
Both show larger variance than mean. This is probably enough to justify modelling a separate variance component (using a negative binomial or quasipoisson).

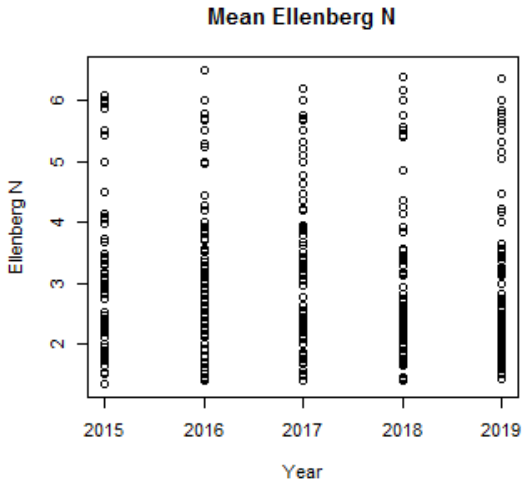
Plots of each variable against time

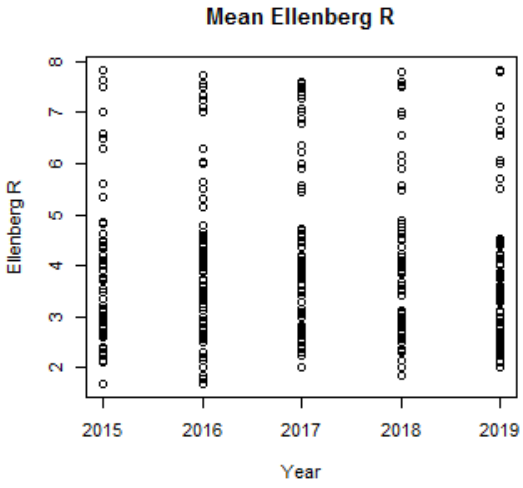
As the focus of this work is to look at trends in these variables over time (i.e. using year as a predictor) it is worth having a look at some exploratory scatterplots to identify any potentially non-linear relationships or datasets where there may not be sufficient temporal replication to calculate a trend.

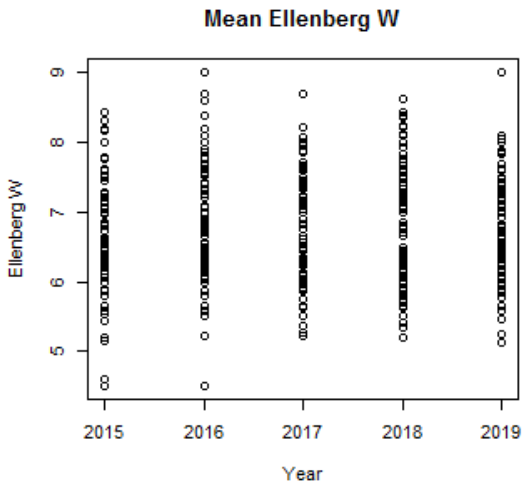
```
par(mfrow=c(3,2))
plot(npms_plot$CSM_POS_site ~ npms_plot$year, main = "CSM positive richnes
s", ylab = "CSM positive richness", xlab = "Year")
plot(npms_plot$CSM_NEG_site ~ npms_plot$year, main = "CSM negative richnes
s", ylab = "CSM negative richness", xlab = "Year")
plot(npms_plot$EBERGN_site ~ npms_plot$year, main = "Mean Ellenberg N", yl
ab = "Ellenberg N", xlab = "Year")
plot(npms_plot$EBERGR_site ~ npms_plot$year, main = "Mean Ellenberg R", yl
ab = "Ellenberg R", xlab = "Year")
plot(npms_plot$EBERGW_site ~ npms_plot$year, main = "Mean Ellenberg W", yl
ab = "Ellenberg W", xlab = "Year")
```











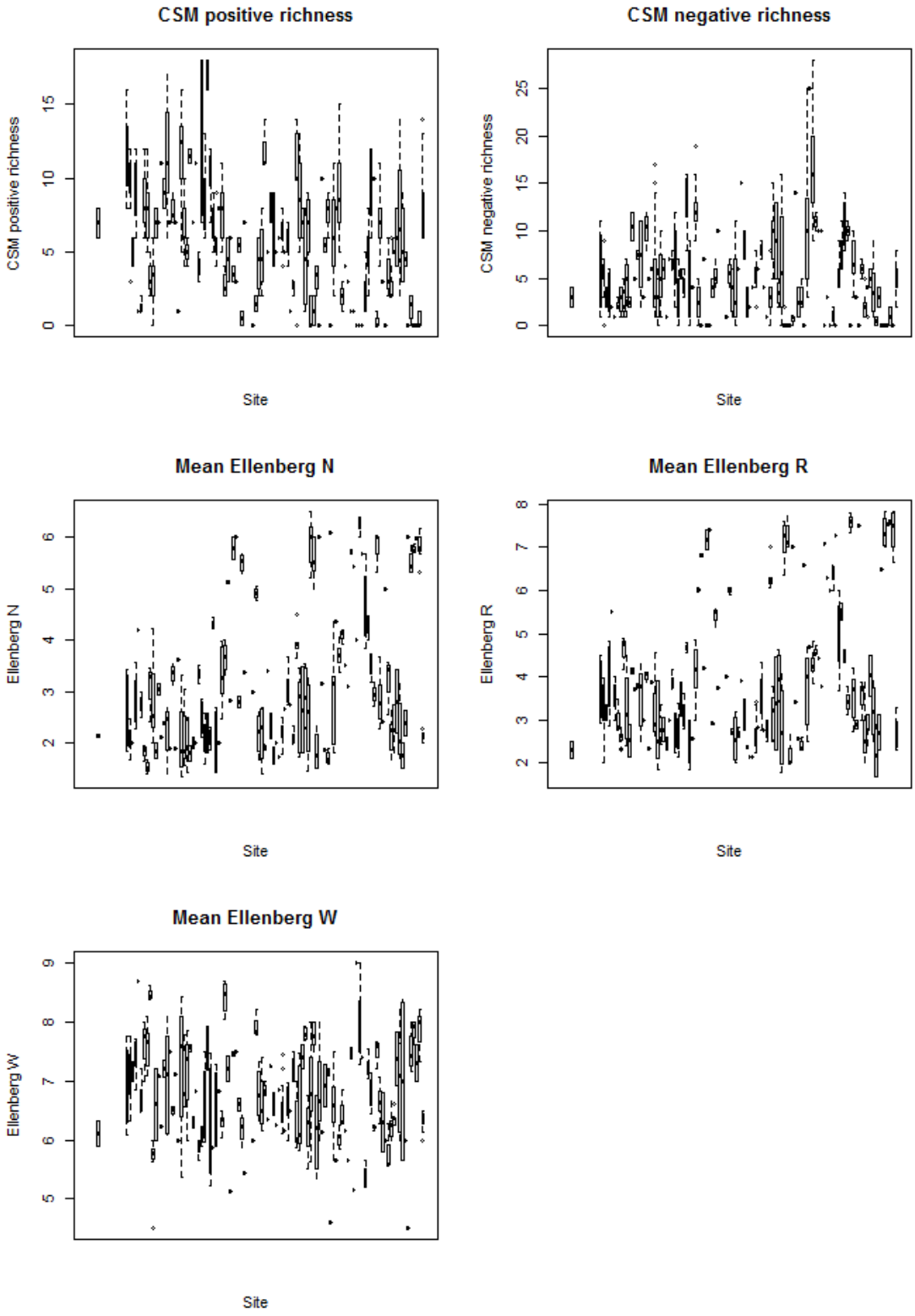
In this dataset the main issue is the small number of survey years. However, fitting a continuous trend against year is probably just about justifiable.

Unsuprisingly there is little evidence of temporal trends in the data, reflecting the short time span covered. There are not enough years to evaluate non-linearity.

Plots of each variable against site

The model we plan to fit considers repeat visits to plots but not nesting of plots within sites or squares. Therefore it is useful to consider how much variation is due to site.

```
par(mfrow=c(3,2))
par(mgp= c(3,1,0))
boxplot(npms_plot$CSM_POS_site ~ npms_plot$site, main = "CSM positive rich
ness", ylab = "CSM positive richness", xaxt =  "n", xlab = "Site")
boxplot(npms_plot$CSM_NEG_site ~ npms_plot$site, main = "CSM negative rich
ness", ylab = "CSM negative richness", xaxt =  "n", xlab = "Site")
boxplot(npms_plot$EBERGN_site ~ npms_plot$site, main = "Mean Ellenberg N",
ylab = "Ellenberg N", xaxt =  "n", xlab = "Site")
boxplot(npms_plot$EBERGR_site ~ npms_plot$site, main = "Mean Ellenberg R",
ylab = "Ellenberg R", xaxt =  "n", xlab = "Site")
boxplot(npms_plot$EBERGW_site ~ npms_plot$site, main = "Mean Ellenberg W",
ylab = "Ellenberg W", xaxt =  "n", xlab = "Site")
```



There is quite a lot of variation due to site i.e. plots within sites are generally more likely to be similar. This reflects all sorts of between-site differences we've not yet accounted for e.g. climate, N deposition. We will need to repeat this exercise once we've fit models accounting for these factors to assess whether there is still lots of variation due to site, or whether this has been explained by adding in the covariates.

Plots of each variable in space

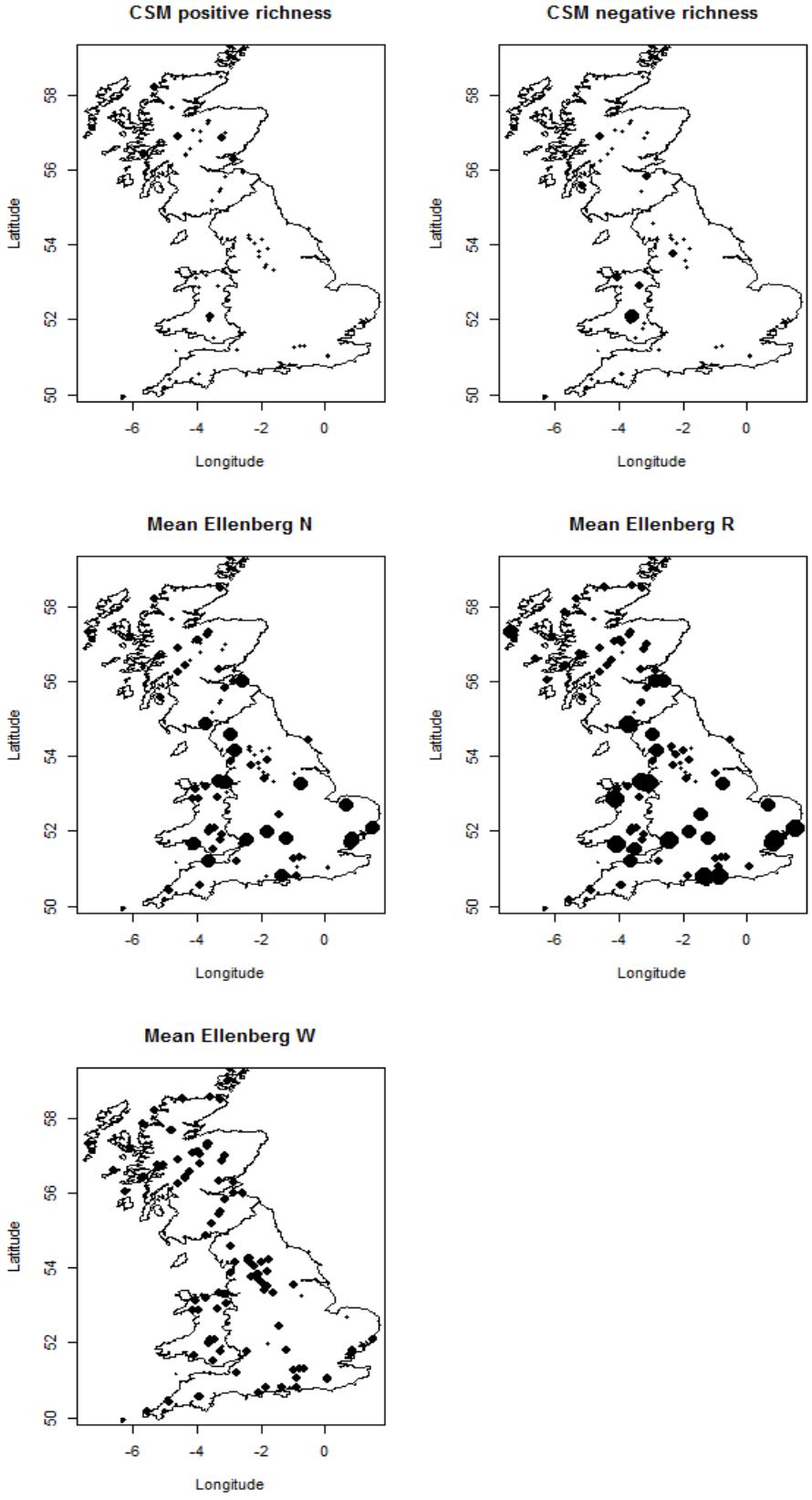
It is quite useful to think about plotting the variables in space, even in a very simplistic way. This will achieve two things:

1. Identify the spatial distribution of the data i.e. how well is the domain of interest covered?

2. Identify any potential spatial patterns in the response variables e.g. are CSM positive counts higher in the south for some reason? At this stage we've not included any covariates so spatial patterns may be due to climate, for example. We can investigate this later once we have the covariate data

The plots below show the locations of each measurement of the data, with the size of the point relative to the value of the response variable

```
#set up for figure by reading in file with GB outline

GB=read.table("GBoutline_latlong.txt",header=T)

par(mfrow=c(3,2))
par(mgp= c(3,1,0))
cex_ind <- round(npms_plot$CSM_POS_site/10)
plot(npms_plot$LATITUDE ~ npms_plot$LONGITUDE, pch = 20, cex = cex_ind, ma
in = "CSM positive richness", ylab = "Latitude", xlab = "Longitude")
lines(GB)
cex_ind <- round(npms_plot$CSM_NEG_site/10)
plot(npms_plot$LATITUDE ~ npms_plot$LONGITUDE, pch = 20, cex = cex_ind, ma
in = "CSM negative richness", ylab = "Latitude", xlab = "Longitude")
lines(GB)
cex_ind <- round(npms_plot$EBERGN/2)
plot(npms_plot$LATITUDE ~ npms_plot$LONGITUDE, pch = 20, cex = cex_ind, ma
in = "Mean Ellenberg N", ylab = "Latitude", xlab = "Longitude")
lines(GB)
cex_ind <- round(npms_plot$EBERGR/2)
plot(npms_plot$LATITUDE ~ npms_plot$LONGITUDE, pch = 20, cex = cex_ind, ma
in = "Mean Ellenberg R", ylab = "Latitude", xlab = "Longitude")
lines(GB)
cex_ind <- round(npms_plot$EBERGW/4)
plot(npms_plot$LATITUDE ~ npms_plot$LONGITUDE, pch = 20, cex = cex_ind, ma
in = "Mean Ellenberg W", ylab = "Latitude", xlab = "Longitude")
lines(GB)
```



A few things to note from these basic maps:

1. The distribution of NPMS plots at heath and bog sites is reasonably good. There is probably an over-representation of this habitat in England and an under-representation in Scotland but this is not too dramatic.

2. The distribution of sites defined as heath and bog seems reasonable given the distributions of these habitat types (i.e. no plots in central London, plots largely concentrated in upland areas)

3. For Ellenberg N and Ellenberg R there is some indication that more southerly sites tend to have larger values. However, it is difficult to conclude much from these patterns as at each location there are likely to be multiple plot values overlaid.

Summaries of data structure

There are a couple of other useful things we can extract about the data.

1. It would be useful to know how many times each plot has been revisited. The plot is going to be the unit over which we estimate the temporal autocorrelation so if a lot of plots have only been visited once then they won't contribute to this estimation

```
#calculate number of repeat visits
repvis <- tapply(npms_plot$year, npms_plot$plotID, function (x) length(uni
que(x)))

#summarise
table(repvis)

## repvis
##  1  2  3  4  5
## 72 51 26 23 17
```

38% plots have only a single visit meaning they won't contribute to calculating the autocorrelation coefficient. Only 9% of plots have all five visits. This means we do not expect particularly good estimates of temporal autocorrelation in this dataset.

2. The number of plots per site might vary quite a lot between schemes so it would be good to extract some statistics about this

```
plotspersite <- tapply(npms_plot$plotID, npms_plot$site, function (x) leng
th(unique(x)))

table(plotspersite)

## plotspersite
##  1  2  3  4  5
## 51 27 15  9  1
```

This suggests that the majority of NPMS squares have a single plot. This might contrast substantially with other schemes where the number of plots may be much higher.

BANGOR
UK Centre for Ecology & Hydrology
Environment Centre Wales
Deiniol Road
Bangor
Gwynedd
LL57 2UW
United Kingdom
T: +44 (0)1248 374500
F: +44 (0)1248 362133

EDINBURGH
UK Centre for Ecology & Hydrology
Bush Estate
Penicuik
Midlothian
EH26 0QB
United Kingdom
T: +44 (0)131 4454343
F: +44 (0)131 4453943

LANCASTER
UK Centre for Ecology & Hydrology
Lancaster Environment Centre
Library Avenue
Bailrigg
Lancaster
LA1 4AP
United Kingdom
T: +44 (0)1524 595800
F: +44 (0)1524 61536

WALLINGFORD (Headquarters)
UK Centre for Ecology & Hydrology
Maclean Building
Benson Lane
Crowmarsh Gifford
Wallingford
Oxfordshire
0X10 8BB
United Kingdom
T: +44 (0)1491 838800
F: +44 (0)1491 692424

enquiries@ceh.ac.uk

www.ceh.ac.uk